



5 PRÁCTICAS RECOMENDADAS PARA DESARROLLAR ALMACENES DE DATOS

ÍNDICE

- 3** Introducción
- 4** Cree un modelo de datos
- 7** Adopte una metodología ágil para almacenes de datos
- 8** Opte por ELT en vez de ETL
- 9** Adopte una herramienta de automatización
- 10** Forme al personal sobre nuevos enfoques
- 11** Resumen
- 12** Acerca de Snowflake



INTRODUCCIÓN

La tecnología de la nube ha revolucionado la forma en que las empresas almacenan, analizan y acceden a los datos. Independientemente de si su organización pretende crear una nueva plataforma de datos desde cero o reestructurar un sistema de almacén de datos heredado para sacar partido a nuevas capacidades, el proyecto tendrá más posibilidades de éxito si conoce algunas directrices y prácticas recomendadas. Si bien algunas de estas prácticas pueden resultar obvias, con demasiada frecuencia se dedica poco tiempo a definir y documentar ciertas decisiones al principio, lo que más tarde deriva en quebraderos de cabeza y falta de eficiencia.

En este ebook, presentamos cinco recomendaciones para estructurar una estrategia de datos y coordinar toda la empresa en torno a ella, de forma que el almacén de datos resultante satisfaga las necesidades empresariales actuales y futuras. Gracias a estas prácticas recomendadas para desarrollar almacenes de datos, será más fácil que todas las partes interesadas de la empresa extraigan más valor del almacén creado. Además, permitirán sentar las bases de una plataforma de datos más amplia, que podrá crecer y adaptarse a su empresa a medida que esta cambie.

1. CREE UN MODELO DE DATOS

El primer paso de cualquier programa de datos es crear un modelo de datos: una representación abstracta que organiza los elementos de datos y describe la relación que hay entre ellos y con las propiedades de sus entidades reales. Un modelo de datos establece una definición y conceptualización comunes de qué información es importante para la empresa, así como su panorama de datos general. Disponer de un modelo de datos ofrece un método para documentar los conjuntos de datos que se incorporarán en el almacén de datos, la relación que existe entre esos conjuntos y los requisitos empresariales que se pretenden satisfacer con la plataforma.

Entonces, ¿sería posible crear un almacén de datos sin un modelo de datos? Sí, pero si decide saltarse este paso básico, perderá mucha información valiosa. Crear un modelo de datos completo suele ser un ejercicio revelador para las empresas, ya que obliga a diversos equipos funcionales a acordar la definición y descripción de los activos de datos y los requisitos empresariales respecto al almacén de datos antes de iniciar su desarrollo.

Un modelo de datos bien definido genera un impacto positivo mucho después de que el almacén de datos (o “data mart”) se haya puesto en funcionamiento. Por ejemplo, un modelo de datos establece el linaje de datos de todos los objetos del almacén de datos, lo que facilita la incorporación de nuevos miembros al equipo o de nuevos objetos de datos en el almacén a medida que cambian las necesidades de la empresa.

