



SO ERSTELLEN SIE IHR DATA MESH AUF SNOWFLAKE

Data Mesh¹ hat sich in den letzten Jahren zu einem immer beliebteren Ansatz für die Datenverwaltung entwickelt. Unternehmen aller Branchen entscheiden sich für ein Data Mesh zur dezentralen Datenverwaltung, um die Agilität der Daten zu verbessern und die organisatorischen Engpässe zu vermeiden, die oft mit zentralisierten und monolithischen Ansätzen verbunden sind.

In diesem Whitepaper wird der Ansatz von Snowflake für das Data Mesh erläutert. Beschrieben werden einige der wichtigsten Funktionen von Snowflake für ein Data Mesh. Dabei werden typische Architekturoptionen vorgestellt, die unsere Kunden gewählt haben, um eine Self-Service-Datenplattform zu implementieren, die verteilte Domänen unterstützt.

Das Data Mesh ist in erster Linie ein organisatorischer Ansatz, der die Zuständigkeiten und die Koordination zwischen den einzelnen Domänen-Teams und ihren Datenprodukten definiert. Es wird jedoch die richtige Technologie benötigt, damit die Domänen dem Data-Mesh-Ansatz auf praktikable Weise folgen können.

„ Beim Data Mesh geht es nicht um Technologie [...], aber Sie brauchen die richtige Technologie, um die Datenproduktteams mit einer Vielzahl von Möglichkeiten auszustatten. Nicht jedes Domänen- und Datenproduktteam muss das Rad neu erfinden und seine Daten- und Analyseplattform von Grund auf neu aufbauen. Ebenso müssen wir es für die Datenproduktteams einfach machen. Sonst gibt es kein Empowerment und keine Dezentralisierung.“

- OMAR KHAWAJA, Global Head BI, Roche (2022)

Snowflake wird von vielen Unternehmen, die einen Data-Mesh-Ansatz verfolgen, erfolgreich als Datenplattform eingesetzt. Es gibt keine einzige Technologieplattform, die eine vollständige End-to-End-Lösung zur Unterstützung des Data-Mesh-Ansatzes bietet. Snowflake bietet jedoch viele der Funktionen, die für eine Self-Service-Datenplattform erforderlich sind. Es ermöglicht eine verteilte, bereichsorientierte Architektur und bietet Funktionen, die bei der Implementierung von Daten als Produkt und einer föderierten Computer-Governance hilfreich sind.

DER ANSATZ VON SNOWFLAKE FÜR DAS DATA MESH

Nach der Zusammenarbeit mit zahlreichen Kunden bei ihren Data-Mesh-Initiativen hat sich Snowflake für den folgenden Ansatz entschieden:

- Wir wissen, dass das Data Mesh in erster Linie eine organisatorische Umstrukturierung bietet. Diese Umstrukturierung hat viele nicht-technische Auswirkungen, erfordert aber oft auch Änderungen auf der Ebene der IT-Architektur und der Technologie.
- Bleiben Sie pragmatisch. Wir raten unseren Kunden, sich nicht die Implementierung des „perfekten“ Data Mesh zum Ziel zu setzen, sondern sich an ihren spezifischen Problemen und Zielen zu orientieren. So sind beispielsweise polyglotte Speicherung und multimodaler Zugriff nützliche Konzepte, aber Unternehmen sollten sich auf ihre tatsächlichen Anforderungen konzentrieren, um die Wirkung zu maximieren.
- Beginnen Sie klein, erweitern Sie schrittweise und arbeiten Sie sich im Laufe der Zeit entlang der Data-Mesh-Reifekurve nach oben. Beginnen Sie beispielsweise mit einer oder zwei Domänen und Datenprodukten, um einen unmittelbaren Geschäftsbedarf zu decken, und nutzen Sie dann den frühen Erfolg, um das Netz zu erweitern.
- Achten Sie auf die Kosten und die Komplexität. So hat es sich zum Beispiel als vorteilhaft erwiesen, die Palette der Tools in der Self-Service-Datenplattform so klein und einheitlich wie möglich zu halten und gleichzeitig alle kritischen Domänenanforderungen zu erfüllen.
- Definieren Sie frühzeitig Anreize und Erfolgskriterien, einschließlich messbarer KPIs für Domänen, Datenprodukte, die Self-Service-Datenplattform und Governance-Kontrollen.
- Es gibt keine sofort einsatzbereite Data-Mesh-Lösung. Wir nutzen unser umfangreiches Partnernetzwerk, um gemeinsam Lösungen zu entwickeln, die den Anforderungen unserer Kunden entsprechen. So sind beispielsweise Tools für Data Governance, Automatisierung, DevOps und andere Bereiche häufig Teil einer Data-Mesh-Lösung, auch wenn sie in diesem Artikel nicht näher erläutert werden.

¹ www.thoughtworks.com/en-us/what-we-do/data-and-ai/data-mesh

RELEVANTE FUNKTIONEN VON SNOWFLAKE

Snowflake bietet eine Reihe von Schlüsselfunktionen, die unsere Kunden beim Aufbau der Self-Service-Datenplattform für ein Data Mesh als hilfreich empfunden haben.

Snowflake ist weit mehr als ein Cloud Data Warehouse

Snowflake ist ein integrierter Cloud-Service-Anbieter, der eine breite Palette an Funktionen für Data Engineering, Data Lakes, Data Warehousing, Data Sharing und wesentliche Teile eines typischen Lebenszyklus für maschinelles Lernen bietet.

Insbesondere können Benutzer:innen Datentransformations-Pipelines erstellen und automatisieren, um verschiedene Eingabedaten in kontrollierte Datenprodukte zu verwandeln. Snowflake kann gängige Dateiformate in Ihren Cloud Storage Buckets ebenso einfach verarbeiten

wie Eingabeströme (z. B. aus Kafka) oder relationale Tabellen. Zu den unterstützten Dateiformaten gehören JSON, XML, Parquet, AVRO, Delta Lake², Apache Iceberg³ und andere. Snowflake bietet außerdem Unterstützung für unstrukturierte Daten wie Bilder, Videos oder andere Binärformate. Daten können in der Snowflake-Plattform mit SQL, Python⁴, Scala, Java und Javascript oder durch den Aufruf externer Funktionen auf der breiteren Cloud-Plattform manipuliert werden.

Snowflake bietet vielleicht nicht alle Funktionen, die Ihre Domänenteams benötigen, aber es bietet eine beachtliche Bandbreite an Funktionen in einem *einzigem* Dienst, für deren Integration sonst eine *Sammlung* von Cloud-Diensten erforderlich wäre. Eine solche Integration kann komplex und zeitaufwendig sein und erfordert hochqualifizierte Fachkräfte.



ABBILDUNG 1: SNOWFLAKE ALS EINE EINZIGE PLATTFORM FÜR VERSCHIEDENE ARTEN VON DATEN UND WORKLOADS

² Zum Zeitpunkt der Veröffentlichung, August 2022, in Public Preview.

³ Zum Zeitpunkt der Veröffentlichung, August 2022, in Private Preview.

⁴ Zum Zeitpunkt der Veröffentlichung, August 2022, in Public Preview.

