

APRENDIZAJE SENCILLO

2.^a edición especial de Snowflake

Data warehouse en la nube

para
dummies[®]



¿Qué es un
data warehouse
en la nube?

Comparación de soluciones
de data warehouse

Cómo elegir un
data warehouse
en la nube

Por gentileza de:



Joe Kraynak
David Baum

Acerca de Snowflake

Snowflake surgió con una visión muy clara: hacer que el data warehouse moderno sea efectivo, asequible y accesible para todos los usuarios de datos. Snowflake facilita a las empresas que hacen un gran uso de los datos una elasticidad instantánea, intercambio de datos seguro y fijación de precios por segundo, a través de varias nubes. Como las soluciones tradicionales a nivel local y en la nube tienen problemas con esto, Snowflake desarrolló un nuevo producto con una nueva arquitectura construida para la nube que combina el poder del data warehouse, la flexibilidad de las plataformas de big data y la elasticidad de la nube a una fracción del coste de las soluciones tradicionales. Snowflake: tus datos sin límites.

Para obtener más información, visita **Snowflake en [snowflake.com](https://www.snowflake.com)**.



Data warehouse en la nube

2.ª edición especial de Snowflake

por Joe Kraynak y David Baum

para
dummies[®]

Data warehouse en la nube para Dummies®, 2.ª edición especial de Snowflake

Una publicación de
John Wiley & Sons, Inc.
111 River St., Hoboken, NJ 07030-5774
www.wiley.com

Copyright © 2020 de John Wiley & Sons, Inc., Hoboken, Nueva Jersey

Ninguna parte de esta publicación se puede reproducir ni almacenar en un sistema de recuperación o transmitir de ninguna forma o por ningún medio, electrónico, mecánico, fotocopiado, grabación, escaneado o de cualquier otra manera, salvo lo permitido en los apartados 107 o 108 de la Ley de Propiedad Intelectual de los Estados Unidos de 1976, sin el permiso previo por escrito del editor. Si deseas solicitar el permiso del editor, debes escribir a Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, o en línea en <http://www.wiley.com/go/permissions>.

Marcas comerciales: Wiley, For Dummies, el logotipo Dummies Man, Dummies.com y cualquier otra imagen comercial relacionada son marcas comerciales o marcas comerciales relacionadas de John Wiley & Sons, Inc. o sus filiales en Estados Unidos y otros países, y no se pueden utilizar sin permiso por escrito. Snowflake y el logotipo de Snowflake son marcas comerciales o marcas comerciales registradas de Snowflake Inc. Todas las demás marcas comerciales son propiedad de sus respectivos propietarios. John Wiley & Sons, Inc. no está asociado con ningún producto ni proveedor que se mencione en este libro.

LÍMITE DE RESPONSABILIDAD/EXCLUSIÓN DE GARANTÍAS: EL EDITOR Y EL AUTOR NO OFRECEN NINGUNA REPRESENTACIÓN NI GARANTÍA SOBRE LA PRECISIÓN O INTEGRIDAD DE LOS CONTENIDOS DE LA MISMA Y ESPECÍFICAMENTE RENUNCIA A TODAS LAS GARANTÍAS, INCLUIDAS SIN LIMITACIÓN LAS GARANTÍAS IMPLÍCITAS DE APTITUD PARA UN PROPÓSITO PARTICULAR. NINGUNA GARANTÍA PODRÁ CREARSE NI EXTENDERSE A TRAVÉS DE MATERIALES DE VENTA O PROMOCIONALES. LOS CONSEJOS Y ESTRATEGIAS QUE SE INCLUYEN EN LA PRESENTE OBRA PUEDEN NO SER APTOS PARA TODAS LAS SITUACIONES. ESTA OBRA SE VENDE ENTENDIÉNDOSE QUE EL EDITOR NO SE DEDICA A INTERPRETAR ASUNTOS LEGALES, CONTABLES NI PROFESIONALES. SI SE NECESITA AYUDA PROFESIONAL, ES PRECISO SOLICITAR LOS SERVICIOS DE UN PROFESIONAL COMPETENTE. NI EL EDITOR NI EL AUTOR SERÁN RESPONSABLES DE POSIBLES DAÑOS DERIVADOS DE LA PRESENTE OBRA. EL HECHO DE QUE SE HAGA REFERENCIA A UNA ORGANIZACIÓN O SITIO WEB EN ESTA OBRA O SE LOS MENCIONE COMO POSIBLES FUENTES DE INFORMACIÓN ADICIONAL NO SIGNIFICA QUE EL AUTOR O EL EDITOR ESTÉ DE ACUERDO CON LA INFORMACIÓN QUE LA ORGANIZACIÓN O SITIO WEB PUEDA PROPORCIONAR NI CON LAS RECOMENDACIONES QUE PUEDA HACER. ASIMISMO, LOS LECTORES DEBEN SABER QUE PUEDEN EXISTIR SITIOS WEB QUE SE MENCIONEN EN ESTE LIBRO QUE HAYAN CAMBIADO O DESAPARECIDO ENTRE EL MOMENTO EN QUE SE ESCRIBIÓ ESTA OBRA Y EL MOMENTO DE LEERLA.

Para obtener información general sobre nuestros productos y servicios, o sobre cómo crear un libro *For Dummies* personalizado para su empresa u organización, ponte en contacto con el Departamento de Desarrollo Empresarial en Estados Unidos en el teléfono 877-409-4177, ponte en contacto con info@dummies.biz o visita www.wiley.com/go/custompub. Para obtener información sobre cómo otorgar licencias a la marca *For Dummies* para productos o servicios, ponte en contacto con BrandedRights&Licenses@Wiley.com.

ISBN 978-1-119-71455-2 (pbk); 978-1-119-71466-8 (ebk)

Producido en los Estados Unidos de América

10 9 8 7 6 5 4 3 2 1

Reconocimientos del editor

Estamos orgullosos de este libro y de las personas que han trabajado en él. Para obtener información sobre cómo crear un libro *For Dummies* personalizado para tu empresa u organización, ponte en contacto con info@dummies.biz o visita www.wiley.com/go/custompub. Para obtener información sobre cómo otorgar licencias a la marca *For Dummies* para productos o servicios, ponte en contacto con BrandedRights&Licenses@Wiley.com.

Entre algunas de las personas que contribuyeron a lanzar este libro al mercado se incluyen las siguientes:

Editora de desarrollo: Nicole Sholly
Editor del proyecto: Martin V. Minner
Editor ejecutivo: Steve Hayes
Director editorial: Rev Mengle
Representante de desarrollo editorial: Karen Hattan

Editor de producción:
Mohammed Zafar Ali
Equipo de colaboradores de Snowflake:
Vincent Morello, Clarke Patterson,
Leslie Steere, Kent Graziano

Índice

INTRODUCCIÓN	1
CAPÍTULO 1: Puesta al día en data warehouse en la nube	3
Definición de data warehouse.....	3
La evolución del data warehouse.....	4
Por qué necesitas un data warehouse en la nube	8
CAPÍTULO 2: Descubrimos por qué surgió el data warehouse moderno	9
Tendencias de datos: volumen, variedad y velocidad	9
Tendencias en informes y análisis	12
Necesidades tecnológicas de cualquier data warehouse moderno	15
CAPÍTULO 3: Criterios para seleccionar un data warehouse moderno	17
Satisface las necesidades actuales y futuras.....	17
Almacena e integra todos los datos en un solo lugar	18
Compatibles con las habilidades, herramientas y experiencia existentes	18
Ahorra dinero a tu organización.....	20
Proporciona resistencia y recuperación de datos	20
Asegura los datos en reposo y en tránsito.....	21
Agiliza la canalización de datos.....	22
Optimiza tu tiempo de generación de valor	22
CAPÍTULO 4: Data warehouse en la nube frente al tradicional (on-premise)	23
Evaluación del tiempo de generación de valor	23
Contabilidad de los costes de procesamiento y almacenamiento.....	24
Dimensionamiento, equilibrio y ajuste	25
Tener en cuenta la preparación de datos y los costes de extracción, transferencia y carga.....	26
Agregar el coste de herramientas especializadas de análisis empresarial.....	27
Tener en cuenta el escalado y la elasticidad	27
Reducir los retrasos y el tiempo de inactividad.....	29
Pensar en los costes de los problemas de seguridad.....	29
Pagar por la protección y recuperación de datos.....	30

CAPÍTULO 5:	Comparación entre distintas soluciones de data warehouse en la nube	31
	Distintas soluciones para el data warehouse en la nube.....	31
	Comparación de arquitecturas.....	32
	Evaluar la gestión de la diversidad de datos.....	33
	Calcular la escala y elasticidad	33
	Comparar las capacidades de concurrencia.....	34
	Garantizar el soporte para SQL y otras herramientas.....	34
	Comprobar el soporte de copia de seguridad/recuperación.....	35
	Confirmar la resistencia y disponibilidad	35
	Optimizar el rendimiento.....	36
	Evaluar la seguridad de los datos en la nube	36
	Trabajo de administración.....	37
	Compartir datos de forma segura	37
	Permitir la replicación global de los datos.....	38
	Asegurar el aislamiento de las cargas de trabajo	38
	Habilitar todos los casos de uso.....	38
CAPÍTULO 6:	Poder compartir datos	39
	Afrontar los retos técnicos.....	40
	Compartir datos con éxito.....	41
	Monetizar los datos.....	41
CAPÍTULO 7:	Maximizar las opciones con una estrategia multinube	43
	¿Qué significa «a través de las nubes»?	44
	Aprovechar la replicación global	44
CAPÍTULO 8:	Asegurar tus datos	47
	Analizar los aspectos clave	47
	Insistir en una actitud integral frente a la seguridad.....	52
CAPÍTULO 9:	Minimizar los costes del data warehouse	53
	Minimizar el coste de almacenamiento.....	53
	Maximizar la eficiencia informática	54
CAPÍTULO 10:	Seis pasos para comenzar a usar el data warehouse en la nube	55

Introducción

Como profesional, gerente o analista, sabes muy bien que el conocimiento es poder y que los datos analizados adecuadamente de manera oportuna proporcionan la información necesaria para tomar decisiones bien informadas y lograr una ventaja competitiva. Hoy en día, las organizaciones tienen una colección mucho mayor de datos relevantes que la que han tenido jamás. Esto incluye una amplia gama de fuentes, internas y externas, incluidos los *data marts*, las aplicaciones basadas en la nube y los datos generados por los equipos.

Desafortunadamente, la arquitectura de los almacenes de datos de los últimos 30 años continúa esforzándose con la carga de conjuntos de datos tremendamente grandes y diversos. Los analistas suelen esperar 24 horas o más para que los datos fluyan al data warehouse antes de que estén disponibles para el análisis. Incluso pueden tener que esperar aún más para ejecutar consultas complejas con esos datos. En muchos casos, los recursos informáticos y de almacenamiento necesarios para procesar y analizar esos datos son insuficientes. Esto hace que los sistemas se bloqueen o fallen. Para evitarlo, los usuarios y las cargas de trabajo deben dejarse en cola, lo que retrasa las cosas todavía más. En tiempos más recientes, han surgido enfoques alternativos, como distintas formas de lagos de datos. Sin embargo, estas soluciones conllevan sus propias limitaciones.

Para seguir siendo eficaces y competitivas, las organizaciones deben ser capaces de aprovechar el poder de la gran cantidad de datos que se generan constantemente y realizar análisis complejos con esos datos. Afortunadamente, la comercialización de la informática en la nube surgió hace más de diez años y ofrece avances en hardware, arquitectura y software que pueden ayudar a tu organización a afrontar este reto y superar tus expectativas.

Acercas de este libro

Bienvenido a la segunda edición de *Data warehouse en la nube para Dummies*, en donde descubrirás cómo tu organización puede aprovechar el poder de grandes cantidades de datos de manera cómoda y asequible para mejorar la eficacia y transformar los datos sin procesar en inteligencia empresarial útil.

