



CÓMO TEJER SU DATA MESH EN SNOWFLAKE

En los últimos años, la data mesh¹ se ha convertido en un enfoque de gestión de datos cada vez más popular. Diversas empresas de todos los sectores eligen la data mesh para descentralizar la gestión de datos con el fin usarlos de una forma más ágil y evitar los cuellos de botella organizativos, que suelen estar relacionados con enfoques centralizados y monolíticos.

En este documento se analiza el enfoque de data mesh de Snowflake. En él se describen algunas de las funcionalidades de Snowflake más importantes que debe tener una data mesh y se presentan las opciones de arquitectura típicas que nuestros clientes han elegido para implementar una plataforma de datos de autoservicio que respalde los dominios distribuidos.

El enfoque de la data mesh es principalmente organizativo y define las responsabilidades y la coordinación entre equipos de dominios independientes y sus productos de datos. Sin embargo, se necesita la tecnología adecuada para que los dominios adopten el concepto de data mesh de una manera factible.

“ La data mesh no es solo tecnología [...]; pero se necesita la tecnología adecuada para dotar de una variedad de funcionalidades a los equipos de productos de datos. Esos equipos de productos de datos y dominios no tienen que reinventar la rueda, empezar desde cero ni crear su propia plataforma de datos y análisis. Tenemos que simplificarles la vida a los equipos de productos de datos. Si no se hace, no hay capacitación ni descentralización.”

OMAR KHAWAJA, Global Head BI, Roche (2022)

Muchas empresas que aplican un enfoque de data mesh están utilizando con éxito Snowflake como plataforma de datos. No existe ninguna plataforma tecnológica que proporcione una solución completa e integral para respaldar el concepto de data mesh. Sin embargo, Snowflake proporciona muchas de las funciones que debe tener una plataforma de datos de autoservicio, que hacen posible una arquitectura distribuida basada en dominios y ofrecen capacidades que ayudan a implementar los datos como producto y la gobernanza computacional federada.

EL ENFOQUE DE DATA MESH DE SNOWFLAKE

Después de trabajar con numerosos clientes en sus iniciativas de data mesh, Snowflake ha adoptado el siguiente enfoque:

- Somos conscientes de que adoptar la data mesh es, ante todo, una transformación organizativa. Dicha transformación tiene muchas implicaciones que no son técnicas, pero también suele requerir cambios en términos de tecnología y arquitectura de TI.
- Se debe ser pragmático. Aconsejamos a nuestros clientes que no intenten implementar la data mesh “perfecta”, sino que aborden sus puntos débiles y objetivos específicos, y los usen como guía. Por ejemplo, el almacenamiento políglobo y el acceso multimodal son conceptos útiles, pero las empresas deben centrarse en sus requisitos reales para maximizar el impacto.
- Se debe comenzar poco a poco, ampliar gradualmente e ir avanzando por la curva de madurez de la data mesh a lo largo del tiempo. Por ejemplo, puede empezar con uno o dos dominios y productos de datos para satisfacer una necesidad empresarial inmediata y, después, aprovechar el éxito inicial para ampliar la data mesh.
- Hay que tener en cuenta el coste y la complejidad. Por ejemplo, que el conjunto de herramientas de la plataforma de datos de autoservicio sea lo más pequeño y uniforme posible en todos los dominios, siempre que cumpla todos los requisitos críticos de dichos dominios, ha demostrado ser beneficioso.
- Los incentivos y los criterios de éxito se deben definir al principio. Esto incluye indicadores clave de rendimiento cuantificables para los dominios, los productos de datos, la plataforma de datos de autoservicio y los controles de gobernanza.
- No existe ninguna solución de data mesh lista para usar. Aprovechamos nuestra amplia red de partners para crear soluciones conjuntas que satisfagan los requisitos de los clientes. Por ejemplo, las herramientas de gobernanza de datos, automatización, DevOps y otras áreas suelen formar parte de una solución de data mesh, aunque esto no se trata en detalle en este documento.

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

FUNCIONES RELEVANTES DE SNOWFLAKE

Snowflake ofrece funcionalidades clave que a nuestros clientes les han resultado útiles al crear la plataforma de datos de autoservicio para una data mesh.

Snowflake es mucho más que un almacenamiento de datos en la nube

Snowflake es un proveedor de servicios de nube integrado que ofrece una amplia gama de funciones para ingeniería de datos, data lakes, almacenamiento de datos, data sharing y partes significativas del ciclo de vida típico de aprendizaje automático.

En concreto, los usuarios pueden crear y automatizar flujos de transformación de datos para convertir diversos datos de entrada en productos de datos gobernados. Snowflake puede funcionar con los formatos de archivo comunes que tenga en sus contenedores de almacenamiento en la nube con

la misma facilidad que con flujos de entrada (por ejemplo, de Kafka) o tablas relacionales. Los formatos de archivo compatibles son JSON, XML, Parquet, AVRO, Delta Lake² y Apache Iceberg³, entre otros. Además, Snowflake admite datos no estructurados, como imágenes, vídeo u otros formatos binarios. Los datos se pueden gestionar en la plataforma de Snowflake mediante SQL, Python⁴, Scala, Java y JavaScript, o invocando funciones externas en la plataforma de nube general.

Puede que Snowflake proporcione, o no, todas las funciones que necesitan sus equipos de dominio; pero ofrece una amplia gama de prestaciones en un *único* servicio que, de lo contrario, requeriría la integración de *varios* servicios en la nube. Dicha integración puede ser compleja y requerir mucho tiempo, así como personal altamente cualificado.



FIGURA 1: SNOWFLAKE COMO PLATAFORMA ÚNICA PARA TIPOS DE DATOS Y WORKLOADS DIFERENTES

² En vista previa pública en el momento de la publicación, agosto de 2022.

³ En vista previa privada en el momento de la publicación, agosto de 2022.

⁴ En vista previa pública en el momento de la publicación, agosto de 2022.

Snowflake es una plataforma distribuida, no monolítica

Snowflake es una plataforma distribuida, pero interconectada, que evita los silos y permite a los equipos distribuidos intercambiar datos de una manera controlada y segura. ¿Cómo funciona? Una empresa puede crear una o varias cuentas de Snowflake que pueden residir en la misma o en diferentes regiones y plataformas de nube (figura 2). Cada cuenta puede albergar varias bases de datos independientes cuyos recursos de procesamiento y almacenamiento se pueden implementar y escalar por separado de forma distribuida.