Snowflake ist eine verteilte Plattform, kein Monolith

Snowflake ist eine verteilte, aber miteinander vernetzte Plattform, die Silos vermeidet und es verteilten Teams ermöglicht, Daten auf kontrollierte und sichere Weise auszutauschen. Wie funktioniert das? Ein Unternehmen kann ein Konto oder mehrere Konten bei Snowflake erstellen, die sich in derselben oder in verschiedenen Cloud-Regionen und -Plattformen befinden können (Abbildung 2). Jedes Konto kann mehrere separate Datenbanken beherbergen, für die Rechen- und Speicherressourcen unabhängig voneinander bereitgestellt und skaliert werden können, und zwar auf verteilte Weise.

Verschiedene Domänenteams können autonom arbeiten, indem sie unabhängige Rechenleistung in separaten Datenbanken oder sogar in separaten Konten nutzen, während sie gleichzeitig die zugrunde liegende Snowflake-Plattform verwenden, um Daten-Assets miteinander zu teilen. Beachten Sie, dass das Snowflake-Konzept einer „Datenbank“ nicht nur eine traditionelle relationale Datenbank umfasst, sondern auch alle anderen funktionalen Fähigkeiten in Snowflake wie Data Engineering, Data Lake, Data Warehousing, Data Sharing und Data Science. Die Verwendung von Rechen-Clustern zur Kombination und Verarbeitung von Daten aus mehreren Datenbanken oder Konten ist eine Kernfunktion der Plattform von Snowflake.

Snowflake verfügt über integrierte Funktionen für den Datenaustausch und Marktplätze

Datenproduzenten in Snowflake können Daten, Datendienste oder Anwendungen mit anderen Konten teilen, indem sie Metadaten („Listings“)

veröffentlichen. Mithilfe der Discovery-Kontrollen für Auflistungen können Produzenten ihre Inhalte privat mit anderen Konten oder einer Gruppe von Konten oder öffentlich über den Snowflake Marketplace teilen. Datenproduzenten können SLAs oder SLOs für die von ihnen freigegebenen Daten angeben, z. B. die Aktualisierungshäufigkeit, den Umfang der Historie, die zeitliche Granularität der Daten und andere Eigenschaften, die helfen, die Daten als Produkt zu beschreiben.

Andere Teams können nach relevanten, für sie verfügbaren Datenbeständen suchen und Zugang erhalten oder anfordern. Solche Data Consumers erhalten Live-Zugriff auf die gemeinsam genutzten Daten, die unter der Kontrolle des Produzenten bleiben, der die Zugriffsrichtlinien anpassen oder den Zugriff jederzeit widerrufen kann. Der Zugriff auf gemeinsam genutzte Daten erfordert keinen ETL- oder Datenverschiebungsprozess, der vom Produzenten oder Nutzer implementiert werden muss. Produzenten können auch externe Tabellen veröffentlichen und gemeinsam nutzen. Dabei handelt es sich um „Ansichten“ von Dateien, die außerhalb von Snowflake gespeichert sind und die optional Delta Lake- und Iceberg-Formate enthalten können. Produzenten können sogar Daten mit Dritten außerhalb des Unternehmens austauschen, selbst wenn diese Parteien keine aktiven Snowflake-Kunden sind. So kann ein Datenproduzent beispielsweise Daten über ein sogenanntes Snowflake-Leserkonto und die gesamte Bandbreite der unterstützten APIs extern freigeben. Sie können aber auch regelmäßig (partitionierte) Daten in einen Cloud Storage Bucket exportieren und dabei eines der heute gängigen Dateiformate verwenden.

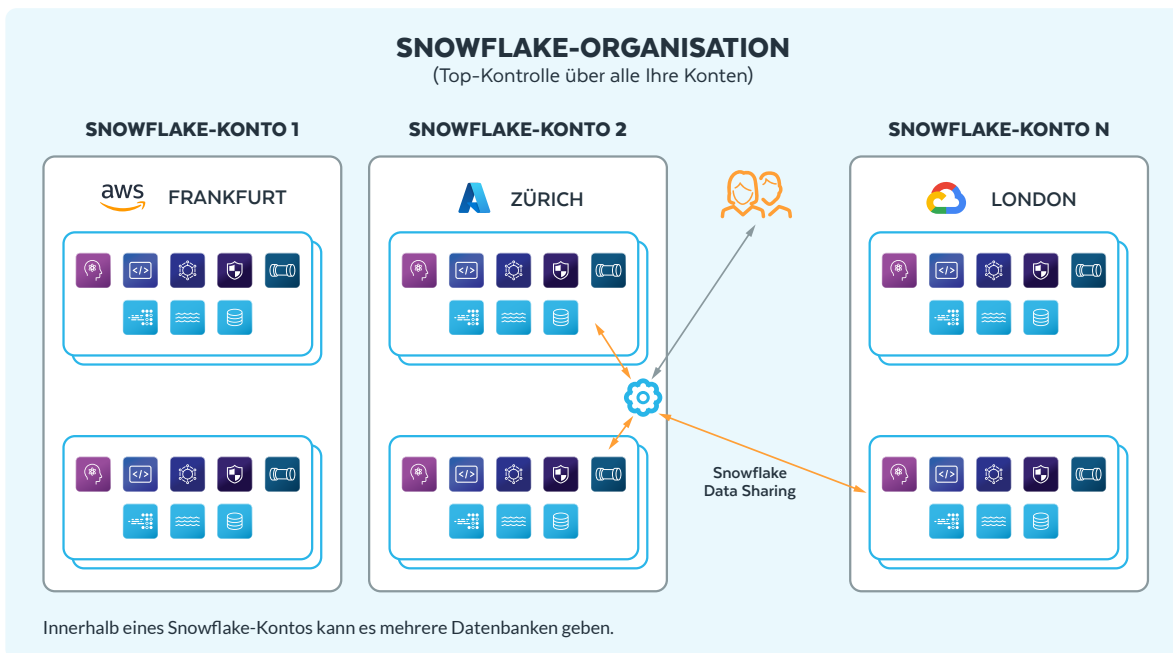


ABBILDUNG 2: SNOWFLAKES ORGANISATION, KONTEN UND DATENBANKEN UNTERSTÜTZEN EINE VERTEILTE ARCHITEKTUR

Snowflake bietet eine breite Palette an Sicherheits- und Governance-Funktionen

Föderierte Governance ist wohl eine der größten Herausforderungen bei der Entwicklung eines Data Mesh und erfordert oft ein oder mehrere Tools in Kombination, um alle Anforderungen zu erfüllen. Auf der Plattformebene unterstützt Snowflake rollenbasierte Zugriffskontrolle, Richtlinien für den Zugriff auf Zeilenebene, Datenmaskierung auf Spaltenebene, externe Tokenisierung sowie Data Lineage (Datenabstammung), Audit-Funktionen und mehr. Benutzer:innen können auch ein oder mehrere Metadaten-Tags (Schlüssel-Wert-Paare) für fast alle Arten von Objekten in Snowflake zuweisen, z. B. Konten, Datenbanken, Schemas, Tabellen, Spalten, Rechen-Cluster, Benutzer, Rollen, Aufgaben, Freigaben und andere Objekte. Tags werden über die Objekthierarchie vererbt und können zum Auffinden, Verfolgen, Einschränken, Überwachen und Prüfen von Objekten auf der Grundlage von benutzerdefinierten Semantiken verwendet werden. Darüber hinaus können Benutzer:innen mithilfe von Tag-basierten Richtlinien⁵ eine Zugriffsbeschränkung mit einem Tag verknüpfen, sodass die Richtlinie automatisch auf jedes passende Datenobjekt angewendet wird, das mit dem entsprechenden Tag versehen ist.