Más datos abren la puerta a más y mayores oportunidades, que casi siempre van acompañadas de retos igualmente grandes. Para aprovechar estas grandes oportunidades, debes implementar una solución de data warehouse que pueda almacenar y organizar datos en diversos formatos, proporcionar un acceso cómodo a ellos y mejorar la velocidad a la que puedes analizarlos. Y esto debe hacerse de la manera más rentable posible. En este libro te enseñamos cómo.

Iconos utilizados en este libro

A lo largo de este libro, los siguientes iconos resaltan consejos y aspectos importantes que es preciso recordar, entre otras cosas:



CONSEJO

Los consejos te guían hacia formas más fáciles de realizar una tarea o formas mejores de usar el data warehouse en la nube en tu organización.



RECUERDA

Este icono resalta conceptos que merece la pena recordar a medida que te adentras en la comprensión y aplicación del data warehouse en la nube.



ESTUDIO DE CASOS

En los estudios de casos de este libro te contamos cómo las organizaciones aplicaron el data warehouse en la nube para ahorrar dinero y mejorar significativamente la velocidad y el rendimiento de sus análisis de datos.

Más allá del libro

Si te gusta lo que lees en este libro y quieres saber más, te invitamos a visitar www.snowflake.com, donde puedes obtener más información sobre la compañía y lo que ofrece, probar Snowflake gratis, obtener detalles sobre diferentes planes y precios, ver seminarios web, acceder a comunicados de prensa, obtener información sobre próximos eventos, acceder a documentación y otro soporte, y ponerte en contacto con la empresa, porque les encantará tener noticias tuyas.

- » Analizar el data warehouse: del pasado al presente
- » Comprender las ventajas de un data warehouse en la nube
- » Reconocer cómo encaja el data warehouse en la nube en la economía actual

Capítulo 1

Puesta al día en data warehouse en la nube

De una forma u otra, la informática en la nube y el software como servicio (SaaS) han existido durante décadas. Pero el data warehouse en la nube como servicio (DWaaS) se ha convertido recientemente en una alternativa al data warehouse local convencional y soluciones similares. ¿Por qué ahora? ¿Qué ha cambiado? En este capítulo, respondemos a estas preguntas y a muchas más.

Comenzamos definiendo qué es un data warehouse y exploramos la evolución del data warehouse para mostrar cómo ha llegado esta tecnología a la nube. Luego, analizamos de qué forma pueden beneficiarse las organizaciones del DWaaS en la nube y explicamos por qué hay cada vez más empresas que confían en el data warehouse en la nube para competir en la economía actual basada en los datos.

Definición de data warehouse

Un *data warehouse* es un sistema informático dedicado a almacenar y analizar datos para revelar tendencias, patrones y correlaciones que proporcionan información y conocimiento. Tradicionalmente, las organizaciones han utilizado almacenes de datos para almacenar e integrar los datos recopilados de sus fuentes internas (generalmente, bases de datos transaccionales), como marketing, ventas, producción y finanzas. El data warehouse surgió cuando las empresas observaron que analizar

los datos directamente de esas bases de datos transaccionales los ralentizaba (e incluso los bloqueaba) con el esfuerzo de la actividad transaccional habitual y las cargas de trabajo necesarias para analizar esos datos. Por lo tanto, todos esos datos se duplicaban en un data warehouse para su análisis, dejando que la base de datos se centrara en las transacciones.

Con los años, las fuentes de datos se expandieron más allá de las operaciones empresariales internas y las transacciones externas. Ahora incluyen volúmenes exponencialmente mayores, variedad y velocidad de datos de sitios web, teléfonos móviles y aplicaciones, juegos en línea, aplicaciones de banca en línea e incluso máquinas. Más recientemente, las organizaciones están capturando grandes cantidades de datos de dispositivos del Internet of Things (IoT).

La evolución del data warehouse

Históricamente, las empresas recopilaban los datos de formas bien definidas y muy estructuradas a una velocidad y volumen razonablemente predecibles. Incluso a medida que avanzaba la velocidad de las tecnologías más antiguas, el acceso y el uso de los datos se controlaban cuidadosamente y se limitaban para garantizar un rendimiento aceptable para cada usuario, gracias a la escasez de potencia y almacenamiento informáticos a nivel local y a la dificultad para aumentar esos recursos. Esto requería que las organizaciones toleraran ciclos analíticos muy largos.

Los tiempos han cambiado (consulta la figura 1-1). Los avances en tecnología permiten a las organizaciones tomar decisiones empresariales significativas respaldadas por grandes cantidades de datos.

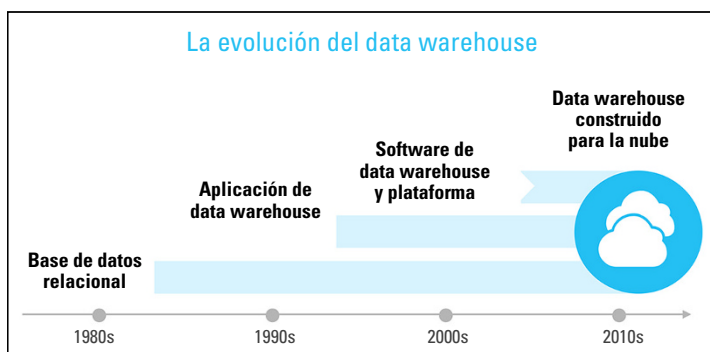


FIGURA 1-1: los sistemas tradicionales hicieron que surgiera el data warehouse en la nube.

No se trata solo de los líderes del mercado o de empresas consolidadas. Las nuevas empresas que llegan al mercado, más pequeñas y ágiles, continúan transformando industrias bien establecidas en cuestión de meses o solo un par de años. Lo hacen con datos para revelar oportunidades y desarrollar productos y servicios que cambian la forma en que los vendedores minoristas y comerciales interactúan con sus clientes.

Reconocer los límites del data warehouse convencional

Los almacenes de datos convencionales no fueron diseñados para manejar el volumen, la variedad y la velocidad de los datos de hoy en día. Los sistemas diseñados más recientemente para abordar estas deficiencias tienen dificultades para acomodar el acceso a los datos y los análisis que necesitan ahora las organizaciones. Los retos de hoy en día revelan que:

- » Las fuentes de datos son más numerosas y variadas, lo que genera estructuras de datos más diversas que deben coexistir en una única ubicación para permitir un análisis exhaustivo y asequible.
- » Las arquitecturas tradicionales provocan de manera inherente competencia entre los usuarios y las actividades de integración de datos, lo que dificulta cargar datos nuevos en el data warehouse y entregar a los usuarios un rendimiento adecuado de forma simultánea.
- » La carga de datos en lotes a intervalos específicos sigue siendo frecuente, pero muchas organizaciones requieren una carga de datos (*microprocesamiento* por lotes) y una transmisión de datos (*carga instantánea*) continuas.
- » Ampliar un data warehouse convencional para satisfacer las crecientes demandas actuales de almacenamiento y carga de trabajo, cuando sea posible, es costoso, complicado y lento.
- » Las plataformas de datos alternativas más recientes a menudo son complejas, y requieren habilidades especializadas y muchos ajustes y configuración. Esto empeora con la creciente cantidad y diversidad de fuentes de datos, usuarios y consultas.

La tecnología y el diseño al rescate

La buena noticia es que la arquitectura de data warehouse y la tecnología (el diseño y los componentes básicos del data warehouse moderno) han evolucionado para satisfacer las demandas de la economía basada en datos con las siguientes innovaciones:

- » **La nube:** un factor clave que impulsa la evolución del data warehouse moderno es la nube. Esto crea acceso a almacenamiento de bajo coste casi infinito, escalabilidad mejorada, la externalización de la gestión y seguridad del data warehouse al proveedor de la nube, y la posibilidad de pagar solo los recursos informáticos y de almacenamiento que realmente se utilizan.
- » **Procesamiento paralelo masivo (MPP):** el MPP, que consiste en dividir una sola operación informática para que se ejecute simultáneamente en una gran cantidad de procesadores informáticos independientes, surgió a principios de la década de 2000. Esta división del trabajo agiliza el almacenamiento y análisis de los datos cuando el software se construye para sacar provecho a esta forma de trabajar.
- » **Almacenamiento en columnas:** tradicionalmente, las bases de datos almacenaban registros en filas, de forma similar al aspecto de una hoja de cálculo. Por ejemplo, podían incluir toda la información sobre un cliente o una transacción minorista. Para recuperar los datos de la manera tradicional, el sistema debía leer toda la fila para obtener un elemento. Esto es laborioso y exige mucho tiempo. Con el almacenamiento en columnas, cada elemento de datos de un registro se almacena en una columna. Con este método, un usuario puede consultar solo un elemento de datos, como los socios del gimnasio que han pagado las cuotas, sin tener que leer el resto de la información de todo el registro, que puede incluir el número de identificación de cada socio, nombre, edad, dirección, ciudad, estado, información de pago, etc. Este método puede proporcionar una respuesta mucho más rápida a este tipo de consultas analíticas.
- » **Procesamiento vectorizado:** esta forma de procesamiento de datos para el *análisis de datos* (la ciencia que examina los datos para sacar conclusiones) aprovecha los diseños recientes y revolucionarios de los chips informáticos. Este método ofrece un rendimiento mucho más rápido en comparación con las soluciones de data warehouse anteriores creadas hace décadas para una tecnología de hardware más antigua y lenta.
- » **Unidades de estado sólido (SSD):** a diferencia de las unidades de disco duro (HDD), las SSD almacenan datos en chips de memoria flash, lo que acelera el almacenamiento, la recuperación y el análisis de los datos. Una solución que aproveche las SSD puede aumentar el rendimiento de manera significativa.

Para obtener más información sobre los avances tecnológicos y otras tendencias que impulsan la evolución del data warehouse, consulta el capítulo 2.

Presentación del data warehouse en la nube

El data warehouse en la nube es una forma rentable que tienen las empresas de aprovechar la última tecnología y arquitectura sin el enorme coste inicial de comprar, instalar y configurar el hardware, software y la infraestructura necesarios. Las distintas opciones de data warehouse en la nube se agrupan generalmente en tres categorías:

- » **Software de data warehouse tradicional implementado en la infraestructura de la nube:** esta opción es similar a un data warehouse local convencional porque reutiliza la base de código original. Requiere conocimientos en TI para construir y administrar el data warehouse. Aunque no tienes que comprar e instalar el hardware ni el software, es posible que aún tengas que realizar una configuración y un ajuste importantes y llevar a cabo distintas operaciones, como copias de seguridad periódicas.
- » **Data warehouse tradicional alojado y administrado en la nube por un tercero como un servicio administrado:** con esta opción, el proveedor externo aporta los conocimientos de TI, pero es probable que tengas muchas de las mismas limitaciones de un data warehouse convencional. El data warehouse se aloja en hardware instalado en un centro de datos administrado por el proveedor. Esto es similar a lo que la industria denomina «proveedor de servicios de aplicaciones» (ASP). Los clientes aún deben especificar de antemano cuánto espacio en disco y recursos informáticos (CPU y memoria) esperan usar.
- » **Un verdadero data warehouse SaaS:** con esta opción, a menudo denominada DWaaS, el proveedor ofrece una solución completa de data warehouse en la nube que incluye todo el hardware y software, y casi elimina todas las tareas relacionadas con el establecimiento y la gestión del rendimiento, el gobierno y la seguridad que necesita un data warehouse. Los clientes generalmente pagan solo por el almacenamiento y los recursos informáticos que utilizan, cuando los utilizan. Esta opción también debe aumentar o disminuir según las necesidades, añadiendo cantidades ilimitadas de potencia informática dedicada a cada carga de trabajo, mientras que una cantidad ilimitada de cargas de trabajo opera simultáneamente sin que el rendimiento se vea afectado.