Los distintos equipos de dominio pueden trabajar de forma autónoma sirviéndose de una potencia de procesamiento independiente —en bases de datos o incluso en cuentas independientes— y seguir utilizando la plataforma de Snowflake subyacente para compartir activos de datos entre sí. Tenga en cuenta que el concepto de “base de datos” de Snowflake no hace referencia simplemente a una base de datos relacional tradicional, sino que también incluye las demás funciones de Snowflake, como ingeniería de datos, data lake, almacenamiento de datos, data sharing y data science. El uso de clústeres de procesamiento para combinar y procesar los datos de varias bases de datos o cuentas es una función esencial de la plataforma de Snowflake.

Snowflake cuenta con funciones integradas de data sharing y Marketplace

Los productores de datos de Snowflake pueden compartir datos, servicios de datos o aplicaciones con otras cuentas mediante la publicación de metadatos

(“listas”). Mediante los controles de detección en fichas, los productores pueden compartir datos de forma privada con otras cuentas, con grupos de cuentas o públicamente a través de Snowflake Marketplace. Los productores de datos pueden especificar acuerdos de nivel de servicio (service-level agreement, SLA) u objetivos de nivel de servicio (service-level objective, SLO) para los datos que están compartiendo, como la frecuencia de actualización, el volumen del historial, el nivel de detalle temporal de los datos y otras propiedades que ayudan a describir los datos como un producto.

Otros equipos pueden hacer búsquedas para identificar activos de datos disponibles que les resulten relevantes y obtener o solicitar acceso a ellos. Estos consumidores de datos obtienen acceso en tiempo real a los datos compartidos. Dichos datos permanecen bajo el control del productor, que puede personalizar las políticas de acceso o revocar el acceso en cualquier momento. El acceso a los datos compartidos no requiere que el productor ni el consumidor implementen un proceso de extracción, transformación y carga (extract, transform, and load; ETL) o de movimiento de datos. Los productores también pueden publicar y compartir tablas externas, que son “vistas” de archivos que se almacenan fuera de Snowflake y que, opcionalmente, pueden incluir formatos Iceberg y Delta Lake. Los productores pueden, incluso, compartir datos con terceros ajenos a la empresa, aunque dichas partes no sean clientes activos de Snowflake. Por ejemplo, un productor de datos puede compartir datos externamente a través de una cuenta de lectura de Snowflake y toda la gama de API compatibles. O bien puede exportar datos (particionados) periódicamente a un contenedor de almacenamiento en la nube utilizando cualquiera de los formatos de archivo más populares hoy en día.

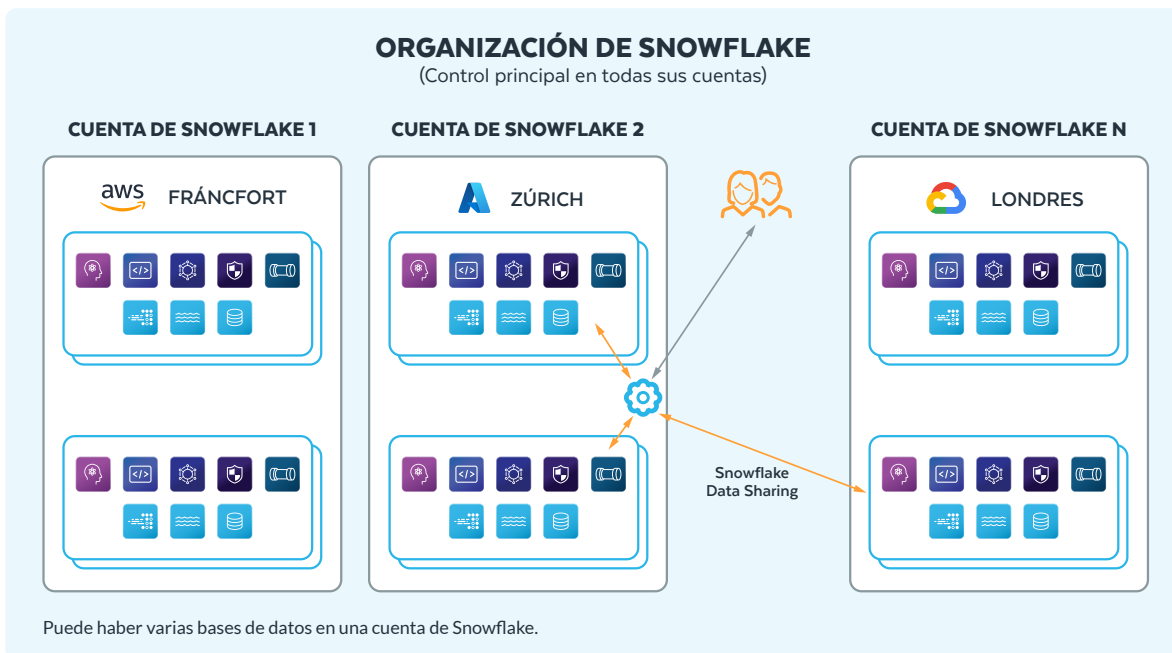


FIGURA 2: LA ORGANIZACIÓN, LAS CUENTAS Y LAS BASES DE DATOS DE SNOWFLAKE ADMITEN UNA ARQUITECTURA DISTRIBUIDA

Snowflake ofrece una amplia gama de funciones de seguridad y gobernanza

La gobernanza federada es una de las cuestiones más difíciles que conlleva la data mesh y, a menudo, debe combinarse con una o varias herramientas para satisfacer todos los requisitos. A nivel de plataforma, Snowflake admite el control de acceso basado en roles, las políticas de acceso a nivel de fila, el enmascaramiento de datos a nivel de columna, la tokenización externa, el linaje de datos, las capacidades de auditoría, etc. Los usuarios también pueden asignar una o varias etiquetas de metadatos (pares clave-valor) a casi cualquier tipo de objeto de Snowflake, como cuentas, bases de datos, esquemas, tablas, columnas, clústeres de procesamiento, usuarios, roles, tareas y recursos compartidos, entre otros. Las etiquetas se heredan a través de la jerarquía de objetos y se pueden aprovechar para detectar, restringir, supervisar y auditar objetos, así como hacer un seguimiento de estos, en función de la semántica definida por el usuario. Además, las políticas de acceso basadas en etiquetas⁵ permiten a los usuarios asociar una restricción de acceso con una etiqueta, de forma que la política de acceso se aplique automáticamente a cualquier objeto de datos que lleve la etiqueta correspondiente.