In Snowflake kann die Definition der meisten Governance-Kontrollen wie Tags, Zugriffsrichtlinien oder Maskierungsregeln getrennt von der Anwendung dieser Kontrollen auf Datenobjekte definiert werden. Dies ermöglicht es den Domänenbesitzern, sich auf gemeinsame Tags oder Richtlinien für alle Domänen zu einigen, während deren Durchsetzung oder Erweiterung jeder einzelnen Domäne überlassen bleibt. Darüber hinaus können sichere Ansichten und Data-Clean-Room-Funktionen für Analysen sensibler Daten verwendet werden, die sonst nicht freigegeben werden könnten.

Aus der Perspektive der Produktnutzung werden Metriken wie Telemetrie- und Verbrauchsdaten gesammelt, die für eine Wirkungsanalyse verwendet werden können. So können die Domänenteams verfolgen, wie und wie oft ihre Datenprodukte von verschiedenen Nutzern verwendet werden.

Snowflake bietet ein nahezu Self-Service-Erlebnis

Einige der häufigsten Gründe für unsere Kunden, sich für Snowflake zu entscheiden, sind die Benutzerfreundlichkeit und die Tatsache, dass es nahezu keinen Wartungsaufwand gibt. Das sind entscheidende Eigenschaften für eine Self-Service-Plattform. So können Benutzer:innen zum Beispiel ganz einfach ihre eigenen Rechen-Cluster instanzieren und skalieren, ohne Unterstützung durch ein IT-Infrastrukturteam. Das Klonen von Entwicklungs- und Testumgebungen ist genauso einfach. Ein Mechanismus zur Erfassung von Änderungsdaten kann mit einer 1-zeiligen SQL DDL-Anweisung eingerichtet werden. Dieser Fokus auf Benutzerfreundlichkeit war ein Leitprinzip für alle Funktionen der Snowflake-Plattform.

⁵ Zum Zeitpunkt der Veröffentlichung, August 2022, in Public Preview.

DATENPRODUKTE IN SNOWFLAKE

Jede Domäne in einem Data Mesh erstellt, verwaltet und besitzt ein oder mehrere Datenprodukte, die mit anderen Domänen und Data Consumers gemeinsam genutzt werden. Die Behandlung von Daten als Produkt erfordert vor allem eine produktorientierte Denkweise, die zu einer organisatorischen Gewohnheit werden muss. Darüber hinaus benötigen Domänen geeignete Self-Service-Tools, welche die Erstellung und Verwaltung von Datenprodukten unterstützen. Im Folgenden erfahren Sie, wie Snowflake Ihnen helfen kann, das Konzept der Daten als Produkt umzusetzen.

Ein Datenprodukt ist definiert als die Kombination aus Daten plus Metadaten, Code und Infrastrukturabhängigkeiten.

- **Daten:** In Snowflake können die Daten eines Datenprodukts in verschiedenen Formen vorliegen, z. B. als Tabellen, Ansichten, Dateien (JSON, XML, Parquet, Avro, CSV usw.) oder externe Tabellen, die als Ansichten über Dateien außerhalb von Snowflake fungieren. Ein einzelnes Datenprodukt kann aus mehreren solcher Objekte bestehen. Eine typische Praxis für Domänen ist die Verwendung eines Schemas pro Datenprodukt zur Gruppierung der Datenobjekte und optional auch des Codes für jedes Datenprodukt. Datenproduzenten können die Daten so modellieren, wie es für die Bedürfnisse der Data Consumers am besten geeignet ist.
- **Metadaten:** Zu den Metadaten eines Datenprodukts gehören die technischen Metadaten seiner Datenobjekte, wie Tabellennamen, Spaltennamen, Datentypen oder Dateiformatdefinitionen. Zu den Metadaten gehören auch Objektabhängigkeiten, die Data Lineage und die Zugriffshistorie. Jedes Objekt kann auch mit Tags versehen werden, bei denen es sich um Schlüssel-Wert-Paare handelt, die beliebige Metadaten wie Datenherkunft, Domäne, Empfindlichkeit, Geschäftsbegriffe, Taxonomie, Kostenstelle oder andere benutzerdefinierte Attribute ausdrücken.

Nach der Veröffentlichung eines Datenprodukts auf dem Snowflake Marketplace wird der Datenproduzent aufgefordert, Unterlagen wie eine Produktbeschreibung, den Geschäftsbedarf, Beispiele, die Nutzungsbedingungen und einen Link zum Support für das Datenprodukt bereitzustellen. Der Produzent wird außerdem aufgefordert, SLOs für Datenprodukte anzugeben, z. B. die Aktualisierungshäufigkeit, die Historie, die zeitliche Granularität der Daten und andere Eigenschaften (siehe Abbildung 3).

- **Code:** Der Code eines Datenprodukts umfasst die Pipelines und Transformationen, mit denen ein Datenprodukt erstellt und aktualisiert wird. In Snowflake kann dieser Code Snowflake Tasks, Pipes, Streams, Stored Procedures⁶, benutzerdefinierte Funktionen usw. umfassen, die allesamt Snowflake-Objekte sind, die in einem Schema pro Datenprodukt gruppiert werden können. Der Code in diesen Objekten kann in SQL, Java, Javascript, Scala oder Python geschrieben sein und läuft nativ auf der Snowflake-Plattform.

Der Code kann auch Richtlinien enthalten. In Snowflake kann es sich dabei um Code für rollenbasierte Zugriffskontrolle, dynamische Richtlinien zur Datenmaskierung, Richtlinien zur Zugriffskontrolle auf Zeilenebene, sichere Ansichten, Object Tagging oder Code zur Klassifizierung oder Anonymisierung/Tokenisierung der Daten handeln.

- **Infrastrukturabhängigkeiten:** So kann beispielsweise ein Snowflake Task, der die Pipeline zur Aktualisierung eines Datenprodukts plant und orchestriert, einen bestimmten Rechen-Cluster für den Auftrag angeben. Dabei kann es sich um eine dedizierte Rechenressource für nur ein Datenprodukt oder um eine gemeinsame Nutzung durch mehrere Datenprodukte handeln. In jedem Fall kann der Cluster bei Bedarf automatisch angehalten und wieder fortgesetzt werden, um nur dann Kosten zu verursachen, wenn Arbeit geleistet wird. Außerdem lassen sich Cluster im Self-Service-Verfahren hoch- und runterskalieren. Tasks, Pipes und andere Operationen können ebenfalls serverlos ablaufen, um die Notwendigkeit expliziter Infrastrukturabhängigkeiten zu reduzieren oder zu beseitigen.

ABBILDUNG 3: FESTLEGEN VON DATENPRODUKT-SLOS FÜR EIN DATENPRODUKT-LISTING

⁶ Zum Zeitpunkt der Veröffentlichung, August 2022, in Public Preview.

Snowflake unterstützt eine Vielzahl von Ein- und Ausgabeports für Datenprodukte, darunter Streaming Ingestion, einen Kafka Connector, einen Spark Connector, eine Dataframe-API, automatische Dateningestion aus Cloud Storage Buckets, eine REST-API, Dateiformate und natürlich SQL-APIs wie JDBC, ODBC, .NET und APIs für viele beliebte Programmiersprachen. Die Kollaborationsfunktionen von Snowflake können auch für den sicheren Zugriff

auf und die nahtlose Bereitstellung von Daten, Datenservices und Anwendungen über Clouds hinweg genutzt werden, ohne dass zusätzliche ETL-Pipelines oder Integrationen erforderlich sind.