Para ver una comparación más detallada de las soluciones de data warehouse en la nube, consulta el capítulo 5.

Por qué necesitas un data warehouse en la nube

Cualquier organización que dependa de los datos para prestar un mejor servicio a sus clientes, agilizar sus operaciones y ser líder en su sector se beneficiará de un data warehouse en la nube. A diferencia de los grandes almacenes de datos tradicionales, la nube consigue que las empresas grandes y pequeñas puedan trazar las dimensiones de su data warehouse para adaptarlas a sus necesidades y presupuesto, y ampliar y reducir dinámicamente el sistema a medida que las cosas cambian de un día a otro y de un año a otro.

Estas son algunas de las áreas en las que la vanguardista tecnología de data warehouse en la nube puede mejorar significativamente las operaciones de una empresa:

- » **Experiencia del cliente:** monitorizar el comportamiento del usuario final en tiempo real puede ayudar a las organizaciones a adaptar productos, servicios y ofertas especiales a las necesidades de los consumidores individuales. Con el análisis de las emociones de los clientes, las empresas comprenden mejor a los clientes al analizar cantidades masivas de publicaciones en redes sociales, tuits y otras actividades en línea.
- » **Garantía de calidad:** las organizaciones también pueden usar la transmisión de datos para monitorizar las señales de alerta temprana de problemas de servicio al cliente o defectos en los productos. Pueden tomar medidas en cuestión de minutos u horas, en lugar de días o semanas, lo que no era posible cuando la única fuente de datos eran los registros de quejas del centro de llamadas.
- » **Eficiencia operacional:** la *inteligencia operacional* (IO) consiste en monitorizar el negocio y analizar eventos para identificar en qué áreas una organización puede reducir costes, aumentar márgenes, agilizar procesos y responder a las fuerzas del mercado con mayor rapidez. Al librar a tu organización de la administración del data warehouse, puedes concentrarte en analizar los datos.
- » **Innovación:** en lugar de analizar solo algunos aspectos para comprender el pasado reciente de una industria, las empresas pueden usar nuevas fuentes de datos y el análisis de datos (predictivo, prescriptivo, aprendizaje automático) para detectar y aprovechar las tendencias, alterando así su industria antes de que un competidor desconocido o imprevisto lo haga primero.



RECUERDA

Casi todos los datos de una empresa se almacenan en una multitud de bases de datos dispares. Las preguntas clave que hay que hacerse son: ¿Hasta qué punto son accesibles esos datos? ¿Cuánto costará extraer, almacenar y analizar todos ellos? ¿Qué ocurrirá si no lo haces? Aquí es donde entra en juego el data warehouse en la nube.

EN ESTE CAPÍTULO

- » Adaptarse a las crecientes demandas de acceso y análisis de datos
- » Ajustarse a cómo se crean y usan los datos hoy en día
- » Afrontar los retos con tecnologías nuevas y mejoradas

Capítulo 2

Descubrimos por qué surgió el data warehouse moderno

El data warehouse en la nube surgió de la convergencia de tres tendencias principales: cambios en las fuentes, volumen y variedad de datos; mayor demanda de acceso y análisis de datos; y mejoras tecnológicas que aumentaron significativamente la eficiencia del almacenamiento, acceso y análisis de los datos. En este capítulo, describimos estas tendencias con mayor detalle y explicamos cómo un data warehouse puede aprovechar las ventajas de la nube para abordarlas.

Tendencias de datos: volumen, variedad y velocidad

Cuando hablamos de datos en este libro, estamos hablando de petabytes. Un petabyte equivale a 1 millón de gigabytes. Esto equivale a aproximadamente 500 000 millones de páginas de texto estándar impreso o a 58 333 películas de alta definición, cada una de aproximadamente dos horas de duración. Los datos provienen de las operaciones diarias de una empresa, de las personas que usan sitios web y aplicaciones de software en sus dispositivos móviles, y de la actividad diaria de dispositivos digitales y mecánicos.

En esta sección, nos centramos en los cambios en los datos y en el uso de los datos que han conducido a la demanda del data warehouse en la nube.

Gestión del tsunami de datos

En un pasado no muy lejano, las empresas generalmente administraban los datos que las personas introducían manualmente en el sistema. También se contaba con datos procedentes de fuentes externas, como clientes y socios. La cantidad de datos era relativamente pequeña y predecible, y los datos se almacenaban, administraban y aseguraban en el centro de datos de una empresa, lo que ahora se conoce como un «método a nivel local».

Hoy en día, el mundo de los negocios está sufriendo un tsunami de datos, ya que los datos llegan de una variedad de fuentes como ya se han mencionado en este libro y de otras fuentes que son demasiado numerosas y variadas para enumerarlas. El volumen y la variedad de estos datos pueden abrumar rápidamente a un data warehouse convencional a nivel local y, a menudo, hace que el procesamiento y análisis de los datos se «cuelgue» o incluso bloquee el sistema, debido a la sobrecarga de usuarios y a las cargas de trabajo que se procesan en un momento dado.

Adaptarse al aumento exponencial de los datos exige un nuevo punto de vista (consulta la figura 2 -1). La conversación debe pasar de qué tamaño debe tener el data warehouse de una organización a si puede ampliarse de manera rentable, sin problemas y de la forma necesaria para manejar grandes volúmenes de datos.

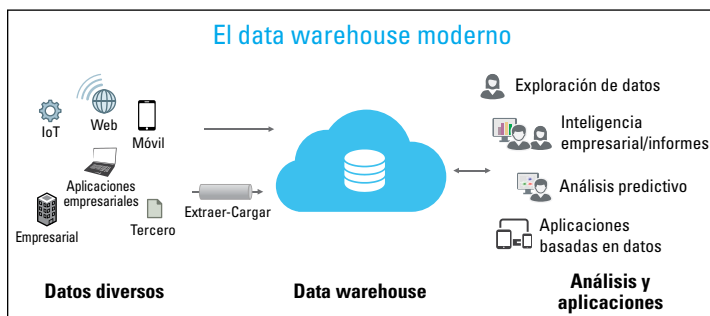


FIGURA 2-1: El data warehouse moderno facilita todos los datos a todos los usuarios.



Los casos de uso que el data warehouse en la nube ha provocado continúan surgiendo. Por ejemplo, las compañías nacidas del SaaS y las grandes empresas que usan la nube para almacenar sus datos están monetizando (vendiendo) esos datos. Los empaquetan como un servicio y los venden a otras organizaciones interesadas en tomar decisiones

comerciales aún mejores a partir de los conocimientos más profundos posibles.

Los beneficios de los datos nacidos en la nube

Las organizaciones han experimentado una rápida adopción del SaaS, incluido el software de gestión de relaciones con los clientes (CRM), paquetes de software de planificación de recursos empresariales (ERP), plataformas de compra de publicidad y herramientas de marketing en línea, entre otros muchos. Gracias a la nube, las nuevas compañías SaaS pueden poner en marcha una tienda por el precio de uno o dos ordenadores portátiles. Estos productos SaaS crean grandes cantidades de datos valiosos que se almacenan en la nube. Además, las organizaciones observan que los proveedores de SaaS ofrecen una mayor seguridad que la que pueden lograr en sus propios centros de datos locales.

La demanda de aplicaciones SaaS/en la nube también ha crecido. La facilidad de la implementación palidece si se compara con lo que requieren las aplicaciones locales para comenzar a funcionar. En el pasado, las empresas podían tener en funcionamiento entre cinco y diez aplicaciones empresariales significativas que generaran datos solamente. Ahora, es normal que incluso las organizaciones medianas tengan cientos o incluso miles de aplicaciones, cada una con el potencial de crear su propio silo de datos (datos de marketing en un sistema, finanzas en otro, información de productos en otro) y ninguno de ellos integrado para poder llevar a cabo un análisis completo y óptimo.

Ahora que la mayoría de los datos de una organización están en la nube, el lugar natural para integrar estos datos también está en la nube. Con el data warehouse en la nube, ya no estás obligado a colocarlo dentro de tu centro de datos, lo que es costoso, requiere mucho tiempo y tiene cada vez menos sentido a medida que crece la cantidad de datos nativos de la nube.

Uso de datos generados automáticamente

Los datos generados automáticamente son un asunto clave relacionado con el Internet of Things (IoT): una colección interminable de dispositivos que comunican datos a través de Internet, incluidos teléfonos inteligentes, termostatos, refrigeradores, plataformas petrolíferas, sistemas de seguridad para el hogar, medidores inteligentes y mucho más. Los datos recopilados y analizados de los dispositivos IoT pueden mejorar productos y procesos, monitorizar equipos y predecir el mantenimiento necesario para evitar averías.

Sin embargo, muchos datos generados automáticamente tienen una relación señal-ruido deficiente. Hay muchos datos valiosos, sí, pero también mucho «ruido». Por lo tanto, a menudo debes guardarlo todo

para encontrar los bits que son útiles. Además, una parte cada vez mayor de estos datos se origina fuera de tu centro de datos. Esto hace que la nube, y su escalabilidad casi infinita, sea la ubicación natural para almacenar e integrar estos datos.

Experimentemos con la exploración de datos

El análisis de los datos comienza con la exploración de datos, identificando conexiones interesantes y útiles que se ponen en manos de los usuarios de datos en forma de informes y análisis. Aunque la exploración de datos no es un concepto nuevo, el crecimiento en el volumen de datos la convierte en un ejercicio que requiere más recursos.

Para la exploración de datos a menudo se necesitan grandes conjuntos de datos. También suele ser de naturaleza experimental, lo que complica la evaluación del ROI necesaria para respaldar el importante coste inicial de la implementación de un data warehouse local tradicional. En respuesta, la nube permite que un data warehouse se amplíe y reduzca según sea necesario y ofrece un modelo de pago por uso que permite a las organizaciones evitar la cuestión de si comprometerse o no con un elevado coste inicial.

Presentación de los lagos de datos

La creciente necesidad de tener cantidades masivas de datos sin procesar en diferentes formatos, todo ello en una sola ubicación, generó lo que ahora se considera el lago de datos heredado. Las organizaciones se dieron cuenta rápidamente de que estas soluciones tenían un coste prohibitivo, ya que era casi imposible transformar esos datos y extraer información útil de ellos.

Pero el interés original en los lagos de datos dejó claro que las compañías querían almacenar todos sus datos en una sola ubicación a un coste razonable. Al añadir un data warehouse moderno en la nube a tu lago de datos existente, o al construir tu lago de datos dentro del data warehouse, puedes lograr fácilmente esa visión original para el lago de datos: cargar, transformar y analizar de forma rentable cantidades ilimitadas de datos estructurados y semiestructurados, con recursos informáticos y de almacenamiento casi ilimitados.

Tendencias en informes y análisis

La toma de decisiones basada en los datos ya no queda relegada únicamente al equipo ejecutivo ni a los científicos de datos. Ahora se usa para mejorar casi todos los aspectos operativos de una empresa. Pero esta creciente demanda de acceso a los datos y análisis en una organización puede ralentizar o bloquear un sistema cuando las cargas de trabajo

compiten por los recursos informáticos y de almacenamiento de los almacenes de datos tradicionales. La eficiencia disminuye, lo que exige a las empresas invertir más tiempo y dinero en una infraestructura adicional para mantener el sistema.

En esta sección, hablamos de algunas de las tendencias que cambian la forma en que las personas acceden a los datos y los utilizan, y cómo esas tendencias impulsan la necesidad de soluciones de data warehouse modernas y construidas para la nube.