En Snowflake, la mayoría de los controles de gobernanza (como las etiquetas, las políticas de acceso o las reglas de enmascaramiento) se pueden definir con independencia de la aplicación de estos controles a objetos de datos. Esto permite a los propietarios de dominios usar etiquetas o políticas comunes a todos los dominios y dejar su aplicación o extensión en manos de cada dominio. Además, las vistas seguras y la funcionalidad de data clean room se pueden utilizar para analizar datos confidenciales que, de lo contrario, no se podrían compartir.

Desde el punto de vista del uso del producto, se recopilan métricas, como datos de telemetría y consumo, que se pueden utilizar para analizar el impacto. Esto permite a los equipos de dominio realizar un seguimiento de la forma y la frecuencia con la que distintos consumidores utilizan sus productos de datos.

Snowflake ofrece una experiencia de autoservicio cercana

Algunas de las razones más comunes por las que nuestros clientes eligen Snowflake son la facilidad de uso y el mantenimiento prácticamente nulo. Estas son propiedades fundamentales en una plataforma de autoservicio. Por ejemplo, los usuarios pueden crear instancias y escalar sus propios clústeres de procesamiento fácilmente sin la ayuda de un equipo de infraestructura de TI. La clonación de entornos de desarrollo y pruebas es igual de sencilla. Se puede configurar un mecanismo de captura de datos modificados con una sentencia de lenguaje de definición de datos (data definition language, DDL) SQL de una línea. La priorización de la facilidad de uso ha sido un principio rector para todas las características y funciones de la plataforma de Snowflake.

⁵ En vista previa pública en el momento de la publicación, agosto de 2022.

PRODUCTOS DE DATOS EN SNOWFLAKE

En una data mesh, cada dominio crea, mantiene y posee uno o más productos de datos que se comparten con otros dominios y consumidores de datos. Tratar los datos como un producto requiere, sobre todo, una mentalidad orientada al producto que debe convertirse en un hábito organizativo. Además, los dominios necesitan herramientas de autoservicio adecuadas que respalden la creación y gestión de productos de datos. A continuación, se describe la forma en que Snowflake puede ayudarle a implementar el concepto de datos como producto.

Un producto de datos se define como la combinación de datos más metadatos, código y dependencias de infraestructura.

- **Datos:** en Snowflake, los datos de un producto de datos pueden adoptar varias formas, como tablas, vistas, archivos (JSON, XML, Parquet, Avro, CSV, etc.) o tablas externas que actúan como vistas de archivos fuera de Snowflake. Un único producto de datos puede constar de varios de estos objetos. Una práctica típica para los dominios es utilizar un esquema por producto de datos para agrupar los objetos de datos y, opcionalmente, el código de cada producto de datos. Los productores de datos pueden modelar los datos de la forma que mejor se adapte a las necesidades de los consumidores de datos.
- **Metadatos:** los metadatos de un producto de datos incluyen los metadatos técnicos de sus objetos de datos, como nombres de tablas, nombres de columnas, tipos de datos o definiciones de formato de archivo. Los metadatos también incluyen las dependencias de objetos, el linaje de datos y el historial de acceso. Cada objeto también puede incluir anotaciones mediante etiquetas, que son pares clave-valor que expresan metadatos arbitrarios, como el origen de los datos, el dominio, la confidencialidad, los términos comerciales, la taxonomía, el centro de costes u otros atributos definidos por el usuario.

Cuando se publica un producto de datos en Snowflake Marketplace, se pide al productor que proporcione documentación, como la descripción del producto, la necesidad empresarial, ejemplos, términos de servicio y un enlace a la asistencia técnica para el producto de datos. También se solicita al productor que especifique los SLO del producto de datos, como la frecuencia de actualización, el volumen del historial, el nivel de detalle temporal de los datos y otras propiedades (véase la figura 3).

- **Código:** el código de un producto de datos incluye los flujos y las transformaciones de creación y actualización de dicho producto. En Snowflake, este código puede incluir tareas, canalizaciones, flujos, procedimientos almacenados⁶, funciones definidas por el usuario, etc. Todos estos elementos son objetos de Snowflake que se pueden agrupar en un esquema por producto de datos. El código de estos objetos puede ser SQL, Java, JavaScript, Scala o Python y se ejecuta de forma nativa en la plataforma de Snowflake.

El código también puede incluir políticas. En Snowflake, puede tratarse de código para el control de acceso basado en roles, políticas de enmascaramiento dinámico de datos, políticas de control de acceso a nivel de fila, vistas seguras, etiquetado de objetos, o código para clasificar, anonimizar o tokenizar los datos.

- **Dependencias de la infraestructura:** por ejemplo, una tarea de Snowflake que programa y organiza el flujo para refrescar un producto de datos puede especificar un clúster de procesamiento determinado para el trabajo. Puede ser un recurso de procesamiento dedicado (para un solo producto de datos) o compartido (entre varios productos de datos). En cualquier caso, el clúster se puede suspender y reanudar automáticamente según sea necesario para incurrir en costes solo cuando esté en funcionamiento. Además, los clústeres se pueden ampliar o reducir en régimen de autoservicio. Las tareas, las canalizaciones y otras operaciones también pueden prescindir del servidor para reducir o eliminar la necesidad de dependencias explícitas de la infraestructura.

The screenshot shows a configuration interface for a data product. On the left, a dark blue sidebar contains the heading 'Attributes' and a sub-section 'Preview' with a list of four checked items: 'Data updated continuously', 'Data for the last month with hourly interval', 'data aggregated by device', and 'data integrated from all countries'. The main area is titled 'Core Attributes' and contains four dropdown menus: 'Update Frequency' (Near Real-time), 'Geographic Coverage' (Not applicable), 'Time Range' (Last month), and 'Timestamp Granularity' (Hourly). Below these is an 'Additional Attributes (optional)' section with three text input fields, the first two containing 'data aggregated by device' and 'data integrated from all countries'. At the bottom right are 'Cancel' and 'Save' buttons.

FIGURA 3: ESPECIFICACIÓN DE LOS SLO DE UN PRODUCTO DE DATOS EN LA FICHA DEL PRODUCTO

⁶ En vista previa pública en el momento de la publicación, agosto de 2022.