Datenprodukte sollten außerdem eine Reihe wichtiger Eigenschaften aufweisen. Tabelle 1 enthält einige Beispiele für Snowflake-Funktionen, mit denen Sie diese Eigenschaften erreichen können.

MERKMALE VON DATENPRODUKTEN	BEISPIELE FÜR SNOWFLAKE-FUNKTIONEN (NICHT VOLLSTÄNDIG)
Sicher	Rollenbasierte Zugriffskontrolle, Zugriffsrichtlinien auf Zeilenebene, dynamische Datenmaskierung, Verschlüsselung, Tokenisierung
Auffindbar	Gezielte Suche/Snowflake Marketplace, optionale Integration mit Drittanbieterkatalog
Adressierbar	Snowflake-Datenfreigaben, standardisierter Zugriff über mehrere Clouds und Regionen hinweg
Verständlich	Benutzerdefinierte Metadaten-Tags, Datenaufstellungen mit Dokumentation, statistische Form der Daten in Snowsight
Vertrauenswürdig	SLOs/SLAs wie Aktualisierungshäufigkeit oder Granularität, Data Lineage, Objektabhängigkeiten, Zugriffshistorie
Nativ zugänglich	SQL, Python, Java, Scala, SQL-APIs, REST-API, Dataframes usw., um auf Daten mit mehreren Modellen zuzugreifen (strukturiert, semistrukturiert, unstrukturiert, verschiedene Dateitypen usw.)
Interoperabel	ANSI SQL-Datentypen, einheitliche Metadaten und gemeinsame APIs über Domänen hinweg, Snowflake Collaboration, Data Sharing, Marketplace, Data Exchange
Für sich genommen wertvoll	Zusammengesetzte Datenprodukte, die aus mehreren Objekten bestehen, Datenprodukte können aus Datenobjekten und Funktionen bestehen, die mit Datenproduktnutzern gemeinsam verwendet werden können

TABELLE 1: IN SNOWFLAKE UNTERSTÜTZTE DATENPRODUKTEIGENSCHAFTEN

ARCHITEKTUROPTIONEN FÜR VERTEILTE DOMÄNEN

Kommen wir nun zu den verschiedenen Snowflake-Topologien, die Unternehmen als Plattform zur Unterstützung verteilter Domänen gewählt haben. Bei diesen Topologien handelt es sich um allgemeine Muster, und die tatsächliche Implementierung kann je nach spezifischen Anforderungen und Präferenzen variieren.

- **„Konto pro Domäne“:** Jede Domain verwendet ein separates Snowflake-Konto.
 - Maximale Isolierung zwischen Domänen.
 - Verschiedene Domänen können in verschiedenen Cloud-Regionen und Cloud-Plattformen betrieben werden.
 - Ermöglicht ein regionen- und cloudübergreifendes Data Mesh mit einem konsistenten Snowflake-Erlebnis und integrierten Funktionen für die gemeinsame Nutzung von Daten zwischen Domänen, basierend auf einem zentralen Austausch von Metadaten, bei dem alle Domänen Datenprodukte veröffentlichen und darauf zugreifen können.
- **„Datenbank pro Domäne“:** Jede Domäne verwendet eine oder mehrere separate Snowflake-Datenbanken.
 - Alle diese Datenbanken werden über ein einziges Snowflake-Konto verwaltet.
 - Vereinfachte Verwaltung von Benutzern, Sicherheit und Governance über Domänen hinweg.
 - Der Zugriff auf Datenprodukte kann ganz einfach durch das Festlegen von Berechtigungen auf Objektebene für alle Datenbanken gewährt werden.
 - Jedes Domänenteam kann nach wie vor eigene Rechen-Cluster aufsetzen und skalieren, unabhängig von anderen Domänen.
- **„Schema pro Domäne“:** Jede Domäne verwendet separate Schemas in einer einzigen Datenbank.
 - Geringster Isolierungsgrad zwischen Domänenumgebungen.
 - Jedes Domänenteam kann nach wie vor eigene Rechen-Cluster aufsetzen und skalieren, isoliert von anderen Domänen.
 - Potenziell höherer Aufwand bei den Namenskonventionen zur Unterscheidung von Objekten aus verschiedenen Domänen.
 - Kann für Subdomänen in einem Domäne-/Subdomäne-Szenario nützlich sein.

- Im weiteren Verlauf dieses Artikels werden wir die Option „Schema pro Domäne“ nicht näher erläutern, aber sie hat viele Ähnlichkeiten mit „Datenbank pro Domäne“.

- **„Heterogene Domänen“:** Domänen können verschiedene IT-Stacks verwenden.
 - Einige Domänen verwenden Snowflake und einige andere Systeme.
 - Einige Domänen befinden sich in der Cloud und einige können möglicherweise in lokalen Rechenzentren angesiedelt sein.
 - Erfordert in der Regel eine höhere Komplexität, um heterogene Domänenumgebungen zu ermöglichen.
 - Sollte besonders gut abgewägt werden, da es dem Ziel eines Data Mesh, eine gemeinsame Domänen-agnostische Self-Service-Plattform für alle Domänen zu verwenden, zuwiderlaufen kann.

Weitere Architekturvarianten oder hybride, von diesen Basistypen abgeleitete Ansätze sind möglich und plausibel. Ein Unternehmen könnte zum Beispiel „Datenbank pro Domäne“ wählen und diese Datenbanken in mehreren Konten statt in einem einzigen Konto haben. Außerdem verwenden einige Domänen möglicherweise separate Datenbanken, während andere ein ganzes Snowflake-Konto nutzen. Außerdem besteht die Umgebung, die ein Domänenteam verwendet, oft aus Snowflake und zusätzlichen Tools, die auf die jeweiligen Anforderungen und Kompetenzen abgestimmt sind.

Der springende Punkt ist, dass Snowflake mehrere Architekturen unterstützt, die unterschiedliche Kompromisse zwischen Domänenautonomie und Dezentralisierung einerseits und verschiedenen Graden von Komplexität und Betriebsmanagement andererseits ermöglichen.

Jedes Unternehmen muss das richtige Gleichgewicht zwischen Zentralisierung und Dezentralisierung finden, das für die Größe, die Tradition und die Unternehmenskultur am besten geeignet ist. Das Gleiche gilt für die föderierte Verwaltung, bei der Unternehmen das richtige Gleichgewicht zwischen zentraler Kontrolle und lokaler Autonomie der Domänen wählen müssen, das für sie am besten funktioniert.

In den folgenden Abschnitten werden diese Architekturoptionen näher erläutert. Der Schwerpunkt dieser Diskussion liegt hauptsächlich auf den Snowflake-Topologien und weniger auf der Integration mit Tools von Drittanbietern, die unsere Kunden häufig zusammen mit Snowflake für ihre Data-Mesh-Initiativen verwenden.