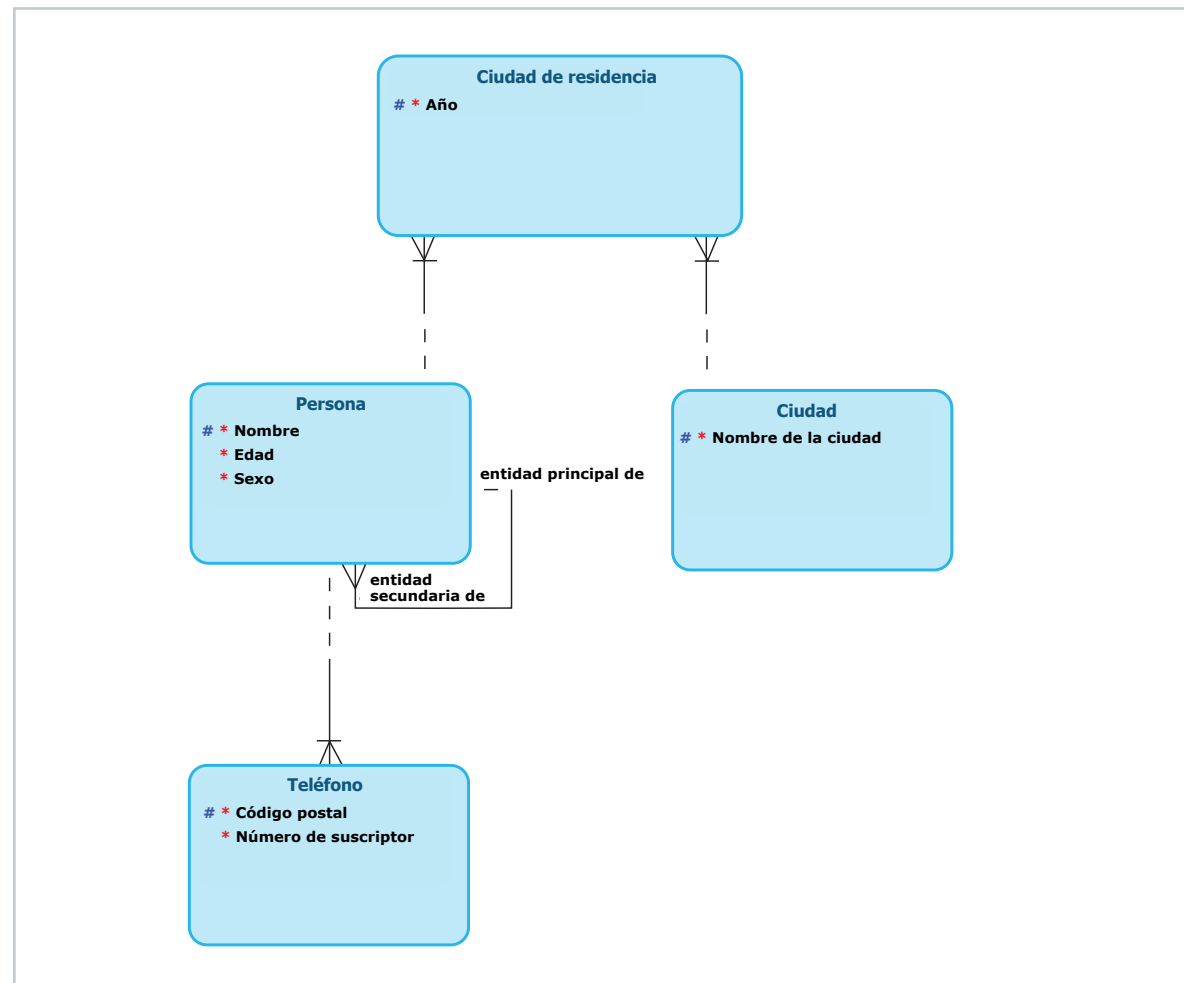


Figura 1: Modelo típico de datos lógicos basado en 3NF

El modelo de datos también proporciona documentación clara sobre el contenido, el contexto y los orígenes. Esto facilita las auditorías y el cumplimiento de nuevos requisitos relativos a los datos, como los que estipula el Reglamento General de Protección de Datos (RGPD), el marco de la Unión Europea que establece directrices para la recopilación y el tratamiento de información personal.

Disponer de un modelo de datos sólido también ayuda a evitar confusiones y las costosas reestructuraciones posteriores. Siempre es buena idea agregar una capa de integración independiente del origen que posibilite el análisis de varios conjuntos de datos en función de las características comunes de dichos conjuntos.