Snowflake es compatible con una gran variedad de puertos de entrada y salida para productos de datos, como la ingesta de transmisión, un conector Kafka, un conector Spark, una API de Dataframe, la ingesta automática de datos desde contenedores de almacenamiento en la nube, una API REST, formatos de archivo y, por supuesto, las API SQL (como JDBC, ODBC, .NET) y las API de muchos lenguajes de programación populares. Las funciones de colaboración de Snowflake también se pueden

utilizar para acceder de forma segura a datos, servicios de datos y aplicaciones, y distribuirlos de forma fluida en varias nubes sin necesidad de contar con un flujo ETL ni integraciones adicionales.

Los productos de datos también deben presentar una serie de propiedades importantes. En la tabla 1 se muestran algunos ejemplos de funciones de Snowflake que pueden ayudarle a obtener esas características.

CARACTERÍSTICAS DE PRODUCTO DE DATOS	EJEMPLOS DE FUNCIONES DE SNOWFLAKE (NO SE MUESTRAN TODAS)
Seguro	Control de acceso basado en roles, políticas de acceso a nivel de fila, enmascaramiento dinámico de datos, cifrado, tokenización
Detectable	Detección específica, Snowflake Marketplace, integración opcional con catálogos de terceros
Direccionable	Data sharing de Snowflake, acceso estandarizado en diversas nubes y regiones
Comprensible	Etiquetas de metadatos personalizadas, fichas de datos con documentación, forma estadística de los datos en Snowsight
Fiable	SLO/SLA, como la frecuencia o el nivel de detalle de las actualizaciones, el linaje de datos, las dependencias de objetos y el historial de acceso
Accesible de forma nativa	SQL, Python, Java, Scala, API SQL, API REST, Dataframes, etc., para acceder a datos (estructurados, semiestructurados, no estructurados, varios tipos de archivos, etc.) de varios modelos
Interoperable	Tipos de datos SQL ANSI, metadatos unificados y API comunes en todos los dominios, Snowflake Collaboration, Data Sharing, Marketplace, Data Exchange
Valioso en sí mismo	Productos de datos compuestos por varios objetos: los productos de datos pueden constar de objetos de datos junto con funciones que se pueden compartir con los consumidores de productos de datos

TABLA 1: CARACTERÍSTICAS DEL PRODUCTO DE DATOS QUE OFRECE SNOWFLAKE

OPCIONES DE ARQUITECTURA PARA DOMINIOS DISTRIBUIDOS

Veamos ahora las diferentes topologías de Snowflake que las empresas han elegido como plataforma para respaldar dominios distribuidos. Estas topologías son patrones generales, así que las implementaciones reales pueden variar en función de los requisitos y las preferencias específicos.

- **“Cuenta por dominio”**: cada dominio utiliza una cuenta de Snowflake independiente.
 - Aislamiento máximo entre dominios.
 - Diferentes dominios pueden funcionar en diferentes regiones y plataformas de nube.
 - Posibilita una data mesh multinube y de varias regiones con una experiencia de Snowflake coherente y funciones integradas de data sharing entre dominios, que se basan en un intercambio centralizado de metadatos mediante el que todos los dominios pueden publicar productos de datos y obtener acceso a ellos.
- **“Base de datos por dominio”**: cada dominio utiliza una o más bases de datos de Snowflake independientes.
 - Todas estas bases de datos se gestionan en una única cuenta de Snowflake.
 - Gestión simplificada de usuarios, seguridad y gobernanza en todos los dominios.
 - El acceso a los productos de datos se puede otorgar fácilmente estableciendo permisos a nivel de objeto en las bases de datos.
 - Cada equipo de dominio puede seguir poniendo en marcha y escalando sus propios clústeres de procesamiento independientemente de otros dominios.
- **“Esquema por dominio”**: cada dominio utiliza esquemas independientes en una única base de datos.
 - Es el menor grado de aislamiento entre entornos de dominio.
 - Cada equipo de dominio puede seguir poniendo en marcha y escalando sus propios clústeres de procesamiento aislados de otros dominios.
 - Las convenciones de nomenclatura podrían requerir un mayor esfuerzo para distinguir los objetos de diferentes dominios.
 - Puede ser útil para subdominios en una situación de dominio/subdominio.

- En este artículo no se aborda la opción “Esquema por dominio” en detalle, pero es muy similar a la opción “Base de datos por dominio”.
- **“Dominios heterogéneos”**: los dominios pueden utilizar diferentes pilas tecnológicas.
 - Algunos dominios usan Snowflake y otros usan otros sistemas.
 - Algunos dominios se encuentran en la nube y otros pueden alojarse on-premise.
 - Generalmente incurre en un mayor grado de complejidad para albergar entornos de dominio heterogéneos.
 - Requiere una consideración especial, ya que puede ser contraria al objetivo de la data mesh de utilizar una plataforma de autoservicio común e independiente del dominio en todos los dominios.

Es posible y plausible usar variaciones de arquitectura o enfoques híbridos derivados de estos tipos básicos. Por ejemplo, una empresa puede elegir “base de datos por dominio” y tener estas bases de datos en varias cuentas en lugar de en una sola. Además, algunos dominios pueden usar bases de datos independientes, mientras que otros usan la totalidad de una cuenta de Snowflake. Por otra parte, el entorno que utiliza un equipo de dominio suele consistir en Snowflake y herramientas adicionales en función de sus requisitos y habilidades.

La clave es que Snowflake admite varias opciones de arquitectura que posibilitan diferentes concesiones de autonomía de dominio y descentralización, por un lado, frente a diferentes grados de complejidad y gestión operativa, por otro.

Cada empresa debe encontrar el equilibrio entre centralización y descentralización que mejor se adapte a su tamaño, infraestructura heredada y cultura organizativa. Lo mismo se aplica a la gobernanza federada, respecto a la que las empresas deben lograr el equilibrio entre control centralizado y autonomía de dominio local que más les convenga.

En las siguientes secciones se explican estas opciones de arquitectura en mayor detalle. El enfoque de esas explicaciones se centra principalmente en las topologías de Snowflake, y no en la integración con herramientas de terceros que nuestros clientes suelen utilizar junto con Snowflake en sus iniciativas de data mesh.

Cuenta única: base de datos por dominio

Se trata de una topología popular, adoptada por muchos de nuestros clientes de data mesh, que precisa una única cuenta de Snowflake en la que los dominios utilizan bases de datos y clústeres de procesamiento independientes como entornos autónomos. A cada dominio se le pueden asignar una o varias bases de datos y clústeres para satisfacer sus necesidades de desarrollo, pruebas y producción. La naturaleza de autoservicio de la plataforma permite a los dominios utilizar las funciones de clonación sin copias de Snowflake para (re)crear entornos de desarrollo y pruebas de forma instantánea y frecuente. Además, se puede permitir que diferentes usuarios de un dominio pongan en marcha y escalen sus propios clústeres de procesamiento para satisfacer sus necesidades respectivas en régimen de autoservicio. No obstante, pueden configurarse la supervisión de costes y consumo, así como cuotas, para dominios u otros niveles de detalle de la jerarquía de usuarios y recursos.