Uso de la elasticidad para permitir los análisis

Estas son algunas situaciones en las que el data warehouse elástico construido en la nube puede permitir hacer más con los datos:

- » La exploración de datos tiene muchas ventajas. Pero nadie sabe realmente de antemano los recursos informáticos necesarios para analizar grandes conjuntos de datos, lo que hace que la escalabilidad elástica en función de la demanda sea ideal para este tipo de análisis.
- » El análisis de datos *ad hoc*, que surge todo el tiempo, responde a una sola pregunta empresarial específica. La elasticidad dinámica y los recursos dedicados para cada carga de trabajo permiten estas consultas sin ralentizar otras cargas de trabajo.
- » Los análisis basados en eventos exige datos constantes. Incorporan nuevos datos para actualizar informes y paneles de información de forma continua, de modo que la alta dirección pueda monitorizar el negocio en tiempo real o casi en tiempo real. La ingesta y el procesamiento de los datos de transmisión exige un data warehouse flexible para manejar las variaciones y los picos en el flujo de datos.

Sustitución de la planificación previa exhaustiva con iteración rápida

Los emprendedores tienen normalmente dos caminos para garantizar la comercialización de una nueva idea: la planificación previa exhaustiva o la iteración rápida. La primera opción es un proceso tradicional que lleva mucho tiempo y que implica pensar en una oportunidad o nueva idea de producto, ideando conceptos una y otra vez, con la esperanza de que se genere una demanda del consumidor. La iteración rápida implica probar rápidamente la idea en el mercado y repetir una y otra vez hasta que una versión viable del producto logre el éxito. A partir de ahí, el proceso comienza de nuevo.

La iteración rápida se ha convertido en el proceso más eficaz para dejar fuera de juego a los competidores bien asentados y cambiar la forma en

que todo un sector hace negocios. Pero requiere una recopilación y análisis de alta velocidad de grandes cantidades de datos precisos para tener éxito. Los avances en el almacenamiento y análisis de datos en la nube han hecho que la iteración rápida sea más práctica, al tiempo que mantienen la precisión de los datos.

SATISFACER LA CRECIENTE DEMANDA DE ANÁLISIS DE DATOS



ESTUDIO DE CASOS

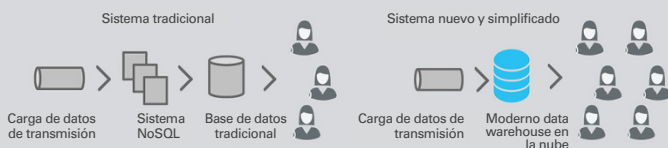
Jana ofrece acceso gratuito a Internet sin restricciones a más de 30 millones de usuarios de teléfonos inteligentes en más de 15 mercados emergentes. Con su aplicación mCent para Android, Jana traslada el coste de Internet móvil de los clientes a más de 4000 empresas a través de contenido patrocinado.

Cuando se presentan nuevas características de contenido de marca o de mCent, Jana analiza y mide las métricas clave, incluida la atención del usuario, el valor del usuario de por vida y los indicadores clave de rendimiento (KPI).

A medida que Jana y sus datos crecían, la arquitectura de análisis inicial de la empresa dejó de servir para el negocio. Las consultas se ralentizaron y los análisis de tablas se volvieron inviables. Para añadir capacidad y sistemas de respaldo y poder administrar el repositorio de datos de código abierto de Jana se necesitaba cada vez más tiempo de administración.

Como se muestra en la figura, Jana actualizó la mayoría de los componentes de su plataforma de datos para optimizar su sistema con un data warehouse construido en la nube para superar estas barreras y obtener los siguientes beneficios:

- Mantener el ritmo de las demandas comerciales de procesamiento y análisis de un flujo de datos dispares en rápido crecimiento.
- Fomentar un mayor uso de los análisis en toda la empresa; el 80 % de los empleados de Jana accede al data warehouse.
- Reducir significativamente los gastos generales de administración.



La transformación de Jana a un data warehouse más rápido, más barato y más eficaz.

Incorporación del análisis

Para muchas empresas, los análisis funcionan como un proceso comercial separado y diferenciado. Pero una tendencia creciente es incorporar los análisis en las aplicaciones empresariales, que se crean cada vez más en la nube. Estas aplicaciones manejan una variabilidad significativa en la cantidad de usuarios que consultan las aplicaciones y en la cantidad de consultas (cargas de trabajo) que los usuarios ejecutan para analizar esos datos. La nube facilita las transferencias de datos desde aplicaciones basadas en la nube al data warehouse en la nube de la organización, donde su escalabilidad y elasticidad son más compatibles con las fluctuaciones en usuarios y cargas de trabajo.

Necesidades tecnológicas de cualquier data warehouse moderno

Las innovaciones tecnológicas pueden mejorar el almacenamiento y el análisis de los datos con respecto a la disponibilidad, simplicidad, coste y rendimiento. En esta sección, nos centramos en las tecnologías clave que deben formar parte de todo data warehouse moderno.

Nube

Las propiedades de la nube la hacen especialmente adecuada para el data warehouse. Las hemos mencionado en otros contextos, pero es importante saber que provienen de la nube:

- » **Recursos ilimitados:** la infraestructura en la nube ofrece recursos casi ilimitados, en función de la demanda y en cuestión de minutos o segundos. Las organizaciones pagan por segundo y solo por lo que usan, lo que permite admitir dinámicamente cualquier ampliación de usuarios y cargas de trabajo sin que el rendimiento se vea comprometido.
- » **Ahorro económico, orientación a los datos:** las empresas que eligen una solución construida en la nube evitan la costosa inversión inicial de hardware, software y cualquier otra infraestructura, y los costes de mantenimiento, actualización y seguridad de un sistema local. En su lugar, se centran en analizar los datos.
- » **Punto de integración natural:** según algunos cálculos, hasta el 80 % de los datos que deseas analizar proviene de aplicaciones que están fuera del centro de datos de tu empresa. Reunir esos datos en la nube es más fácil y barato que construir un centro de datos interno porque no hace falta que compres por adelantado hardware y software por valor de millones de dólares para luego pagar al personal técnico que se encargue del mantenimiento de esos recursos.

Almacenamiento y procesamiento en columnas

Como hemos dicho anteriormente, el almacenamiento en columnas mejora significativamente la eficacia y rendimiento del almacenamiento, la recuperación y el análisis de los datos, lo que permite un acceso más rápido a los resultados para los usuarios del sistema.

Unidades de estado sólido (SSD)

A diferencia de las unidades de disco duro (HDD), las SSD almacenan los datos en chips de memoria flash, que acelera el almacenamiento, la recuperación y análisis de los datos. Estas mejoras aumentan al potencia informática de los almacenes de datos diseñados para usar SSD de manera eficaz.

NoSQL

NoSQL, abreviatura en inglés de «no solo lenguaje de consulta estructurado» (SQL), describe una tecnología que permite el almacenamiento y análisis de formas de datos más novedosas, como los datos generados automáticamente y en redes sociales, para enriquecer y expandir el análisis de datos de una organización. Los almacenes de datos tradicionales no se adaptan muy bien a estos tipos de datos. Por lo tanto, en los últimos años, han surgido enfoques más recientes, como JSON, Avro y XML, para manejar estas formas de datos «semiestructurados».

Algunos de estos sistemas NoSQL se diseñaron con la intención de sustituir a los almacenes de datos tradicionales, pero terminaron complementándolos. Para obtener valor de los datos semiestructurados, las organizaciones a menudo tienen que extraer y transformar los datos de un sistema NoSQL y cargarlos en un data warehouse tradicional para que los usuarios empresariales puedan acceder a ellos fácilmente. Como resultado, esto añade otra capa de complejidad y coste para las empresas (como Jana; consulta el estudio de caso anterior) que intentan capitalizar los beneficios de ambos tipos de sistemas.

Por lo tanto, el data warehouse moderno construido en la nube debe incorporar y optimizar la ingesta y consulta de formatos de datos estructurados (tradicionales) y semiestructurados para que las organizaciones eviten tener que pagar y administrar dos sistemas.

- » **Elegir la solución de data warehouse correcta**
- » **Conseguir una alta relación rendimiento-precio**
- » **Priorizar la seguridad, la protección y el gobierno de los datos**

Capítulo 3

Criterios para seleccionar un data warehouse moderno

Las tendencias de las que hemos hablado en el capítulo 2 han llevado a la necesidad y oportunidad de un nuevo tipo de data warehouse construido para el volumen, la variedad y la velocidad de los datos actuales y para las nuevas formas en que las organizaciones usan sus datos. Dicha solución debe aprovechar las innovaciones tecnológicas clave, incluida la nube.

Cuando estés buscando un data warehouse, debes tener en cuenta una lista de criterios que te ayudará a identificar qué alternativa se adapta mejor a tus necesidades. Considera este capítulo tu lista de verificación para encontrar la mejor solución de data warehouse para tu organización.

Satisface las necesidades actuales y futuras

Una verdadera elasticidad tiene ventajas comerciales, pero también mucho más. Debes poder ampliar tanto los recursos informáticos como el almacenamiento de forma independiente, de modo que no te veas obligado a añadir más almacenamiento cuando en realidad lo único que necesitas es mayor potencia informática (y viceversa). Estas son las capacidades clave de un data warehouse flexible.

Almacena e integra todos los datos en un solo lugar

Los datos no tradicionales o semiestructurados, como hemos explicado en los capítulos anteriores, pueden aportar mucha información a los análisis de datos, más allá de los límites de los datos tradicionales. Pero esto requiere un nuevo enfoque para cargar y transformar estos nuevos tipos de datos antes de que la organización pueda analizarlos. La mayoría de los almacenes de datos tradicionales sacrifican el rendimiento o la flexibilidad para manejar estos tipos de datos. Un data warehouse moderno debe eliminar la necesidad de diseñar y modelar estructuras rígidas y tradicionales desde el principio, para evitar tener que transformar los datos semiestructurados antes de la carga. También debe optimizar el rendimiento de las consultas con respecto a estos tipos de datos mientras aún están en sus formas nativas. En general, el data warehouse debe ser compatible con diversos datos de una forma flexible y evitar problemas de rendimiento.

Cargar de manera eficiente todos tus datos en una sola ubicación resulta crucial. Sin embargo, integrar todos esos diversos tipos de datos para llevar a cabo análisis más precisos es otra cosa. Un data warehouse moderno debe integrar automáticamente tus datos semiestructurados, que estuvieron confinados a los sistemas NoSQL, con datos estructurados inherentes a una base de datos relacional, corporativa y tradicional. No debe haber nada que tengas que instalar ni configurar, y el ajuste y rendimiento deben estar incorporados. Lo más importante, no deberías tener que mantener ni pagar dos sistemas independientes para administrar todos tus datos.

Compatibles con las habilidades, herramientas y experiencia existentes

Los almacenes de datos tradicionales están obsoletos solo porque la tecnología abarca cuatro décadas y no puede rediseñarse fácilmente para la nube. Eso también significa que el lenguaje del que dependen, SQL, sigue siendo un pilar de la industria. Por ello, existe una amplia gama de herramientas consolidadas y nuevas de gestión de datos, transformación de datos, integración, visualización, inteligencia empresarial y análisis que se comunican con un data warehouse SQL. El papel bien establecido del SQL estándar también significa que muchas personas saben cómo trabajar con SQL.

Los almacenes de datos tradicionales son compatibles con SQL, pero no pueden incorporar las capacidades necesarias para almacenar y procesar



ESTUDIO DE CASOS

ANÁLISIS DE DATOS DISPARES

Chime es una banca más inteligente para la generación móvil. Chime reúne y analiza datos a través de plataformas móviles, web y de servidor back-end para mejorar las experiencias de sus miembros al tiempo que aporta valor a tu negocio.

El análisis de las métricas empresariales clave en Chime era laborioso y suponía tener que recopilar y analizar datos de una gran cantidad de servicios, incluidos los servicios de Facebook y Google Ad. Chime también extraía eventos de otras herramientas de análisis de terceros, la mayoría de los cuales proporcionaron datos semiestructurados como JSON.