Einzelkonto: Datenbank pro Domäne

Eine beliebte Topologie, die viele unserer Data-Mesh-Kunden einsetzen, verwendet ein einziges Snowflake-Konto, in dem Domänen separate Datenbanken und separate Rechen-Cluster als autonome Umgebungen betreiben. Jeder Domäne können eine oder mehrere Datenbanken und Cluster für ihre Entwicklungs-, Test- und Produktionsanforderungen zugewiesen werden. Dank des Self-Service-Charakters der Plattform können Domänen die Zero-Copy Cloning-Funktionen von Snowflake nutzen, um Entwicklungs- und Testumgebungen sofort und häufig (neu) zu erstellen. Darüber hinaus können verschiedene Benutzer:innen innerhalb einer Domäne ihre eigenen Rechen-Cluster für ihre jeweiligen Bedürfnisse aufstellen und skalieren – und zwar im Self-Service-Verfahren. Dennoch können Kosten- und Verbrauchsmonitore sowie Quoten für Domänen oder andere Granularitätsebenen in der Benutzer- und Ressourcenhierarchie konfiguriert werden.

Da alle Domänen die Snowflake Data Cloud nutzen, können sie ihre eigenen Umgebungen und Rechenressourcen nutzen, ohne physische Silos zu bilden, die den Zugriff auf Datenprodukte erschweren würden.

Ein gewisses Maß an Governance wird zentral entschieden und mit einem DevOps-Prozess auf alle Datenbanken angewendet. Das kann durch Funktionen wie Object Tags erleichtert werden, um einen einfachen Überblick über die verschiedenen Objekte zu behalten, die den Domänen gehören. Innerhalb der Domänen wird die Governance von den Domänenteams kontrolliert, die eine rollenbasierte Zugriffskontrolle sowie Richtlinien für den Zugriff auf Zeilen- und Spaltenebene anwenden, um Daten abzusichern und Benutzer und Domänen vom unerwünschten Zugriff auf bestimmte Daten auszuschließen.

DRITTANBIETERKATALOG/SNOWFLAKE DATA MARKETPLACE

Inventar gemeinsamer Datenprodukte

METADATEN

SUCHE

ZUGRIFFSANFRAGE-WORKFLOW

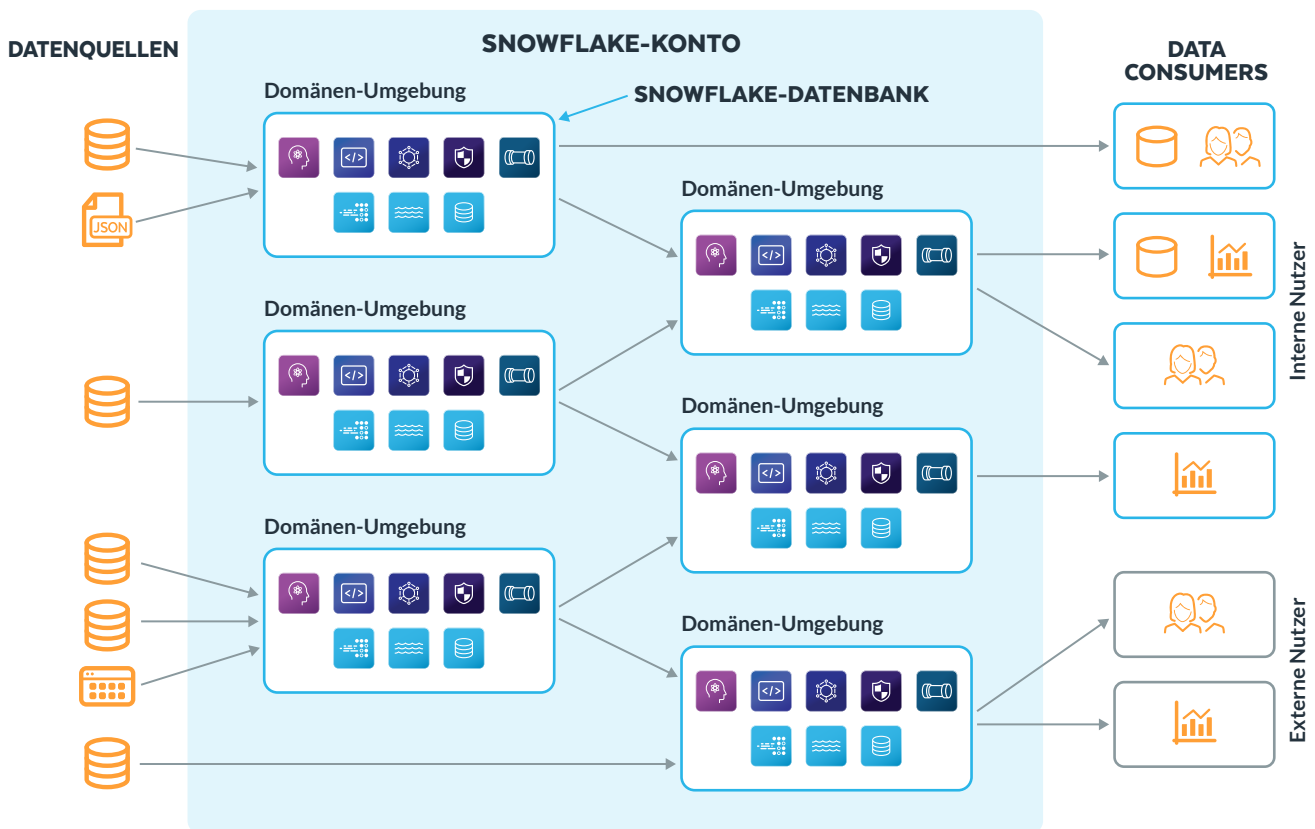


ABBILDUNG 4: EINZELNES SNOWFLAKE-KONTO - DATENBANK PRO DOMÄNE

Jede Domäne kann mehrere Schemas haben, wobei eines als Schicht dient, um Produkte für andere Domänen verfügbar zu machen. Ein anderer Ansatz ist die Verwendung einer gemeinsam Freigabe-Datenbank, in der jede Domäne ein Schema hat, um ihre Datenprodukte als Ansichten zu veröffentlichen (keine Kopien). Bei diesen Produkten kann es sich um strukturierte, halbstrukturierte oder unstrukturierte Daten handeln, je nachdem, was das Produkt umfasst. Die Produkte werden dann in einem Datenkatalog eines Drittanbieters aufgeführt, damit sie auffindbar sind.

Für die Beantragung des Zugriffs auf ein Produkt haben wir mehrere Möglichkeiten gesehen, z. B. einen manuellen Ansatz, bei dem der Antragsteller ein Ticket eröffnen muss, das dann vom Domänenteam bearbeitet und der Zugriff durch Zuweisung einer Zugriffsrolle an den Antragsteller erlaubt oder verweigert wird. Einige Kataloge bieten einen automatischeren Ablauf.

Alle Domänen in einem einzigen Snowflake-Konto zu haben, bietet folgende Vorteile:

- Der Zugriff auf Datenprodukte kann ganz einfach durch das Festlegen von datenbankinternen Berechtigungen erfolgen.
- Die zentralisierte Verwaltung von Netzwerk-, Sicherheits- und Governance-Richtlinien vereinfacht die gesamte Verwaltung.
- Disaster Recovery ist einfacher, da nur ein anderes Konto in einer anderen Region oder Cloud zur Unterstützung benötigt wird.