El hecho de que todos los dominios utilicen Snowflake Data Cloud les permite tener entornos y recursos de procesamiento independientes que no sean silos físicos, que dificultarían el acceso a los productos de datos.

Parte de la gobernanza se decide de forma centralizada y se aplica a todas las bases de datos con un proceso de DevOps. Esto se puede facilitar mediante funciones, como las etiquetas de objeto, para disponer de una visión general sencilla de los diferentes objetos que pertenecen a los dominios. Dentro de los dominios, la gobernanza está controlada por los equipos de dominio que aplican el control de acceso basado en roles, así como por políticas de acceso a nivel de fila y columna para proteger los datos y evitar que ciertos usuarios y dominios obtengan acceso no deseado a determinados datos.

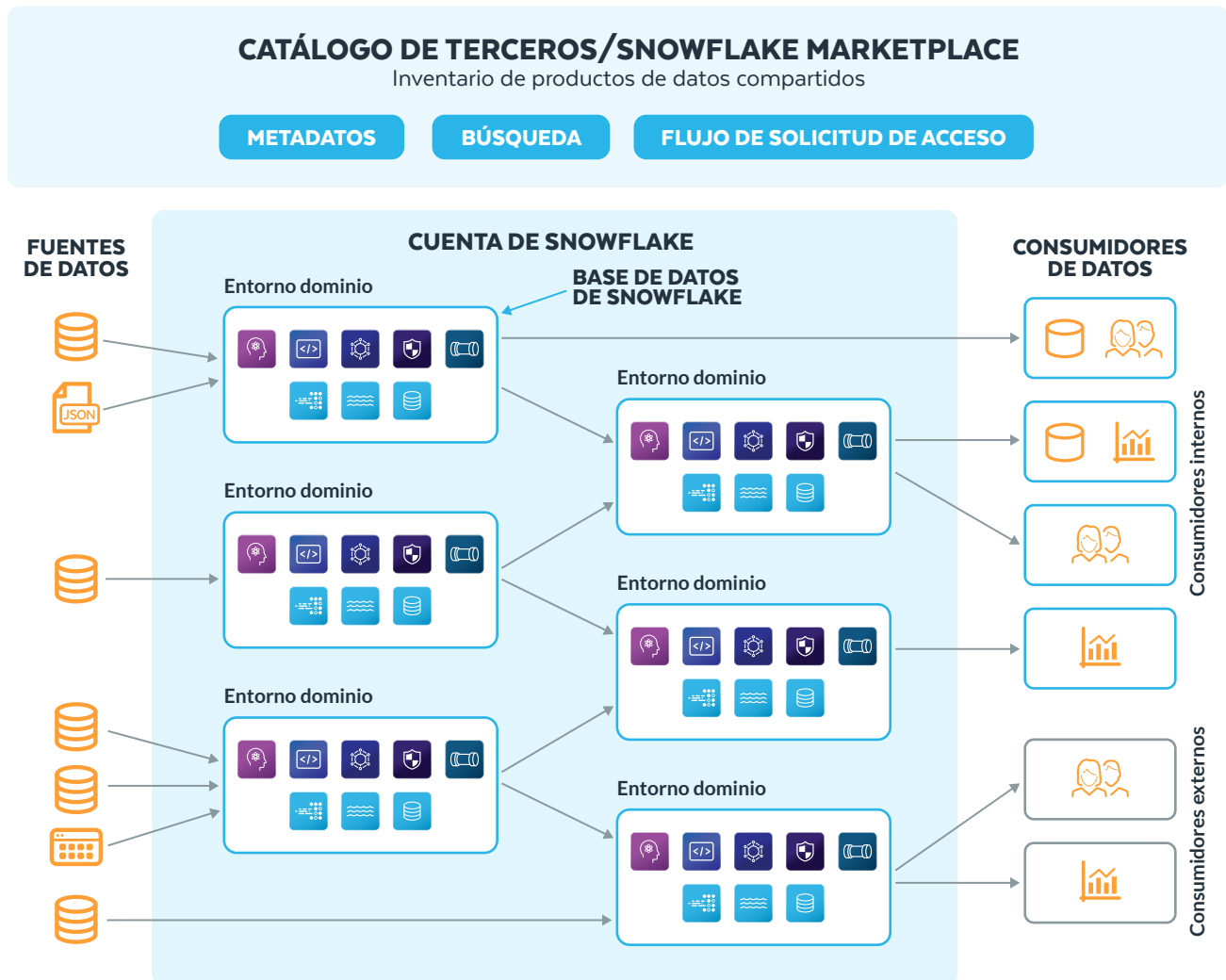


FIGURA 4: CUENTA ÚNICA DE SNOWFLAKE: BASE DE DATOS POR DOMINIO

Cada dominio puede tener varios esquemas, uno de los cuales sirve como capa para que los productos estén disponibles en otros dominios. Otro enfoque sería utilizar una base de datos compartida común, en la que cada dominio tendría un esquema para publicar sus productos de datos como vistas (sin copias). Estos productos pueden ser datos estructurados, semiestructurados o no estructurados, en función de los datos de los que conste el producto. A continuación, los productos se incluyen en un catálogo de datos de terceros para que se puedan detectar.

Hemos visto varias formas de solicitar acceso a un producto. Por ejemplo, siguiendo un enfoque manual, el solicitante debe abrir una incidencia para que la procese el equipo de dominio, que denegará el acceso o lo concederá otorgando el rol correspondiente al solicitante para que pueda acceder. Algunos catálogos proporcionan un flujo más automático.

Tener todos los dominios en una cuenta de Snowflake ofrece las siguientes ventajas:

- El acceso a los productos de datos se puede otorgar fácilmente estableciendo permisos dentro de la base de datos.
- La administración centralizada de políticas de red, seguridad y gobernanza simplifica la gestión general.
- La recuperación ante desastres es más sencilla, ya que solo requiere una cuenta en otra región o nube.

Las convenciones de nomenclatura deben planificarse cuidadosamente, ya que puede haber muchos objetos, y se debe tener en cuenta que cada dominio puede requerir entornos de desarrollo, pruebas, aceptación y producción (Development, Test, Acceptance and Production; DTAP), que pueden crear fácilmente mediante clonación sin copias.