Un almacén de datos reúne muchos orígenes y tipos de datos diferentes, entre los que se incluyen los conjuntos de datos tradicionales, como los datos de gestión de relaciones con los clientes (customer relationship management, CRM) o los datos de planificación de recursos empresariales (enterprise resource planning, ERP), así como los conjuntos de datos relativos a blogs, feeds de Twitter, datos del Internet de las cosas (IdC), e incluso tipos de datos que aún estén por inventar. Por eso, contar con una capa de integración flexible que no esté demasiado vinculada a un solo sistema permitirá preparar su almacén de datos para el futuro.

Un modelo de datos de gran eficacia debe emplear definiciones y estructuras semánticas definidas según el dominio empresarial, no basadas en las definiciones específicas de un solo sistema de origen. Por ejemplo, mientras un sistema de CRM puede referirse a los clientes con el término “cust”, otro puede referirse a ellos como “cust_ID”. Para sacar el máximo partido al almacén de datos, es fundamental establecer una regla semántica que abarque toda la empresa y defina la forma en que los usuarios deben nombrar, consultar y analizar los datos de todos los conjuntos.

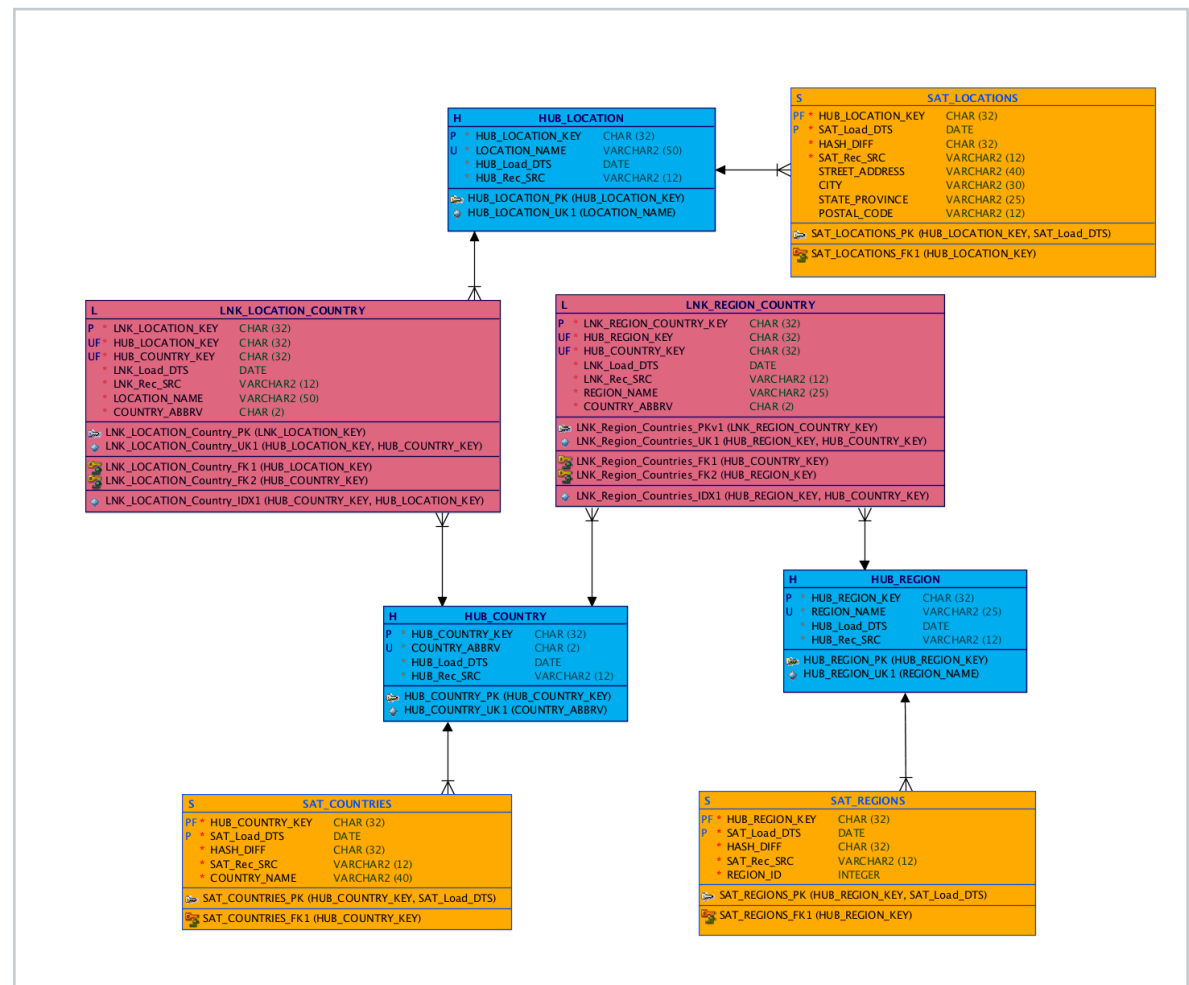


Figura 2: Ejemplo de modelo de datos basado en el enfoque de modelado de DV

A medida que una empresa se enfrenta a cambios, fusiones y adquisiciones, el sistema de CRM que usa en un momento dado puede sustituirse por otro diferente. Si su modelo de datos está estrechamente vinculado a un origen específico, tendrá que realizar una gran reestructuración para integrar un segundo origen que reemplace el sistema antiguo. La asignación de datos se vuelve mucho más sencilla con una capa de integración independiente del origen, que permite reemplazar un sistema de origen antiguo por uno nuevo sin que ello afecte a los informes generados posteriormente ni tener que cambiar el comportamiento de los usuarios.

Una vez implementado el modelo de datos, es esencial seleccionar un enfoque estándar. A continuación, se presentan los principales tipos de estándares de modelado de datos que se usan al diseñar un almacén de datos:

- **Tercera forma normal (third normal form, 3NF):** estándar arquitectónico diseñado para reducir la duplicación de datos y garantizar la integridad referencial de la base de datos.¹
- **Esquema en estrella:** tipo de arquitectura más sencillo y extendido para desarrollar almacenes de datos y data marts dimensionales, que consta de una o varias tablas de hechos que hacen referencia a cualquier cantidad de tablas de dimensiones.²
- **Data Vault (DV):** el modelado de DV, desarrollado específicamente para abordar los problemas de agilidad, flexibilidad y escalabilidad que presentan otros enfoques, se creó como repositorio de datos empresariales históricos, detallado, no volátil, auditable y fácilmente ampliable. Está muy estandarizado y combina elementos de los modelos en estrella y de 3NF.³

Cada arquitectura ofrece sus ventajas, pero la decisión de cuál adoptar dependerá de las necesidades empresariales de la organización.

Sinceramente, no importa tanto qué arquitectura elija una organización, sino el hecho de que elija, documente y respalde continuamente una arquitectura como parte del desarrollo de un modelo de datos para el almacén. Este método posibilitará su eficacia en el futuro, así como una única metodología de asistencia y solución de problemas que facilite a los nuevos miembros de los equipos la puesta en marcha rápida.



2. ADOPTE UNA METODOLOGÍA ÁGIL PARA ALMACENES DE DATOS

Antes, crear un almacén de datos (o incluso un data mart) era una tarea laboriosa, monolítica, de varios trimestres o años y sujeta al tradicional proceso en cascada. En la era moderna ya no suele ser así, ya que muchas organizaciones deciden adoptar un enfoque de diseño más flexible e iterativo, o ágil.

Ahora que las necesidades empresariales cambian a mayor velocidad que nunca y surgen nuevos orígenes de datos en línea rápidamente, las empresas deben ser capaces de adaptarse y aprovechar esos recursos con rapidez y facilidad. Eso implica que deben aprender a crear soluciones de datos y análisis de forma incremental y ágil. Si se cuenta con una planificación adecuada y coordinada con una sola capa de integración independiente del origen, los proyectos de datos de gran envergadura se pueden dividir en tareas más pequeñas que se entreguen en plazos más cortos, lo que permite proporcionar un valor empresarial incremental mucho más rápido.