Die Namenskonventionen sollten sorgfältig geplant werden, da es viele Objekte geben kann, wenn man bedenkt, dass jede Domäne DT(A)P-Umgebungen (Development, Test, Acceptance, Production) benötigt, die sie mit Zero-Copy Cloning leicht erstellen können.

Ein ähnlicher Ansatz ist die Verwendung eines Schemas pro Domäne. Die Konsequenzen dieses Ansatzes sind ähnlich wie beim Ansatz Datenbank pro Domäne, da alles logisch innerhalb eines Snowflake-Kontos organisiert ist. Beachten Sie jedoch, dass es einfacher ist, eine Datenbank pro Domäne zu haben, wenn es darum geht, Daten mit externen Nutzern öffentlich über Snowflake Marketplace oder privat mithilfe von Listing Discovery Controls zu teilen.

Mehrere Konten: Konto pro Domäne

Eine andere mögliche Topologie bietet die Möglichkeit, dass jede Domäne in einem separaten Snowflake-Konto arbeitet. Diese Konten können sich in der gleichen oder in verschiedenen Cloud-Regionen und Cloud-Plattformen befinden. Die globale Snowflake Data Cloud ermöglicht es Unternehmen und Domänen, Daten über Konten, Regionen und Cloud-Plattformen hinweg gemeinsam zu nutzen und auf sichere und geregelte Weise standardisierten Zugriff auf die Datenprodukte der jeweils anderen zu erhalten. Einige unserer Kunden nutzen diese Fähigkeit, um ein Data Mesh für mehrere Regionen und mehrere Clouds zu unterstützen.

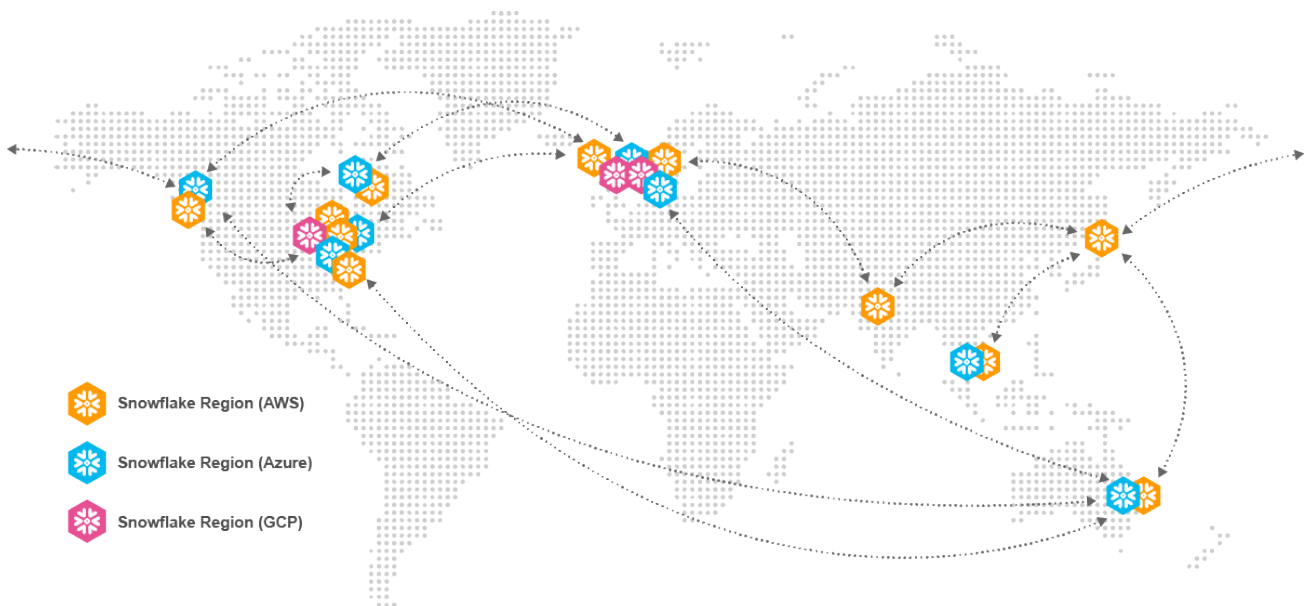


ABBILDUNG 5: DAS SNOWGRID ALS GLOBALE DATA-CLOUD-PLATTFORM

Es gibt verschiedene Gründe, warum sich Unternehmen für diese Topologie entscheiden. Ein Unternehmen kann zum Beispiel global verteilt arbeiten, wobei die verschiedenen Domänen natürlich mit den verschiedenen Standorten und Regionen der Welt übereinstimmen können. Einige Unternehmen sind möglicherweise weltweit tätig und müssen die Anforderungen an die Datenlokalisierung einhalten (z. B. ein internationales Unternehmen, bei dem ein Teil der Daten Europa nicht ohne Anonymisierung, Maskierung oder andere Maßnahmen zur Einhaltung der Datenschutzbestimmungen verlassen darf).

Ein weiterer häufiger Grund sind Fusionen und Übernahmen, die ein Unternehmen dazu zwingen können, Daten über Regionen oder Cloud-Plattformen hinweg auszutauschen. Einige Unternehmen verfolgen bewusst eine Multi-Cloud-Strategie zur Diversifizierung oder um Präferenzen und vorhandene Investitionen zu berücksichtigen, die verschiedene Geschäftsbereiche bereits getätigt haben. Durch die Verwendung separater Konten wird auch eine größere Autonomie der Domänen erreicht (z. B. wenn eine separate Benutzer- und Sicherheitsverwaltung für jede Domäne erforderlich ist).

Die sich daraus ergebende Topologie (Abbildung 6) ist logisch gesehen der Verwendung einer separaten Datenbank pro Domäne sehr ähnlich, mit dem Unterschied, dass jede Domäne nun ein separates Snowflake-Konto „besitzt“ und die Snowflake-Funktionen für die gemeinsame Nutzung von Daten und den Marketplace nutzt, um Datenprodukte für andere zugänglich zu machen.

Folgende Vorteile können sich gegenüber dem Ansatz der Datenbank pro Domäne ergeben:

- Data Sharing und Kollaborationsfunktionen können domänenübergreifend genutzt werden.
- Globale Benennungsstandards können einfacher angewendet werden, da jedes Konto ein unabhängiger Namespace ist.
- Cloud-Plattform und regionale Präferenzen werden unterstützt.
- Für jedes Konto gibt es eine separate Sicherheits- und Benutzerverwaltung.