El enfoque de esquema por dominio es similar a este. Las implicaciones de este enfoque son similares al enfoque de base de datos por dominio, ya que todo está organizado de manera lógica en una cuenta de Snowflake. Sin embargo, debe tenerse en cuenta que disponer de una base de datos por dominio es más fácil a la hora de compartir datos con consumidores externos de forma pública a través de Snowflake Marketplace, o de forma privada mediante controles de detección de fichas.

Varias cuentas: cuenta por dominio

Se trata de otra posible topología, que permite que cada dominio funcione en una cuenta de Snowflake independiente. Estas cuentas pueden alojarse en la misma o en diferentes regiones y plataformas de nube. Snowflake Data Cloud global permite a las empresas y los dominios compartir datos entre cuentas, regiones y plataformas de nube, y obtener fácilmente acceso estandarizado a los productos de datos de los demás de forma segura y gobernada. Algunos de nuestros clientes utilizan esta capacidad para respaldar una data mesh multinube y de varias regiones.

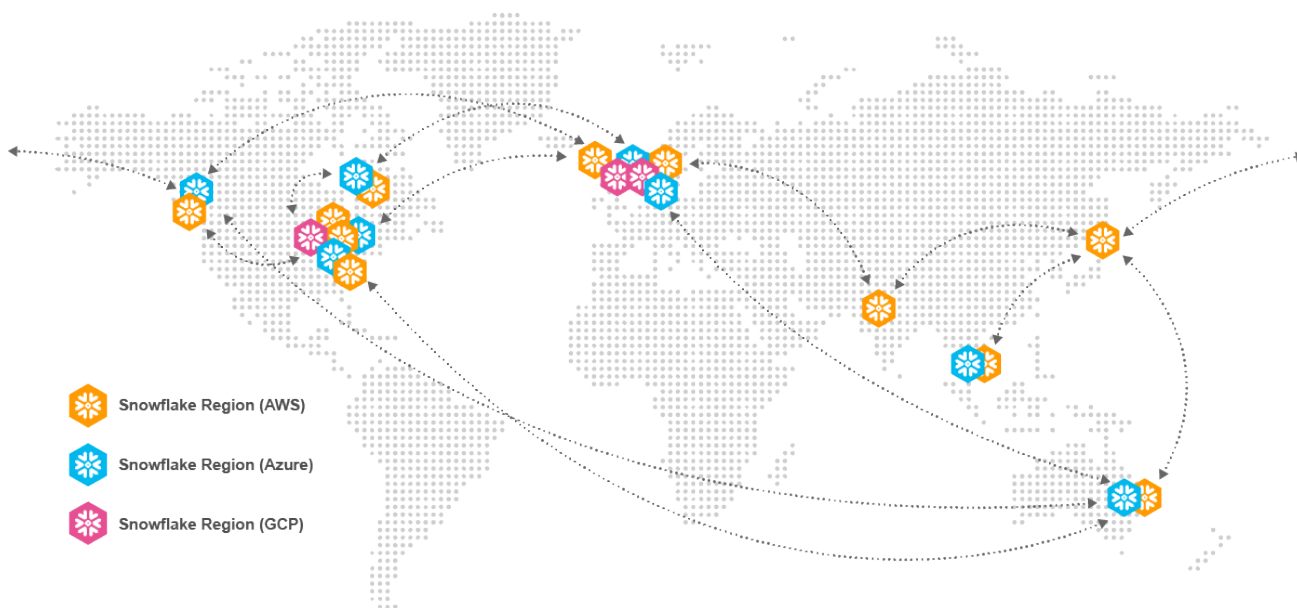


FIGURA 5: SNOWGRID COMO PLATAFORMA DE DATA CLOUD GLOBAL

Existen varias razones por las que las empresas eligen esta topología. Por ejemplo, una empresa puede operar de una forma distribuida globalmente en la que diferentes dominios podrían coordinarse de manera natural con distintas ubicaciones y regiones del mundo. Es posible que algunas empresas operen a nivel mundial y necesiten contemplar los requisitos de ubicación de los datos (como una multinacional que tenga ciertos datos que no puedan salir de Europa sin anonimización, enmascaramiento u otras medidas para garantizar el cumplimiento de los reglamentos de privacidad de los datos).

Otro motivo común son las fusiones y adquisiciones que pueden obligar a una empresa a intercambiar datos entre regiones o plataformas de nube. Algunas empresas tratan deliberadamente de encontrar una estrategia multinube con fines de diversificación o para dar cabida a las preferencias y las inversiones existentes de diferentes unidades de negocio. El uso de cuentas independientes también ofrece una mayor autonomía de los dominios; por ejemplo, si se necesita gestionar los usuarios y la seguridad de forma independiente en cada dominio.

La topología resultante (figura 6) es lógicamente muy similar al uso de una base de datos independiente por dominio, excepto que cada dominio ahora “posee” una cuenta de Snowflake independiente y utiliza las funciones de Snowflake Marketplace y data sharing de Snowflake para que otros usuarios puedan acceder a los productos de datos.

En comparación con el enfoque de base de datos por dominio, puede ofrecer las siguientes ventajas:

- Las funciones de colaboración y data sharing se pueden utilizar en diversos dominios.
- Los estándares de nomenclatura globales se pueden aplicar más fácilmente, ya que cada cuenta es un espacio de nombres independiente.
- Se admiten las preferencias regionales y de plataforma de nube.
- La seguridad y los usuarios se gestionan en cada cuenta de forma independiente.