Para lograr este objetivo, hay arquitectos de almacenamiento de datos que están adoptando la metodología ágil, que surgió inicialmente en el ámbito del desarrollo de software. De acuerdo con la metodología ágil, los requisitos y las soluciones evolucionan con la labor colaborativa de clientes y equipos interdisciplinarios que se autoorganizan. Al aplicar esta metodología a la concepción y construcción del almacén de datos, las empresas pueden activar nuevos conjuntos de datos y resolver nuevos desafíos empresariales con mayor rapidez.⁴

Como parte de la metodología ágil, ha surgido una serie de enfoques que contribuyen a aportar valor más rápido; por ejemplo:

- **Scrum:** este marco de proceso, bautizado con el nombre de la formación de rugby por la que los jugadores entrelazan los brazos y avanzan (también llamada “melé”), es el más extendido en el contexto de desarrollo ágil. Se trata de un marco ligero que mejora la comunicación diaria y la reevaluación flexible de los planes que se ejecutan en fases breves e iterativas de trabajo.⁵ Ralph Hughes codificó la aplicación de Scrum al almacenamiento de datos en una serie de influyentes obras que resultan de utilidad para las empresas que adopten este enfoque.
- **Kanban:** se trata de un método para gestionar la creación de productos, con un énfasis en la entrega continua sin sobrecargar al equipo de desarrollo. Al igual que Scrum, Kanban es un proceso diseñado para ayudar a los equipos a colaborar de forma más eficaz. Recibe su nombre de las tarjetas Kanban, un sistema creado por Taiichi Ohno (ingeniero industrial de Toyota) para realizar el seguimiento de la producción de una fábrica y, así, mejorar la eficacia de la fabricación.
- **Business Event Analysis and Modelling (BEAM):** Lawrence Corr y Jim Stagnitto desarrollaron esta revolucionaria metodología en su libro “Agile Data Warehouse Design”. BEAM se centra en los eventos empresariales, en lugar de en los requisitos conocidos de elaboración de informes, para modelar toda el área de procesos de la empresa. Consta de siete dimensiones para identificar y, posteriormente, detallar los eventos

empresariales que se denominan las “siete uves dobles” en inglés: “who” (quién), “what” (qué), “when” (cuándo), “where” (dónde), “how” (cómo), “how many” (cuántos) y “why” (por qué).⁶

Para sacar más partido al desarrollo ágil, es muy útil disponer de una plataforma de datos ágil. Las plataformas de datos basadas en la nube proporcionan una flexibilidad y una elasticidad estructurales gracias a las que pueden escalarse a medida que evolucionen las necesidades empresariales. Este tipo de plataforma requiere menos esfuerzo, mantenimiento y administración para resultar útil, y se puede ampliar y adaptar a los cambios de los requisitos de la empresa. Gracias a un servicio en la nube moderno, los equipos pueden dedicar menos tiempo a ajustar consultas y aprovisionar el almacenamiento, y más tiempo a abordar los desafíos empresariales inmediatos y aportar valor a la empresa.

No obstante, adoptar metodologías y estructuras ágiles no es una tarea sencilla. Requiere un compromiso con la cultura de la organización y suele suponer un cambio significativo en la mentalidad y el flujo de trabajo respecto a los procesos tradicionales de almacenamiento de datos. Reorganizar un equipo de TI para que pueda trabajar cómodamente en un entorno ágil puede llevar entre 6 y 12 meses, lo que puede resultar paradójico, puesto que el objetivo de la metodología ágil es ofrecer valor más rápido. Esta transición se puede acelerar con la ayuda de un asesor experimentado en dicha metodología. Una vez que se ha efectuado el cambio, los equipos pueden empezar a aplicar más cambios al almacén de datos de forma incremental en cuestión de semanas, en vez de meses.