Con su nuevo data warehouse en la nube, Chime pudo cumplir los siguientes requisitos:

- Entregar con eficiencia datos estructurados y semiestructurados y permitir que estén disponibles para consultas en tiempo casi real, usando tablas de bases de datos SQL estándar.
- Simplificar la canalización de datos sin la necesidad de diseñar un nuevo modelo para cada nuevo tipo de datos que se cargara en su data warehouse.
- Ampliar y reducir las capacidades para satisfacer las demandas de las cargas de trabajo y controlar los costes.
- Integrarse rápidamente y sin problemas con herramientas de análisis de datos de terceros.
- Habilitar SQL en lugar de otras opciones que requieren lenguajes de programación complicados para extraer y analizar datos.

Los analistas de Chime modelan ahora más escenarios para mejorar los servicios para miembros, pasan menos tiempo esperando los resultados de las consultas y dedican más tiempo al análisis de datos.

eficazmente datos semiestructurados. Por lo tanto, muchas organizaciones han recurrido a enfoques alternativos, como las soluciones NoSQL. Las limitaciones de estos sistemas plantean otro problema. Requieren habilidades y conocimientos especializados que no están ampliamente disponibles y podrían no ser compatibles con SQL. Un data warehouse moderno debe estar diseñado con tecnología líder, pero construido sobre estándares inclusivos y establecidos (como SQL), y debe ser compatible con otras habilidades y herramientas que se encuentran con frecuencia en la industria, como los lenguajes informáticos Spark, Python y R.

Ahorra dinero a tu organización

Un data warehouse convencional puede costar millones de dólares en licencias, hardware y servicios, además del tiempo y la experiencia necesarios para configurar, administrar, implementar y ajustar el almacén, más los costes para asegurar y hacer copias de seguridad de los datos. Además, construir un data warehouse que satisfaga los requisitos empresariales y aproveche al máximo el volumen y la variedad de los datos actuales a menudo es prohibitivo para cualquier organización.

Un data warehouse moderno debe afrontar estos retos a un precio mucho más bajo. Por ejemplo, ¿amplía el almacenamiento y la potencia informática por separado para que pagues solo por los recursos que necesitas? ¿Amplía también las cargas de trabajo y la concurrencia? ¿Será compatible con diversas estructuras de datos e integrará datos diversos en un solo lugar? ¿Experimentará un tiempo de inactividad mínimo o nulo y ofrecerá la opción de actualizarse automáticamente o por etapas? Y, finalmente, ¿puede hacer todo esto automáticamente sin la complejidad, el gasto y la molestia de tener que ajustar y reajustar manualmente el sistema para obtener el mejor rendimiento? Consulta el capítulo 5 para comparar almacenes de datos en la nube.



RECUERDA

Con el data warehouse en la nube, tu tarifa de servicio debe cubrir todo por una pequeña fracción del coste de una solución local convencional. Pero no todas las soluciones en la nube son iguales. Sus diferencias también determinan cuánto debe pagar un cliente, de una forma u otra, para obtener información útil de los datos.

Proporciona resistencia y recuperación de datos

Muchos tipos de errores en el data warehouse pueden causar pérdidas o discrepancias de los datos. Por lo tanto, tu data warehouse debe mantener tus datos seguros, actualizados y disponibles. Los almacenes de datos tradicionales suelen proteger los datos llevando a cabo copias de seguridad periódicas, que consumen recursos informáticos valiosos e interfieren con las cargas de trabajo en curso. Las copias de seguridad periódicas también requieren almacenamiento adicional y, a menudo, no incluyen los datos más recientes, lo que provoca discrepancias de datos.

Un data warehouse moderno debe autoadministrarse cuando se trata de garantizar la durabilidad, resistencia y disponibilidad del sistema. No debe interferir con las cargas de trabajo en curso, disminuir el rendimiento ni dar como resultado la falta de disponibilidad del servicio debido a los procesos de copia de seguridad que se estén ejecutando en un segundo plano. Y debe ser económico y ofrecerte formas inteligentes de conservar tus datos sin tener que copiarlos y moverlos a otro lugar.

Por último, tener una arquitectura de varias nubes te brinda portabilidad para reubicar los datos y las cargas de trabajo a medida que tu negocio se expande, tanto entre regiones geográficas como entre los principales proveedores de la nube, como Amazon, Microsoft y Google.

Asegura los datos en reposo y en tránsito

La seguridad de los datos cubre las siguientes dos áreas principales:

- » **Confidencialidad:** evita el acceso no autorizado a los datos.
- » **Integridad:** garantiza que los datos no se modifiquen ni se dañen, que se controlen adecuadamente y continúen teniendo la misma calidad.

Un data warehouse moderno también debe ser compatible con el *control de acceso basado en roles (RBAC)* multinivel. Esto garantiza que los usuarios tengan acceso solo a los datos que pueden ver. Para una mayor seguridad, requiere *autenticación multifactor (MFA)*. Con la AMF, cuando un usuario inicia sesión, el sistema envía una solicitud de verificación secundaria, a menudo a un teléfono móvil. A continuación, el usuario debe introducir el código de acceso que se ha enviado al teléfono. Esto garantiza que una persona no autorizada con un nombre de usuario y contraseña robados no pueda acceder al sistema.

El gobierno de datos garantiza que el acceso a los datos de una empresa y el uso que se haga de los mismos sean adecuados, y que todos los datos se administren y protejan para impedir las infracciones y cumplir con las regulaciones particulares. También requiere una supervisión rigurosa para mantener la calidad de los datos que tu empresa comparte con sus integrantes. Los datos incorrectos pueden conducir a decisiones empresariales perdidas o deficientes, pérdida de ingresos y aumento de los costes. Los administradores de datos, encargados de supervisar la calidad de los datos, pueden identificar los datos cuando estén dañados o sean inexactos, cuando no se actualicen con la frecuencia suficiente para continuar siendo relevantes, o cuando se analicen fuera de contexto.

Cifrar los datos, lo que significa aplicar un algoritmo de cifrado para traducir el texto claro en texto cifrado, es otra característica de seguridad requerida. Una parte más importante de la solución es la «gestión de claves». Una vez que cifras tus datos, usarás una clave de cifrado para descifrarlos. Además de proteger los datos, debes proteger la clave que decodifica los datos. ¿Cuánto tiempo tienes que usar la misma clave? ¿Qué sucede si la clave se ve comprometida? Todo esto hay que gestionarlo. El data warehouse debe usar un enfoque jerárquico de ajuste de claves para cifrar las claves de cifrado y un proceso sólido de rotación de claves para limitar la cantidad de veces que se usa una misma clave.

Además, el proveedor de soluciones de un data warehouse moderno en la nube debe llevar a cabo pruebas de seguridad periódicas, conocidas como «pruebas de penetración», para verificar proactivamente las vulnerabilidades. El proveedor debe aplicar estas medidas de manera coherente y automática sin que se vea afectado el rendimiento.

Para ver un análisis completo de la seguridad y el gobierno del data warehouse en la nube, consulta el capítulo 8.



RECUERDA

Elige un data warehouse con seguridad de extremo a extremo y que sea estándar de la industria. Busca una solución que haya superado auditorías de seguridad como SOC 1/SOC 2 tipo II e ISO/IEC 27001.

Agiliza la canalización de datos

Por *canalización de datos* se entiende principalmente el proceso de *extracción, transformación y carga* por el que se importan datos al almacén en un formato que admite consultas. Una canalización de datos lenta obliga a los usuarios, como los analistas, a pasar demasiado tiempo esperando para acceder a los datos. El rápido crecimiento en la diversidad, cantidad y tamaño de la transmisión de datos no relacionales que llegan desde múltiples fuentes agrava el problema.

Un data warehouse moderno debe reducir la complejidad general del proceso para que la canalización de datos sea más rápida. Las soluciones modernas deben poder cargar eficientemente los datos semiestructurados en su formato nativo y hacer que estén disponibles de inmediato para consultas sin que se necesiten sistemas adicionales e intrincados, como NoSQL, para transformar los datos. Esto permite a los usuarios acceder inmediatamente a los datos de la misma manera que consultan una base de datos SQL. Dichas soluciones pueden proporcionar acceso a nuevos datos de manera exponencial y con mayor rapidez, reduciendo el proceso de gestión y transformación de un día a menos de una hora.

Optimiza tu tiempo de generación de valor

La implementación de una solución no tiene por qué ser una gran tarea, y los aspectos cruciales que antes eran manuales deben automatizarse. Sobre todo, la solución que elijas debe estar disponible todo el tiempo para todos los usuarios y abarcar todos los tipos de datos a una fracción del coste de los sistemas tradicionales. Dicho sistema debe proporcionar información inmediata proveniente de los datos para ayudar a agilizar la organización y aumentar su capacidad para prestar servicio a los clientes y liderar su industria.

- » Acortar el tiempo de generación de valor
- » Reducir los costes de almacenamiento y procesamiento
- » Aprovechar la elasticidad dinámica
- » Externalizar la administración y la seguridad

Capítulo 4

Data warehouse en la nube frente al tradicional (on-premise)

Cuando estés buscando un nuevo data warehouse, lo primero en lo que tienes que pensar es dónde deseas ubicar el data warehouse: en el centro de datos de tu organización o en la nube en la modalidad de software como servicio (SaaS). El data warehouse local tradicional es una tecnología madura y bien establecida que se diseñó mucho antes de que la nube se convirtiera en una plataforma viable. Con la rápida adopción de la nube, existe la necesidad de soluciones de data warehouse que puedan aprovechar al máximo lo que ofrece la nube. En este capítulo, presentamos los aspectos clave a tener en cuenta para el data warehouse en la nube, y lo comparamos con los sistemas tradicionales a nivel local.

Evaluación del tiempo de generación de valor

La implementación de un data warehouse convencional (consulta el capítulo 3) puede llevar al menos un año y convertirse en un proyecto de varios años antes de que puedas extraer información de tus datos. La agilidad de los negocios hoy en día significa que las partes interesadas clave que respaldan el proyecto, así como los facilitadores técnicos y comerciales clave responsables del éxito del proyecto, podrían irse del

equipo o la empresa antes de que el proyecto se ponga en marcha. Un ciclo tan largo también expone al proyecto a recesiones económicas, déficit de ingresos de la compañía y el riesgo de no implementar el proyecto nunca debido a la variación en su alcance.

Además, las soluciones locales no están diseñadas para manejar los datos semiestructurados actuales. Eso requiere añadir una plataforma NoSQL de código abierto, que agregue otra capa de complejidad y alargue la fase de implementación de un nuevo data warehouse .

Si se hace bien, un data warehouse en la nube puede estar en funcionamiento en semanas o en tan solo unos meses. Por lo tanto, la mayor parte del tiempo requerido para comenzar a funcionar debe emplearse extrayendo datos de tus otras fuentes de datos y configurando una herramienta de análisis front-end para extraer información del data warehouse .

Contabilidad de los costes de procesamiento y almacenamiento

Los almacenes de datos locales son caros en términos de hardware, software y administración. Los costes de hardware pueden incluir los costes de los servidores, dispositivos de almacenamiento adicionales, espacio en el centro de datos para alojar el hardware, una red de alta velocidad para acceder a los datos y la alimentación y las fuentes de alimentación redundantes necesarias para mantener el sistema en funcionamiento. Si tu almacén es de misión crítica, añade los costes para configurar un sitio de recuperación ante desastres. Las organizaciones también pagan con frecuencia cientos de miles de dólares en tarifas de licencias de software de data warehouse y paquetes complementarios. Los usuarios finales adicionales, incluidos los clientes y proveedores que tienen acceso al data warehouse , pueden aumentar significativamente esos costes. Después, añade el coste continuado de los contratos de soporte anual, que a menudo suponen el 20 % del coste de la licencia original. Además, un data warehouse local necesita personal especializado en tecnología de la información (TI) para implementar y mantener el sistema. Esto crea un posible atasco cuando surgen problemas y hace que la responsabilidad del sistema recaiga en el cliente, no en el proveedor.