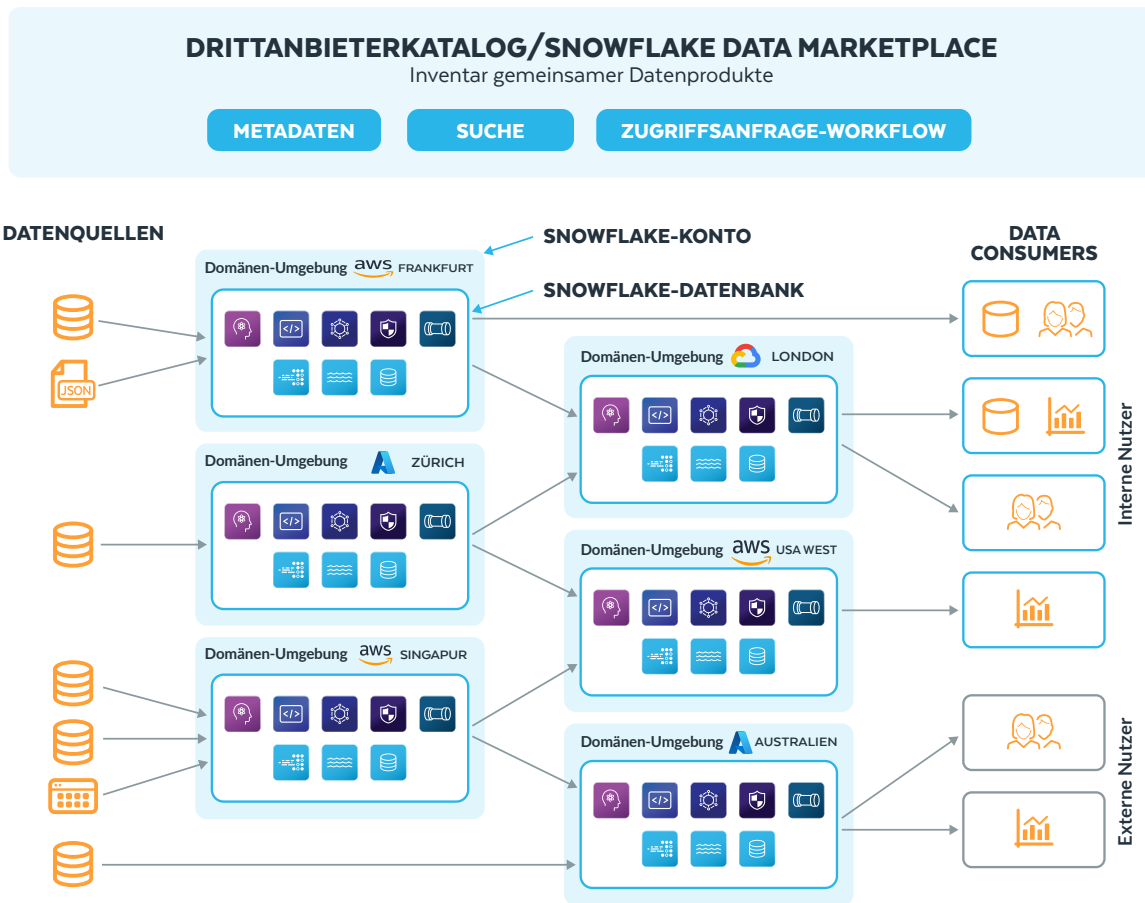


ABBILDUNG 6: MEHRERE KONTEN - EIN KONTO PRO DOMÄNE

Heterogene Architektur

Einige Kunden haben uns gefragt, wie sie andere Domänen-Umgebungen, die nicht zu Snowflake gehören, in die oben beschriebenen Topologien integrieren können. Solche Integrationen führen zu einer heterogenen Architektur, bei der nicht alle Domänen dieselbe domänenagnostische Datenplattform zur Implementierung ihrer Pipelines und Datenprodukte verwenden. Oft wird dies durch den Wunsch motiviert, mehrere unterschiedliche Repositorys oder Technologie-Stacks wiederzuverwenden, die bereits in verschiedenen Teilen des Unternehmens vorhanden sind.

Wir haben festgestellt, dass eine solche heterogene Architektur in der Regel die Kosten und die Komplexität einer Data-Mesh-Anschaffung in die Höhe treibt. Der Grund dafür ist, dass eine größere Heterogenität der teilnehmenden Systeme es schwieriger macht, die Einheitlichkeit von Governance, Sicherheit, Metadaten, Interoperabilitätsstandards, Leistung, erforderlichen Fähigkeiten, IT-Support und anderen kritischen Bereichen zu gewährleisten. Daher empfehlen wir unseren Kunden, die *Rolle* der verschiedenen Systeme und Repositorys, die sie in ein Data Mesh integrieren möchten, sorgfältig zu prüfen. Werden diese Systeme

wirklich als Domänenumgebungen verwendet, die Datenprodukte erstellen und bereitstellen? Oder sollten sie eher als Datenquellen betrachtet werden, die Domänen als Input nutzen? Im letzteren Fall können die Kunden oft auf die oben beschriebenen Topologien zurückgreifen.

Ein Ansatz für die Integration von Implementierungen anderer Domänen als Snowflake besteht darin, dass sie Datenbestände oder „fast fertige Datenprodukte“ in eine Zwischenschicht verschieben, für die Snowflake als „Proxy“ fungieren kann, der die Datenprodukte mit einheitlicher Governance, Sicherheit, Interoperabilität usw. für den Rest des Data Mesh bereitstellt.

Bei dieser Zwischenschicht könnte es sich beispielsweise um Kafka-Themen handeln, die über kontinuierliche Ingestion in Snowflake eingespeist werden, gefolgt von automatischen Aktualisierungen der Datenprodukte in Snowflake. Die Zwischenschicht könnte auch ein oder mehrere Cloud Storage Buckets in Amazon S3, Azure Blob Storage, Azure Data Lake Storage oder Google Cloud Storage sein. Als Datenformate sind JSON, XML, Parquet, AVRO, Apache Iceberg, Delta Lake und andere möglich. Snowflake kann entweder