3. OPTE POR ELT EN VEZ DE ETL

Antes, para desarrollar almacenes de datos, se recurría a un enfoque de extracción, transformación y carga (extract-transform-load, ETL), por el que se extraían los datos de los sistemas de origen que se iban a importar al almacén de datos, se limpiaban o se les aplicaban reglas empresariales en un servidor externo, y se cargaban en el almacén de datos de destino. El aumento de las capacidades y la potencia de procesamiento de las plataformas de datos han dado como resultado un nuevo enfoque preferido: extracción, carga y transformación (extract-load-transform, ELT).

Según el enfoque ELT, se extraen datos sin procesar del origen y se cargan relativamente intactos en el área de almacenamiento provisional del almacén de datos. Pueden añadirse metadatos, la fecha de carga o información del origen a los datos que, después, se incorporan directamente en el almacén de datos. Una vez que los datos están en el almacén, las empresas pueden usar las capacidades de la base de datos para transformarlos, ya sea aplicando un modelo de datos para modificar su estructura, aplicando reglas empresariales o tomando medidas de calidad para limpiar los datos (por ejemplo, corregir direcciones incompletas, estandarizar nombres de campos de datos y eliminar duplicados).

El método de ELT tiene dos ventajas: ahorro de costes y mayor trazabilidad. Este enfoque ayuda a ahorrar costes, ya que permite a las empresas aprovechar las capacidades de la plataforma de datos para transformar los datos, en lugar de tener que usar un servidor externo. La capacidad de procesamiento basada en la nube suele ser mucho más económica que la transformación y manipulación de los datos en un servidor externo, así que trasladar los datos a la nube es mucho más rápido y barato. Con el enfoque de ELT también es más sencillo auditar y realizar un seguimiento de los datos en el futuro, ya que proporciona una imagen de los datos originales directamente en la plataforma de datos. De esta forma, el propio almacén de datos puede desempeñar la función de lo que se denomina “data lake”, donde los datos sin procesar se almacenan de forma persistente.



4. ADOPTE UNA HERRAMIENTA DE AUTOMATIZACIÓN

El objetivo del almacén de datos es activar y entregar datos con mayor rapidez para fundamentar decisiones empresariales y extraer mayor valor de ellos. Una forma de acelerar la entrega es adoptar la metodología ágil. Otra forma es adoptar herramientas de automatización que pueden contribuir a desarrollar e implementar código a mayor velocidad. Dado que muchas tecnologías de almacén de datos se basan en patrones, la codificación necesaria para cargar y estructurar datos suele ser repetible, lo que significa que se puede automatizar. En el mercado hay ciertas herramientas (cada vez más) con las que se pueden automatizar algunas de las tareas de creación o incluso todas.

La automatización permite a las empresas sacar más partido a sus herramientas, realizar iteraciones más rápido y aplicar estándares de codificación con mayor facilidad. Permite crear código estandarizado, algo increíblemente útil en organizaciones en las que, tradicionalmente, los modelos de datos y el código de ETL se desarrollaban a mano. La automatización proporciona un estándar documentado para los distintos artefactos, así como un mecanismo de aplicación y garantía de calidad (quality assurance, QA) para vigilar que todos los desarrolladores y diseñadores sigan dicho estándar.

Las herramientas de automatización que usan plantillas para generar código son especialmente útiles, ya que convierten los estándares en propiedades preferidas dentro de las propias plantillas. Así, la incorporación es más rápida, ya que los desarrolladores y diseñadores pueden usar esas herramientas basadas en estándares, lo que garantiza que la implementación sea consistente y la curva de aprendizaje sea más corta. Que la implementación sea consistente tiene una ventaja añadida: es más fácil realizar pruebas y depurar errores, ya que el código se desarrolla sobre la base de los mismos estándares.