Un data warehouse en la nube reemplaza los gastos de capital iniciales y el coste continuado de un sistema local por precios simples en función de los gastos operativos. Solo pagas una tarifa mensual según la cantidad de recursos de procesamiento y almacenamiento que realmente

utilizas. Hablando de forma conservadora, el coste anual de una solución de data warehouse en la nube puede ser una décima parte del de un sistema local similar.

Dimensionamiento, equilibrio y ajuste

Para conseguir un rendimiento óptimo, el data warehouse local se debe modelar, dimensionar, equilibrar y ajustar, lo que requiere una inversión inicial significativa junto con costes constantes de monitorización y administración. Una configuración así suele incluir:

- » Número y velocidad de las unidades centrales de procesamiento (CPU)
- » Cantidad de memoria
- » Número y tamaño de los discos para la capacidad de almacenamiento requerida
- » *Ancho de banda* de entrada/salida (E/S) (una medida de la cantidad de datos que se pueden transferir en un momento dado)
- » Un modelo de datos personalizado que defina la estructura del almacén, los tipos de datos incluidos y la frecuencia de actualización

Con un data warehouse local, las organizaciones a menudo dimensionan su sistema para un uso máximo, lo que puede representar solo un corto periodo del año. Por ejemplo, una empresa puede necesitar toda la potencia del data warehouse solo al final de cada trimestre o año fiscal. Pero tiene que pagar por esa capacidad máxima las 24 horas del día, todos los días del año, porque el sistema no puede ampliarse o reducirse con facilidad.

El data warehouse en la nube flexible ofrece dos ventajas clave:

- » Las complejidades y el coste de la planificación y administración de la capacidad (dimensionamiento, equilibrio y ajuste del sistema) deben integrarse en el sistema, automatizarse y estar cubiertas por el coste de tu suscripción.
- » Lo mismo ocurre con el aprovisionamiento dinámico de recursos de almacenamiento y procesamiento sobre la marcha para satisfacer las demandas de las cargas de trabajo cambiantes en períodos pico y de uso estable. La capacidad es tener lo que necesites cuando lo necesites. Pero no todas las cargas de trabajo se generan de la misma manera. Un data warehouse en la nube flexible te permite obtener información detallada sobre qué recursos se asignan a qué usuarios y cargas de trabajo.

Tener en cuenta la preparación de datos y los costes de extracción, transferencia y carga

Un data warehouse local debe extraer datos de todas tus fuentes de datos. Después, debe transformar esos datos para que se adhieran a la estructura de datos, a menudo rígida, que está en *dentro* del sistema antes de cargarlos en el almacén. Un problema clave es la adhesión a una cantidad finita y costosa de capacidad de procesamiento y almacenamiento. Como resultado, la transformación de los datos debe tener lugar fuera del horario comercial habitual para evitar la concurrencia con otros trabajos de procesamiento de datos. Esto resulta caro. Además, los datos semiestructurados no llegan en filas y columnas coherentes, rasgo inherente a las estructuras de datos tradicionales. También son datos de gran volumen y alta velocidad.

Las mejores soluciones construidas en la nube pueden cargar los datos semiestructurados directamente sin transformarlos. Estas soluciones pueden proporcionar acceso a datos nuevos hasta 50 veces más rápido que un data warehouse tradicional. Asimismo, el menor coste del almacenamiento ilimitado en la nube proporciona a los analistas de datos acceso a todos los datos en lugar de limitarlos a agregados periódicos de esos datos.



ESTUDIO DE CASOS

OPTIMIZACIÓN DE LA CANALIZACIÓN DE DATOS

DoubleDown, un estudio de juegos en línea, añadió un sistema NoSQL a su canalización de datos con el fin de preparar los datos antes de cargarlos en su data warehouse. Pero este enfoque significó que el registro de eventos diarios de DoubleDown (clics de usuarios y otros datos generados por las actividades de los jugadores) necesitara mucho tiempo de procesamiento. La empresa no podía acceder a los datos de un día hasta las 3 de la tarde del día siguiente. Peor todavía, si uno de los clústeres de procesamiento fallaba, la empresa perdía datos.

DoubleDown eligió un sistema que podía cargar directamente sus datos semiestructurados sin tener que transformarlos primero, haciendo así que esos datos estuvieran disponibles inmediatamente para las consultas. Esto mejoró la calidad y el rendimiento de la canalización de los datos al proporcionar datos a los analistas casi 100 veces más rápido — en 15 minutos frente a 24 horas—, eliminar casi todas las averías frecuentes de la canalización anterior que sufría la empresa, proporcionar a los analistas una granularidad de datos completa en lugar de agregados periódicos, y reducir el coste de la canalización de datos de DoubleDown en un 80 %.

Los analistas de DoubleDown ahora tienen acceso inmediato a los datos de las nuevas versiones de productos para tomar decisiones más rápidas basadas en los datos.

Agregar el coste de herramientas especializadas de análisis empresarial

Como hemos dicho en el capítulo 3, los almacenes de datos locales tradicionales no están diseñados para manejar el volumen, la variedad y la velocidad de los datos actuales. Como resultado, las organizaciones funcionan con dos plataformas de datos: un data warehouse SQL corporativo y a nivel local para el almacenamiento de los datos relacionales tradicionales y una plataforma de big data NoSQL, que puede ejecutarse localmente o en la nube, para almacenar datos no relacionales.

Desafortunadamente, administrar estos sistemas más novedosos es muy complejo y requiere herramientas y conocimientos especializados que no son tan frecuentes como las herramientas y conocimientos de SQL. A fin de cuentas, SQL lleva existiendo décadas, mientras que los sistemas NoSQL son relativamente nuevos.

La solución ideal de data warehouse en la nube ofrece lo mejor de ambos mundos: la flexibilidad para integrar datos relacionales y no relacionales junto con el soporte de las herramientas y habilidades de SQL que se encuentran fácilmente para consultar esos datos.



CONSEJO

Quando estés buscando un nuevo data warehouse, ten en cuenta el coste y la disponibilidad de las habilidades y los conocimientos necesarios para administrar el data warehouse, así como las numerosas herramientas analíticas y de otro tipo que se utilizan junto al data warehouse.

Tener en cuenta el escalado y la elasticidad

Los almacenes de datos convencionales son propensos a ralentizaciones y fallos del sistema ya que los usuarios y los procesos compiten por los recursos limitados. Estos sistemas conectan estrechamente el almacenamiento y el procesamiento en un solo *clúster* de ordenadores (un grupo de ordenadores), lo que hace que sea costoso aumentar uno y no aumentar otro.

Las soluciones de data warehouse más novedosas construidas en la nube proporcionan un procesamiento y almacenamiento prácticamente ilimitados; sin embargo, cuentan con un data warehouse que puede ampliar el almacenamiento por separado del procesamiento (consulta la figura 4-1). Lo ideal es que el data warehouse en la nube se amplíe de tres maneras:

- » **Almacenamiento:** el almacenamiento en la nube es inherentemente escalable, ajustando fácilmente la cantidad de almacenamiento para satisfacer las necesidades cambiantes.

- » **Procesamiento:** los recursos utilizados para procesar las cargas y consultas de datos deberían poder ampliarse o reducirse fácilmente, en cualquier momento, a medida que cambian la cantidad e intensidad de las cargas de trabajo.
- » **Usuarios y cargas de trabajo (conurrencia):** las soluciones con recursos informáticos fijos disminuyen a medida que aumentan los usuarios y las cargas de trabajo. Las organizaciones a menudo se ven obligadas a replicar datos en *data marts* por separado, cambiar algunas cargas de trabajo fuera del horario comercial normal y poner en cola a los usuarios para conservar el rendimiento. Solo la nube puede permitir que un data warehouse se amplíe «horizontalmente» añadiendo clústeres de procesamiento dedicados de cualquier tamaño a un número casi infinito de usuarios o cargas de trabajo que acceden a una sola copia de los datos, sin que ello afecte al rendimiento de los demás.

Busca una solución en la nube que separe el almacenamiento del procesamiento, para que ambos puedan ampliarse de manera fácil e independiente el uno del otro con el fin de reducir los costes. La solución también debe poder ampliarse horizontalmente, para admitir más usuarios y cargas de trabajo sin que ello afecte al rendimiento.

Escalado y elasticidad

- **Escalado elástico para almacenamiento**
Almacenamiento en la nube de bajo coste, totalmente replicado y resistente
- **Escalado elástico para procesamiento**
Ampliación y reducción en función de las cargas de trabajo
- **Escalado elástico para concurrencia**
Añade usuarios y cargas de trabajo sin afectar al rendimiento

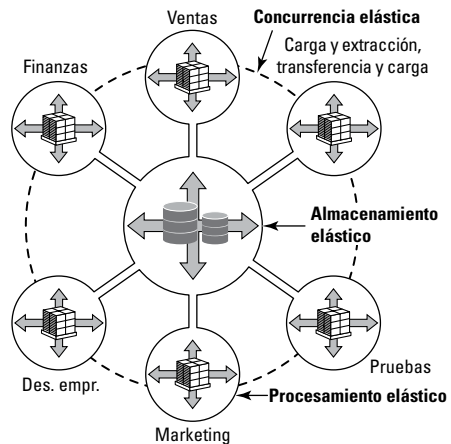


FIGURA 4-1: lo ideal es que el data warehouse se amplíe de tres maneras.

Reducir los retrasos y el tiempo de inactividad

Muchas empresas con soluciones locales se quejan principalmente de dos cosas. Deben esperar horas o más de un día para que los datos recopilados el día anterior estén en el almacén y disponibles. Deben esperar el mismo tiempo para que se ejecute una consulta compleja en un conjunto de datos grande. En algunos casos, cuando hay varios procesos concurrentes, el sistema puede bloquearse, lo que amplía los retrasos y el tiempo de inactividad.

Con un almacenamiento y recursos informáticos virtualmente ilimitados, las soluciones de data warehouse en la nube, diseñadas como dinámicamente elásticas, están mejor equipadas para ampliarse, vertical y horizontalmente, y reducirse para satisfacer las crecientes demandas. Sin embargo, reducir los retrasos y eliminar el tiempo de inactividad no planificado requiere algo más que simplemente aumentar los recursos del sistema. Las mejores soluciones agilizan la canalización de datos y almacenan datos para que las consultas se ejecuten de manera más eficiente sin que sea necesario un ajuste manual.

Busca soluciones que respondan a todos estos tipos de problemas de rendimiento y que minimicen el tiempo de inactividad. La rapidez con la que puedes acceder a tus datos y análisis puede afectar significativamente a tus operaciones y capacidad para mantener una ventaja competitiva.

Pensar en los costes de los problemas de seguridad

Una sola vulneración en los datos puede convertirse rápidamente en una pesadilla para las relaciones públicas y resultar en la pérdida de ventas e importantes multas de las agencias reguladoras. Si bien la nube genera miedo ante los riesgos de seguridad, puede ser más segura que tu centro de datos.

Si optas por un data warehouse local, eres el único responsable de garantizar la seguridad de los datos confidenciales, lo que implica una atención cuidadosa y constante a la protección del firewall, los protocolos de seguridad, el cifrado de los datos (en reposo y en tránsito), los roles y privilegios de usuario, y la monitorización y adaptación a las amenazas de seguridad emergentes.

Implementar una seguridad para los datos que sea eficaz resulta complejo y costoso, especialmente en cuanto a los recursos humanos. Las medidas de seguridad mal implementadas te exponen a costes aún mayores en caso de incumplimiento.