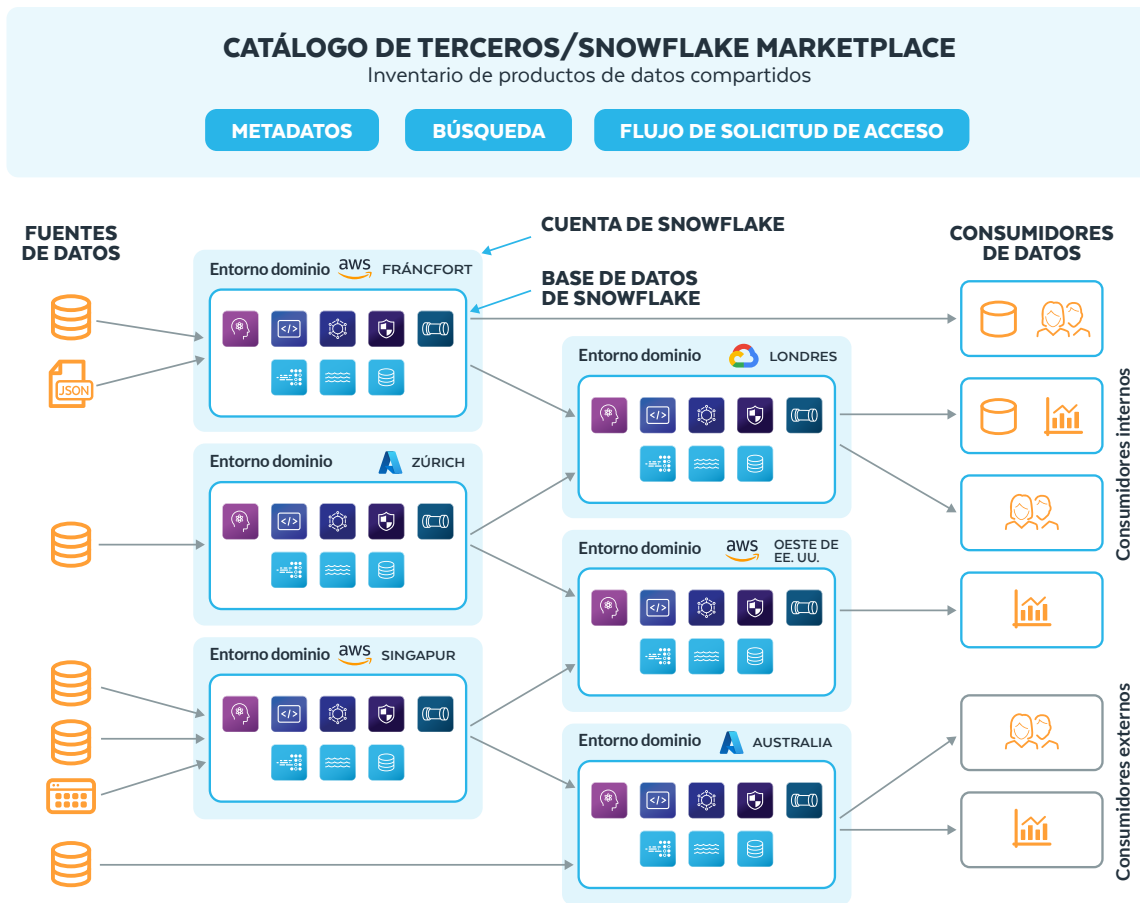


FIGURA 6: VARIAS CUENTAS: UNA CUENTA POR DOMINIO

Arquitectura heterogénea

Algunos clientes nos han preguntado cómo integrar otros entornos de dominio que no son de Snowflake en las topologías mencionadas anteriormente. Estas integraciones dan lugar a una arquitectura heterogénea en la que no todos los dominios utilizan la misma plataforma de datos independiente del dominio para implementar flujos y productos de datos. A menudo, esto se debe al deseo de reutilizar varios repositorios o pilas tecnológicas diferentes que ya existen en diferentes partes de la organización.

Hemos observado que una arquitectura tan heterogénea suele aumentar los costes y la complejidad de la transición a la data mesh. El motivo es que una mayor heterogeneidad de los sistemas implicados obstaculiza la unificación en materia de gobernanza, seguridad, metadatos, estándares de interoperabilidad, rendimiento, habilidades necesarias, asistencia de TI y otras áreas críticas. Por lo tanto, animamos a los clientes a considerar cuidadosamente el rol de los diversos sistemas y repositorios que pretenden integrar en una data mesh. ¿Estos sistemas se utilizan realmente como entornos de dominio que crean y ofrecen productos

de datos? ¿O deben considerarse fuentes de datos que los dominios usan como entrada? Este último caso a menudo permite a los clientes recurrir a las topologías mencionadas anteriormente.

Un enfoque para integrar implementaciones de dominios que no son de Snowflake es que envíen activos de datos o “productos de datos casi listos” a una capa intermedia para la que Snowflake puede actuar como “proxy” que proporcione al resto de la data mesh los productos de datos con una gobernanza, una seguridad, una interoperabilidad, etc. coherentes.

Esta capa intermedia podría ser, por ejemplo, temas de Kafka que se incorporan a Snowflake a través de una ingesta continua, seguida de actualizaciones automáticas de productos de datos en Snowflake. La capa intermedia también podría ser uno o varios contenedores de almacenamiento en la nube de Amazon S3, Azure Blob Storage, Azure Data Lake Storage o Google Cloud Storage. Los formatos de datos pueden incluir JSON, XML, Parquet, AVRO, Apache Iceberg y Delta Lake, entre otros. Snowflake puede incorporar nuevos archivos de contenedores