Con estas herramientas, la iteración también se acelera, puesto que los generadores automatizados de código tienden a no cometer errores de sintaxis. La actualización de código suele implicar la adición de un nuevo objeto a la herramienta o la modificación de las propiedades de las plantillas en el nivel global, con lo que se genera código nuevo que está disponible inmediatamente para su implementación en el entorno de pruebas y validación.

5. FORME AL PERSONAL SOBRE NUEVOS ENFOQUES

La adopción de la metodología ágil o del desarrollo de código automatizado no implica solo un cambio de habilidades, sino también de mentalidad. Se necesita formación y educación para garantizar que el equipo aproveche los nuevos enfoques y tecnologías de forma eficaz. Esto puede implicar la incorporación de expertos externos para que formen a los equipos sobre las prácticas recomendadas de Scrum, o bien educarlos respecto a las ventajas, las reglas y las prácticas recomendadas relativas a cualquier arquitectura estándar que haya adoptado la empresa para su plataforma de datos.

Hay muchos recursos del sector disponibles para que la transición a la metodología ágil sea más sencilla. **Agile Alliance**, una organización de afiliados sin ánimo de lucro dedicada a la promoción de los conceptos de desarrollo de software ágil tal y como se describe en el “Manifiesto por el desarrollo ágil de software, ofrece muchas opciones de formación para familiarizarse con la metodología ágil. **Scrum Alliance** ofrece certificaciones y formación básica y avanzada sobre Scrum. Por su parte, hay una selección de partners que ofrecen formación y certificación a través de **Data Vault Alliance**.

Tal y como ocurre con cualquier nuevo proceso y cambio en la cultura, las organizaciones deben gestionar la curva de adopción para garantizar que el nuevo enfoque se aplique de manera consistente y efectiva en las operaciones diarias. Identificar proyectos piloto o de prueba de concepto para iniciar a los equipos en los nuevos enfoques garantizará que los profesionales desarrollen y perfeccionen sus habilidades en contextos protegidos pero reales que acelerarán sus competencias y capacidades.



RESUMEN

Las prácticas recomendadas incluidas en este ebook requieren una inversión inicial para alcanzar el valor empresarial a largo plazo que pueden ofrecer. A cambio, el retorno de esa inversión es doble: sentará las bases de un programa de análisis de datos de éxito desde el principio y acelerará la aportación de valor empresarial incremental a su entorno de datos hasta mucho después del primer lanzamiento de producción.

Dado que los requisitos empresariales cambian, y el deseo de obtener más valor de más datos y tipos de datos sigue aumentando, implementar estas prácticas recomendadas le ayudará a concebir casos de uso de almacenamiento de datos que van más allá de los modelos tradicionales. Gracias a una base sólida y una plataforma ágil, podrá expandirse a otros ámbitos de datos y satisfacer nuevas demandas ampliando el programa para que respalde la data science, el aprendizaje automático, la inteligencia artificial y, quizás, incluso la monetización de datos. Con los recursos flexibles y escalables en la nube de hoy en día, puede lograr cualquier objetivo con sus datos.





ACERCA DE SNOWFLAKE

Snowflake permite a cualquier organización movilizar sus datos con Snowflake Data Cloud. Los clientes utilizan el 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 573 de las empresas que figuran en Forbes Global 2000 (G2K) (2022), a fecha de 31 de enero de 2023, utilizan Snowflake Data Cloud para impulsar sus negocios.

Más información en [snowflake.com](https://www.snowflake.com)



© 2023 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.

CITAS

¹ es.wikipedia.org/wiki/Tercera_forma_normal

² es.wikipedia.org/wiki/Esquema_en_estrella

³ snowflake.com/blog/data-vault-modeling-and-snowflake

⁴ [Agiledata.org/essays/dataWarehousingBestPractices.html](https://agiledata.org/essays/dataWarehousingBestPractices.html)

⁵ scrum.org/resources/what-is-scrum

⁶ bystembuilders.com/beam