Como los proveedores de data warehouse en la nube prestan servicio a varios clientes, estos pueden permitirse contar con los conocimientos y recursos para proporcionar seguridad en el data warehouse de extremo a extremo de nivel industrial. Busca un proveedor que garantice el cifrado de extremo a extremo estándar de la industria para proteger los datos tanto en reposo como en tránsito.

Pagar por la protección y recuperación de datos

Los almacenes de datos locales son vulnerables a la pérdida de datos por fallos en los equipos, cortes de energía o sobretensión, robo o vandalismo y desastres (incendios, inundaciones y terremotos, entre otros). Para proteger tus datos, debes hacer una copia de seguridad con regularidad y almacenar las copias de seguridad en una ubicación remota. También necesitas una fuente de alimentación de reserva para evitar la pérdida de datos y garantizar que el data warehouse esté siempre disponible para procesar datos y consultas entrantes. Si se produce un desastre, necesitarás personal cualificado para recuperar los datos, utilizando las copias de seguridad más recientes. Si tu data warehouse es de misión crítica, es posible que necesites también un centro de recuperación ante desastres en otra ubicación geográfica (un centro de datos adicional) junto con el software, licencias y procesos para garantizar la conmutación automática por error para que no haya interrupción en el servicio.

La nube proporciona una solución ideal para la protección y recuperación de datos. Por su naturaleza, almacena datos fuera de las instalaciones. Algunas soluciones en la nube hacen copias de seguridad de los datos automáticamente en dos o más ubicaciones físicas separadas. Si los centros de datos están aislados geográficamente, también proporcionan recuperación ante desastres integrada. Los centros de datos en la nube tienen fuentes de alimentación redundantes, por lo que permanecen en funcionamiento incluso durante largos periodos de cortes de suministro. Los proveedores de la nube pueden ofrecer estas protecciones a un precio mucho menor que el que tendrías que pagar tú, ya que distribuyen los costes entre miles de clientes.



CONSEJO

Si no deseas administrar tus propias copias de seguridad de datos, asegúrate de preguntar a tu potencial proveedor de data warehouse en la nube de qué forma configura su servicio. Del mismo modo, si necesitas protección de recuperación ante desastres, asegúrate de que la arquitectura del proveedor utiliza centros separados geográficamente. Pregunta también si tu proveedor ofrece su solución a través de varios proveedores de la nube en caso de que un desastre requiera que cambies a una instancia de tu data warehouse en otra nube.

- » Factores que afectan al rendimiento
- » Elegir una solución que garantice la protección y seguridad de los datos
- » Calcular los ahorros en costes administrativos

Capítulo 5

Comparación entre distintas soluciones de data warehouse en la nube

La creciente adopción de la nube ha provocado que los proveedores locales existentes y los que acaban de acceder al mercado ofrezcan versiones en la nube de sus productos de data warehouse. Por supuesto, no hay dos soluciones iguales. En este capítulo, explicamos algunas de las diferencias y qué buscar en los almacenes de datos en la nube.

Distintas soluciones para el data warehouse en la nube

Las siguientes soluciones en la nube ofrecen capacidades de data warehouse significativamente diferentes:

- » **Infraestructura como servicio (IaaS):** requiere que el cliente instale software tradicional de data warehouse en ordenadores proporcionados por el proveedor de la plataforma en la nube. El cliente gestiona todos los aspectos del hardware de la nube y el software para el data warehouse. Las capacidades del data warehouse son idénticas a las del mismo software implementado en equipos locales.

- » **Plataforma como servicio (PaaS):** con esta solución híbrida, el proveedor del data warehouse proporciona el hardware y el software como un servicio en la nube, y el proveedor gestiona la implementación del hardware, la instalación del software y su configuración. El cliente administra, ajusta y optimiza el software.
- » **Software como servicio (SaaS):** el proveedor del data warehouse proporciona todo el hardware y software, incluidos todos los aspectos de la administración del hardware y el software. Normalmente se incluyen en el servicio las actualizaciones de software y hardware, seguridad, disponibilidad, protección de datos y optimización.

En todos estos casos, la tarea de comprar, implementar y configurar el espacio del centro de datos, y el hardware que necesita el data warehouse, se transfiere del cliente al proveedor. Más allá de esa ventaja, los beneficios y los inconvenientes de las diferentes ofertas varían desde la facilidad de uso hasta la seguridad y la disponibilidad.



RECUERDA

Si un proveedor de data warehouse simplemente proporciona acceso a su data warehouse tradicional a través de la nube, es probable que la solución se parezca a su arquitectura y funcionalidad originales locales.

Comparación de arquitecturas

Muchos proveedores ofrecen un data warehouse en la nube originalmente diseñado e implementado para entornos locales. Estas arquitecturas tradicionales se crearon mucho antes de que existiera la nube y sus beneficios surgieron como una opción viable. Alternativamente, cualquier solución de data warehouse creada para la nube debe aprovechar los beneficios de la nube (consulta la figura 5-1). Para encontrar una solución basada en una arquitectura optimizada para la nube, busca las siguientes características:

- » Almacenamiento centralizado para todos los datos
- » Escalado independiente de recursos informáticos y de almacenamiento
- » Concurrencia casi ilimitada sin competir por los recursos
- » Carga y consulta de datos simultánea sin que se vea afectado el rendimiento
- » Réplica de datos en múltiples regiones y nubes para mejorar la continuidad del negocio y simplificar la expansión
- » Uso compartido de los datos sin configurar API ni establecer procedimientos de extracción, transferencia y carga engorrosos
- » Un servicio de metadatos robusto que abarque todo el sistema. Los *metadatos* son datos sobre otros datos, como el tamaño del archivo,

el autor y cuándo se creó. Una arquitectura optimizada para la nube también aprovecha el almacenamiento como servicio, donde el data warehouse se amplía y reduce de forma automática y transparente para el usuario. El data warehouse diseñado para arquitecturas más antiguas es costoso y tiene una escalabilidad limitada.

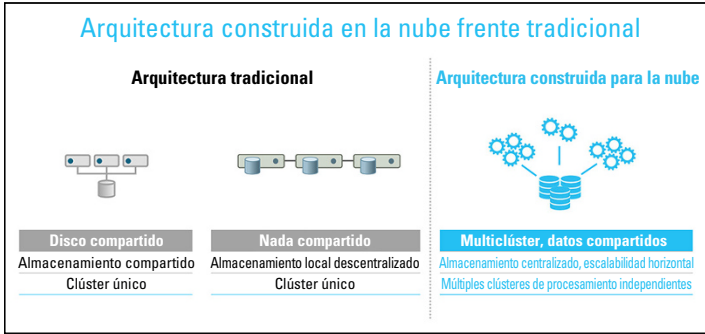


FIGURA 5-1: cómo una arquitectura optimizada para la nube agiliza el rendimiento.

Evaluar la gestión de la diversidad de datos

Un factor clave que impulsa la adopción del data warehouse en la nube proviene del creciente volumen de datos que se origina en la nube, fuera del centro de datos de una empresa. En la mayoría de los casos, estos datos no relacionales deben transformarse antes de cargarse en un data warehouse tradicional a nivel local o en la nube. Este enfoque añade una complejidad significativa y retrasos para acceder a los nuevos datos.

Con este mayor volumen y variedad de datos, la nube se ha convertido en un punto de integración natural. La mejor forma de abordar este problema es con un data warehouse en la nube que pueda manejar datos relacionales y no relacionales, sin tener que transformar los datos no relacionales o comprometer el rendimiento durante la carga de datos o el procesamiento de consultas.



RECUERDA

Los datos deben transformarse antes de cargarse en un almacén tradicional en la nube. Como alternativa, la organización debe comprar y mantener un sistema adicional para manejar los datos no relacionales.

Calcular la escala y elasticidad

No todos los almacenes de datos en la nube tienen el mismo tipo de elasticidad. Las soluciones avanzadas pueden ampliarse y reducirse, sobre la marcha, y sin desconectar el sistema o dejarlo en modo de solo lectura.



CONSEJO

Piensa en los inconvenientes de las soluciones que no se amplían con facilidad:

- » Un data warehouse en la nube que requiere una reconfiguración manual implica planificarse y coordinarse de manera concienzuda con el proveedor para ampliar los recursos.
- » La ampliación puede requerir tiempo de inactividad o cambiar al modo de solo lectura para redistribuir datos y reconfigurar el sistema.
- » La mayoría de las ofertas de data warehouse en la nube combinan procesamiento y almacenamiento en el mismo nodo, lo que requiere que los clientes amplíen ambas cosas cuando en realidad solo necesitan aumentar una u otra.
- » La mayoría son versiones llevadas a la nube de soluciones locales, por lo que tendrás que comprar una configuración de tamaño excesivo, pero infrutilizada, para cuando tengas picos de uso. A la larga, superarás los recursos disponibles y te enfrentarás a costosas actualizaciones.

Comparar las capacidades de concurrencia

La *concurrencia* es la capacidad de realizar dos o más tareas simultáneamente o permitir que dos o más usuarios accedan a una solución informática. En un data warehouse tradicional, los recursos fijos de procesamiento y almacenamiento limitan la concurrencia. Con la nube, sin embargo, el procesamiento y el almacenamiento no son fijos. Las arquitecturas optimizadas para la nube admiten la concurrencia de las siguientes dos formas:

- » Varios usuarios pueden consultar los mismos datos simultáneamente sin que se vea afectado el rendimiento.
- » La carga y las consultas de datos pueden realizarse simultáneamente, lo que permite cargas de trabajo múltiples y simultáneas sin contención de recursos.

Garantizar el soporte para SQL y otras herramientas

Casi toda la inteligencia empresarial, extracción, transformación y carga y las herramientas de análisis de datos pueden comunicarse con un data warehouse que sea compatible con un SQL estándar. Sin embargo, no todas las soluciones de data warehouse en la nube son totalmente compatibles con un SQL estándar. Por ejemplo, las soluciones de big data

posicionadas como «almacenes de datos en la nube» a menudo son soluciones NoSQL y solo son parcialmente compatibles con SQL o compatibles con un SQL no estándar. Aunque la compatibilidad con estas nuevas herramientas analíticas es importante, SQL sigue siendo el estándar de la industria para consultar datos. Tu data warehouse debe ser compatible con herramientas SQL para la gestión de datos, transformación de datos, integración de datos, visualización, inteligencia empresarial y otros tipos de análisis.

Comprobar el soporte de copia de seguridad/recuperación

Con las soluciones locales y muchas soluciones de data warehouse en la nube, los clientes deben proteger sus propios datos con herramientas de copia de seguridad y replicación de datos. Sin embargo, algunas soluciones de data warehouse en la nube incluyen la protección de datos como parte del servicio.



RECUERDA

Para lograr una protección óptima, busca una solución que guarde automáticamente versiones anteriores de datos o que duplique automáticamente los datos para usarlos como copia de seguridad en línea. La solución también debe permitir la recuperación de autoservicio de datos perdidos o dañados mediante la replicación en distintas regiones del mismo proveedor de la nube o a través de múltiples proveedores de la nube para garantizar la completa continuidad del negocio.

Confirmar la resistencia y disponibilidad

La *resistencia* es la capacidad del data warehouse para continuar funcionando automáticamente cuando falla un componente, red o incluso el centro de datos. La *disponibilidad* es la capacidad que tienen los usuarios para acceder al sistema en todo momento (lo que se conoce como «tiempo de actividad»). Los servicios de data warehouse en la nube varían en función de hasta qué punto es responsable el cliente de la *disponibilidad* y la *resistencia*. En el nivel más básico, un servicio de data warehouse en la nube puede requerir que el cliente maneje la monitorización del sistema para detectar y, posiblemente, impedir un fallo. El cliente también puede tener que proveer la replicación de los datos, para que esté disponible una copia duplicada del data warehouse en caso de fallo. En el otro extremo del espectro, el proveedor proporciona monitorización, replicación y conmutación por error automática como parte del servicio.