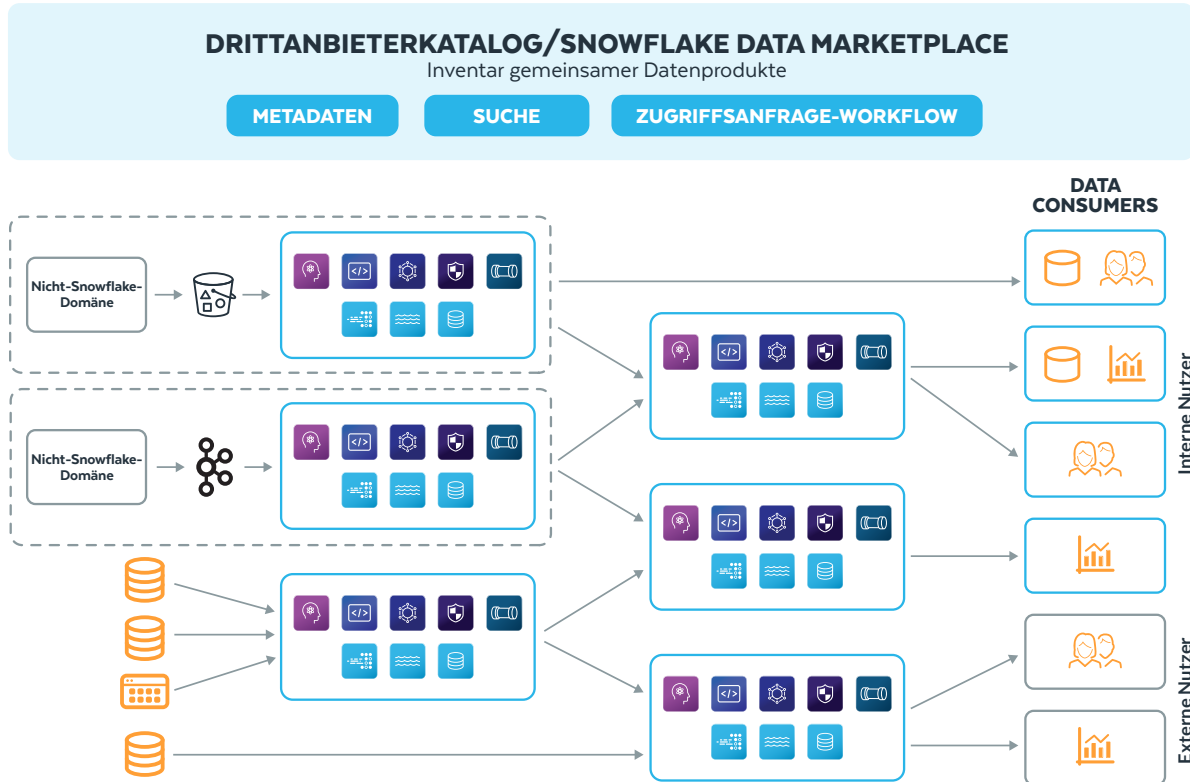


ABBILDUNG 7: HETEROGENE ARCHITEKTUR

kontinuierlich neue Dateien aus Storage Buckets automatisch testen, um die beste Leistung, Sicherheit und automatische Verwaltung zu gewährleisten, oder den Lesezugriff auf solche Dateien als externe Tabellen für den Rest des Data Mesh freigeben. Die externen Tabellen von Snowflake sind im Wesentlichen Ansichten von Daten, die sich in Dateien außerhalb von Snowflake befinden. Dennoch sind externe Tabellen erstklassige Datenobjekte in Snowflake, die wie andere Datenobjekte in Snowflake abgesichert und verwaltet, verbunden und sogar über Snowflake Collaboration gemeinsam genutzt werden können. Dadurch kann Snowflake als Integrationsschicht fungieren, die externe Daten auf einheitliche und geregelte Weise offenlegen kann, ohne dass die Daten zwangsläufig aufgenommen und dupliziert werden müssen.

Verschiedene andere Optionen für die Integration von Nicht-Snowflake-Umgebungen in die Datenplattform sind vorhanden, liegen aber außerhalb des Rahmens dieses Dokuments.

Einige Unternehmen betrachten die Datenvirtualisierung als potenzielle Lösung für die Integration verschiedener Domänen-Umgebungen. Es gibt zwar durchaus sinnvolle Anwendungsfälle für die Datenvirtualisierung, aber wir haben festgestellt, dass

sie auch eine Reihe von Herausforderungen mit sich bringt. Eine Herausforderung ist die Performance, wenn Daten aus mehreren verschiedenen Repositories zusammengeführt werden müssen. Dies erfordert in der Regel eine Datenverschiebung, um die Daten für die Berechnung der Zusammenführung an einen gemeinsamen Ort zu bringen, auch wenn andere Prädikate zu den Datenquellen heruntergeladen werden können.

Dies kann die Virtualisierung für performanceabhängige Anwendungsfälle verhindern. Mit Snowflake hat eine Join-Zusammenführung zwischen mehreren Datenobjekten in einer Datenbank annähernd die gleichen Performanceeigenschaften, wie wenn sich diese Datenobjekte in separaten Datenbanken oder sogar in separaten Snowflake-Konten befinden, was eine bedeutende Eigenschaft der Snowflake-Plattformarchitektur ist. Eine weitere Herausforderung, die wir bei der Virtualisierung in einigen Unternehmen gesehen haben, besteht darin, dass sie Teams dazu ermutigen kann, weiterhin auf ihren jeweiligen Technologieinseln zu arbeiten, die oft sehr domänenspezifisch sind, anstatt eine gemeinsame und domänenunabhängige Self-Service-Plattform anzustreben.

ZUSAMMENFASSUNG

Das Data Mesh ist kein Allheilmittel für alle Herausforderungen der Datenverwaltung und Datenintegration. Wenn Sie jedoch feststellen, dass das Data Mesh der richtige Ansatz für Ihr Unternehmen ist, sollten Sie sich auf die organisatorischen und nicht-technischen Fragen konzentrieren, die für den Erfolg unerlässlich sind. Einige Beispiele hierfür sind organisatorische Veränderungen, Rollen und Verantwortlichkeiten, Personalausstattung, Anreize und Verantwortlichkeit, die Zustimmung der wichtigsten Interessengruppen oder die Umstellung auf das Produktdenken.

Schließlich müssen Sie eine Self-Service-IT-Architektur entwerfen, die verteilte Domänen und Datenprodukte mit föderierter Governance unterstützen kann. Snowflake kann diese Schlüsselrolle als einfach zu bedienende Self-Service-Plattform für Domänenteams spielen. Snowflake unterstützt verschiedene Topologien, mit denen Unternehmen den gewünschten Grad an Dezentralisierung und Domänenautonomie wählen können, während gleichzeitig sichergestellt wird, dass die Domänen miteinander verbunden und interoperabel bleiben. Snowflake ermöglicht sowohl Single-Account-Topologien als auch Multi-Region- und Multi-Cloud-Architekturen und unterstützt die Integration von externen Domänen oder unternehmensübergreifenden Kollaborations-Setups. Die zugrundeliegende Snowflake-Plattform mit dem globalen Snowgrid und dem Snowflake Marketplace fungiert als Bindeglied, das Unternehmen dabei hilft, das Risiko der Bildung von Datensilos zu vermeiden.

Darüber hinaus bietet Snowflake eine breite Palette von Funktionen, damit Unternehmen die Konzepte von Daten als Produkt und föderierter Governance umsetzen können. Snowflake lässt sich außerdem problemlos in eine Vielzahl von Drittanbieter-Tools integrieren. Dadurch können zusätzliche Plattformfunktionen bereitgestellt werden. Als Ergänzung zu den organisatorischen Veränderungen und Prozessen, die für eine erfolgreiche Data-Mesh-Umwandlung erforderlich sind, ist die Snowflake Data Cloud eine hervorragende technologische Wahl. Weitere Informationen über die Möglichkeiten von Snowflake finden Sie unter snowflake.com/data-mesh

ÜBER SNOWFLAKE

Die Snowflake Data Cloud bietet jedem Unternehmen die Möglichkeit, seine Daten zu mobilisieren. Mithilfe der Data Cloud können Kunden Datensilos zusammenführen, Daten entdecken und sicher freigeben, Datenapplikationen unterstützen sowie verschiedene KI/ML- und analytische Workloads ausführen. Wo auch immer sich Daten oder Benutzer befinden, Snowflake bietet eine einheitliche Datenlösung, die sich über mehrere Clouds und geografische Regionen erstreckt. Tausende von Kunden in zahlreichen Branchen nutzen die Snowflake Data Cloud und bringen so ihre Unternehmen voran.

Darunter fallen auch 590 Unternehmen der Forbes Global 2000 (G2K) aus dem Jahr 2022 (Stand: 30. April 2023). Erfahren Sie mehr unter [snowflake.com](https://www.snowflake.com).



© 2022 Snowflake Inc. Alle Rechte vorbehalten. Snowflake, das Logo von Snowflake und alle sonstigen hier erwähnten Namen von Produkten, Funktionen und Services von Snowflake sind eingetragene Marken oder Marken von Snowflake Inc. in den USA und anderen Ländern. Alle anderen erwähnten oder verwendeten Markennamen oder Logos dienen ausschließlich der Identifikation und können die Marken ihrer jeweiligen Eigentümer sein. Snowflake darf nicht mit diesen Eigentümern in Verbindung gebracht oder von diesen unterstützt oder gefördert werden.