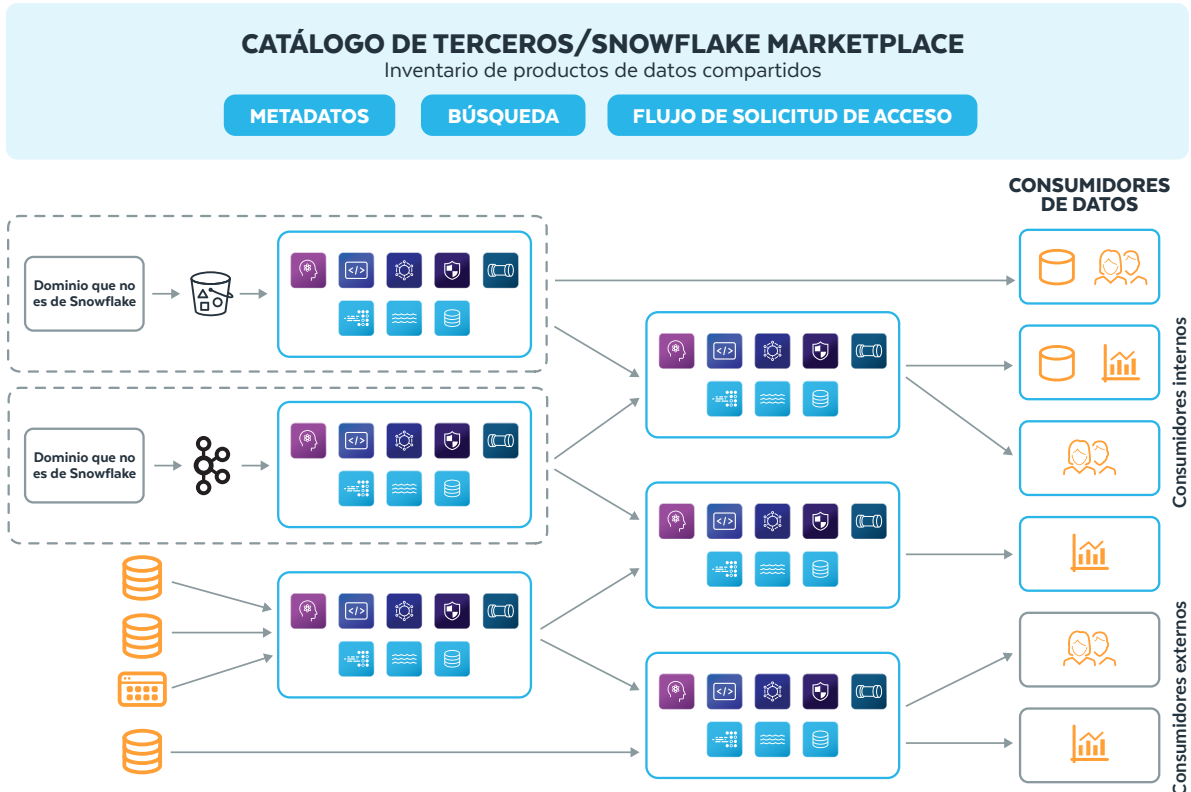


FIGURA 7: ARQUITECTURA HETEROGÉNEA

de almacenamiento de forma automática y continua para obtener el mejor rendimiento, seguridad y gestión automática, o bien exponer el acceso de lectura a los archivos de ese tipo, como tablas externas, al resto de la data mesh. Las tablas externas de Snowflake son, en esencia, vistas de datos que residen en archivos fuera de Snowflake. Sin embargo, las tablas externas son objetos de datos de primera clase en Snowflake que se pueden proteger, gobernar, combinar e, incluso, compartir a través de las funciones de colaboración de Snowflake, de forma muy similar a otros objetos de datos en Snowflake. Esto permite a Snowflake actuar como una capa de integración que puede exponer los datos externos de forma coherente y gobernada, sin necesidad de ingerir ni duplicar los datos.

Existen otras opciones para integrar entornos que no son de Snowflake en la plataforma de datos; pero no los trataremos en este documento.

Algunas empresas consideran que la virtualización de datos es una posible solución para integrar un conjunto diverso de entornos de dominio. Aunque existen casos de uso válidos de virtualización de datos, hemos observado que esta también plantea

una serie de desafíos. Uno de ellos es el rendimiento cuando es necesario combinar datos de varios repositorios diferentes. Para hacerlo, normalmente hay que trasladarlos a un lugar común para procesar la combinación, incluso aunque otros predicados puedan enviarse a las fuentes de datos.

Esto puede impedir la virtualización en casos de uso sensibles respecto al rendimiento. Con Snowflake, la combinación de varios objetos de datos en una única base de datos tiene aproximadamente las mismas características de rendimiento que cuando estos objetos de datos están en bases independientes, o incluso en cuentas de Snowflake independientes, que es una propiedad importante de la arquitectura de la plataforma de Snowflake. Otro desafío que hemos observado en algunas empresas es que la virtualización puede alentar a los equipos a seguir trabajando en sus respectivas islas tecnológicas, que suelen ser muy específicas de cada dominio, en lugar de esforzarse por conseguir una plataforma de autoservicio común e independiente del dominio.

RESUMEN

La data mesh no es la panacea para todos los desafíos de gestión e integración de datos. No obstante, si determina que la data mesh es el enfoque adecuado para su empresa, asegúrese de centrarse en las cuestiones organizativas y no técnicas requeridas para alcanzar el éxito. Algunos ejemplos incluyen cambios organizativos, funciones y responsabilidades, personal, incentivos y rendición de cuentas, apoyo de las partes interesadas clave o cambios de mentalidad en términos de concepción de productos.

Con el tiempo, tendrá que diseñar una arquitectura de TI de autoservicio que admita dominios distribuidos y productos de datos con gobernanza federada. Snowflake puede desempeñar esa función clave como plataforma de autoservicio fácil de usar para los equipos de dominio. Snowflake admite diferentes topologías que permiten a las empresas elegir el grado deseado de descentralización y autonomía de dominio, al tiempo que garantiza que los dominios mantengan la interconexión y la interoperabilidad. Snowflake permite topologías de cuenta única, así como arquitecturas multinube y de varias regiones, y respalda la integración de dominios externos o configuraciones de colaboración entre varias empresas. La plataforma de Snowflake subyacente, con Snowgrid global y Snowflake Marketplace, actúa como un tejido conectivo que ayuda a las organizaciones a evitar el riesgo de que se creen silos de datos.

Además, Snowflake proporciona una amplia gama de funciones que ayudan a las empresas a implementar los conceptos de datos como producto y gobernanza federada. Snowflake también se integra fácilmente con numerosas herramientas de terceros que pueden proporcionar funciones de plataforma adicionales. Snowflake Data Cloud es una excelente opción tecnológica con la que complementar los cambios y procesos organizativos necesarios para que la transformación de la data mesh sea satisfactoria. Para obtener más información sobre cómo puede ayudarle Snowflake, visite snowflake.com/data-mesh.

ACERCA DE SNOWFLAKE

Snowflake permite a cualquier organización movilizar sus datos con Snowflake Data Cloud. Los clientes utilizan Data Cloud para unificar, descubrir y compartir datos de forma segura, y ejecutar diversos workloads analíticos. Independientemente de la ubicación de los datos o de los usuarios, Snowflake ofrece una experiencia de datos única que abarca varias nubes y regiones geográficas.

Miles de clientes de numerosos sectores, incluidas 543 de las empresas que figuran en Forbes Global 2000 (G2K) (2022), a fecha del 31 de octubre de 2022, utilizan Snowflake Data Cloud para impulsar sus negocios. Más información en [snowflake.com](https://www.snowflake.com).



© 2022 Snowflake Inc. Todos los derechos reservados. Snowflake, el logotipo de Snowflake y el resto de nombres de productos, funciones y servicios de Snowflake mencionados en este documento son marcas registradas o marcas comerciales de Snowflake Inc. en Estados Unidos y otros países. El resto de logotipos o nombres de marcas mencionados o utilizados en este documento se usan únicamente con fines identificativos, y pueden ser las marcas comerciales de sus respectivos titulares. Snowflake puede no estar asociado con, patrocinado o apoyado por cualquiera de dichos titulares.