La disponibilidad también es un factor para las actualizaciones de software. Los distintos proveedores adoptan enfoques diferentes durante la actualización:

- » **Básico:** los clientes gestionan las actualizaciones y el tiempo de inactividad relacionado.
- » **Mejor:** el proveedor gestiona las actualizaciones e informa a los usuarios de las próximas actualizaciones, para que puedan planificar el tiempo de inactividad.
- » **Óptimo:** el proveedor proporciona actualizaciones transparentes sin involucrar a los usuarios ni someterlos a ningún tiempo de inactividad. El proveedor también permite a los clientes optar por las actualizaciones automáticas o no, para que puedan recibirlas cuando lo deseen.



CONSEJO

Busca la cantidad de «nuevos» en cuestión de disponibilidad que tiene la solución de data warehouse en la nube (99,9XX % de tiempo de actividad).

Optimizar el rendimiento

Una de las grandes promesas de la nube es la capacidad de tener grandes cantidades de recursos disponibles que puedes pagar solo cuando los necesitas. Busca una solución de data warehouse en la nube que pueda optimizar el rendimiento en función de la demanda y que elimine el esfuerzo administrativo para incorporar nuevos recursos.



RECUERDA

Aléjate de los almacenes de datos que interrumpen o retrasan la actividad para añadir o eliminar recursos. Algunas soluciones también requieren trabajo administrativo, incluida la redistribución de los datos y el nuevo cálculo de los metadatos.

Evaluar la seguridad de los datos en la nube

La nube suele percibirse como menos segura que el data warehouse local; sin embargo, las soluciones en la nube han ido ganando aceptación debido a los allanamientos en centros de datos físicos «seguros». Estos incidentes indican que las empresas tienen una capacidad limitada para proteger sus propios datos. Las ofertas de data warehouse en la nube transfieren la responsabilidad de la seguridad del centro de datos físico al proveedor de la solución, pero cuidado: las características de seguridad varían según los proveedores:

- » Las ofertas de data warehouse en la nube básicas proporcionan solo algunas capacidades de seguridad, dejando cosas como el cifrado,

control de acceso y monitorización de seguridad en manos del cliente.

- » Otras soluciones ofrecen características como el cifrado y los controles de acceso, que los clientes pueden elegir activar, pero dejan el sistema vulnerable si no se habilitan.
- » Las ofertas de almacenes de datos en la nube que están más orientadas al servicio incorporan características de seguridad y proporcionan cifrado, gestión de claves de cifrado, rotación de claves, detección de intrusiones y mucho más como parte del servicio.

Trabajo de administración

Los almacenes de datos tradicionales requieren una cantidad significativa de tiempo, esfuerzo y conocimientos por parte del cliente. Uno o más administradores de bases de datos deben llevar a cabo los parches y actualizaciones del software, creación de particiones y reparticiones de los datos, administración de índices, administración de cargas de trabajo, actualizaciones estadísticas, administración y monitorización de seguridad, copias de seguridad y replicación, y ajuste y reconfiguración de consultas entre otras cosas.

A nivel básico, una solución de data warehouse en la nube que se basa en tecnología local más antigua aún requiere que el cliente administre todos estos aspectos. Las nuevas ofertas de data warehouse reducen o eliminan gran parte de esta sobrecarga administrativa mediante nuevos diseños y automatización.

Compartir datos de forma segura

Muchas empresas pueden mejorar sus operaciones aprovechando los repositorios de datos, servicios y transmisiones de terceros. Los métodos tradicionales para compartir datos, como FTP, API y el correo electrónico, requieren que copies los datos y los envíes a los consumidores. Estos métodos engorrosos, costosos y arriesgados se basan en compartir datos estáticos, que se convierten en obsoletos rápidamente y deben actualizarse continuamente con versiones más actuales. En el capítulo 6 se detalla cómo un data warehouse construido en la nube permite compartir datos activos de manera gobernada y segura.



CONSEJO

Los sólidos métodos actuales para compartir datos te permiten intercambiar datos activos sin moverlos de un lugar a otro.

Permitir la replicación global de los datos

La *replicación de datos* crea múltiples copias de tus datos en la nube. Tener este tipo de huella global no solo es esencial para la recuperación ante desastres y la continuidad del negocio, sino que también es útil si deseas compartir datos con una base de clientes de todo el mundo sin configurar canalizaciones de extracción, transferencia y carga en las distintas regiones. Los principales proveedores de almacenes de datos te permiten compartir fácilmente datos entre distintas regiones geográficas y en varias nubes, como Amazon Web Services (AWS), Microsoft Azure y Google Cloud Platform (GCP). Estas capacidades de replicación global amplían tus mercados, facilitan la participación de socios y permiten un ecosistema más completo para el análisis y el uso compartido de los datos.

Asegurar el aislamiento de las cargas de trabajo

Un factor clave en la velocidad y el rendimiento de un data warehouse es su capacidad para aislar las cargas de trabajo. Para ser eficaz, el data warehouse en la nube debe configurar fácilmente varios grupos de recursos informáticos (de diferentes tamaños) para separar las cargas de trabajo de los usuarios y los procesos que deben ejecutarse simultáneamente. Esto elimina la contención y proporciona recursos del tamaño adecuado a cada carga de trabajo. Lo ideal es que estas cargas de trabajo separadas accedan a los mismos datos simultáneamente y se activen y desactiven fácilmente, en función de las necesidades.

Habilitar todos los casos de uso

En los entornos tradicionales, los diferentes sistemas de datos manejan diferentes casos de uso: un data warehouse para informes operativos, *data marts* para informes y análisis de distintos departamentos de la empresa, lagos de datos para la exploración de datos y herramientas especializadas para actividades como el análisis predictivo. Cada uno de estos casos de uso necesita hardware, una copia de los datos y una administración individual entre otras cosas.

Para reunir estos casos de uso tan diversos en la nube, un data warehouse debe poder trabajar con formas rápidas y eficientes de clonar múltiples copias de tablas, esquemas y bases de datos sin generar el dolor de cabeza y el coste de almacenamiento de las formas de duplicación de datos tradicionales. Un data warehouse en la nube también debe facilitar la recuperación de los errores o problemas creados por las tareas de transformación de datos con características como el viaje en el tiempo, que permite el acceso simple y la reversión a versiones anteriores de datos.

- » Reconocer la importancia de compartir datos
- » Establecer una arquitectura eficiente para compartir datos
- » Aprovechar las oportunidades para compartir datos

Capítulo 6

Poder compartir datos

Compartir datos consiste en proporcionar acceso a los datos, tanto dentro de una empresa como entre empresas que consideran que tienen activos valiosos para compartir. Una organización que pone a disposición sus datos, o los comparte, es un *proveedor de datos*. Una organización que quiere usar los datos compartidos es un *consumidor de datos*. Cualquier organización puede ser un proveedor de datos, un consumidor de datos o ambas cosas.

Además de todos los datos que las organizaciones generan y comparten internamente, muchas mejoran sus operaciones aprovechando los repositorios, servicios y flujos de datos de terceros. Por ejemplo, una organización de servicios financieros podría aprovechar varios indicadores de mercado, financieros y económicos para crear mejores modelos de datos, lo que a su vez ayudará a que cree nuevas ofertas de productos para sus clientes.

Existe una gran cantidad de valor potencial que se puede liberar con las crecientes fuentes de datos del mundo, internamente y a través de mercados e intercambios externos. Hasta hace poco, sin embargo, no existía tecnología para compartir datos sin una cantidad significativa de riesgo, coste, quebraderos de cabeza y retrasos. Aunque el uso comercial de uso compartido de los datos ha existido durante casi un siglo, hasta la fecha todos los métodos han sido restrictivos. Imagina las posibilidades que existirían si todas las organizaciones pudieran tener acceso, a medida que los necesitan, a datos activos listos para usar y pudieran hacer un uso inmediato de ellos. El proveedor de datos ya no tendría que deconstruir los datos, no se trasladarían al consumidor de datos, ni este tendría que reconstruirlos. Estarían accesibles al instante y listos para usarse dentro de un entorno seguro y controlado.

Afrontar los retos técnicos

Los métodos tradicionales para compartir datos, como el protocolo de transferencia de archivos (FTP), el almacenamiento en la nube (Amazon S3, Box, Dropbox y otros), las interfaces de programación de aplicaciones (API) y el correo electrónico requieren que hagas una copia de los datos compartidos y la envíes a tus consumidores de datos. Estos métodos engorrosos, costosos y arriesgados producen datos estáticos, que rápidamente se vuelven obsoletos y deben actualizarse con versiones más actuales, lo que requiere mover y administrar los datos de manera constante.

Las nuevas tecnologías para compartir datos permiten a las organizaciones compartir fácilmente segmentos de sus datos y recibir datos compartidos de una manera segura y controlada. No se necesita trasladar los datos, tecnología de extracción, transformación y carga ni actualizaciones constantes para mantener los datos actualizados. No es necesario transferir datos a través de una FTP ni configurar API para enlazar aplicaciones. Debido a que los datos se comparten en lugar de copiarse, no se requiere almacenamiento adicional en la nube. Con esta nueva arquitectura, los proveedores de datos pueden publicar datos de manera fácil y segura para estudio, consulta y enriquecimiento instantáneos por parte de los consumidores de datos, como se muestra en la figura 6-1.

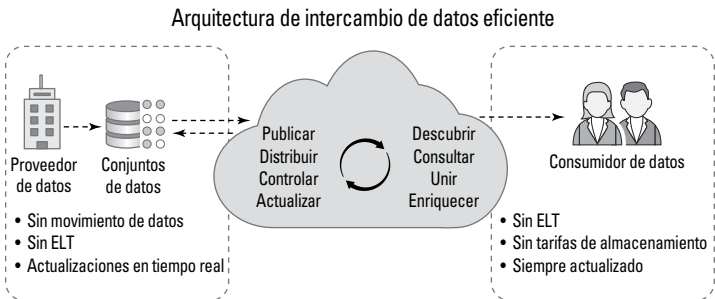


FIGURA 6-1: una arquitectura eficiente para compartir datos en tiempo real.

Un data warehouse multi-tenant o «multiinquilino» construido en la nube es la plataforma ideal para un servicio que permita compartir datos, ya que permite a los miembros autorizados de un ecosistema en la nube aprovechar las versiones activas de solo lectura de los datos. Los proveedores de datos pueden compartir datos con proveedores, socios de la cadena de suministro, socios de logística, clientes y otros muchos agentes. Estas soluciones construidas en la nube aprovechan los últimos avances en informática y data warehouse en la nube. En lugar de transferir físicamente los datos a los consumidores internos o externos, el

almacén permite el acceso de solo lectura a una parte gobernada del conjunto de datos activos a través de SQL.

Compartir datos con éxito

La mayoría de las organizaciones que se embarcan en un viaje para compartir datos siguen una progresión familiar:

1. **Colaboración interna:** los datos se comparten dentro de la empresa entre sus unidades de negocio y filiales, mejorando la colaboración y desglosando los silos de datos.
2. **Conocimientos empresariales:** tener datos más completos mejora la colaboración y genera mejores conocimientos empresariales a medida que compartir datos se convierte en la norma.
3. **Análisis de clientes:** la compañía crea análisis orientados al cliente para mejorar el valor de un producto o servicio, el primer paso hacia la monetización de datos.
4. **Análisis avanzados:** a medida que los clientes solicitan más datos, la compañía desarrolla servicios de análisis personalizados para proporcionarles a los clientes información enriquecida a partir de sus datos.
5. **Servicios de datos:** la compañía aprovecha los conjuntos de datos internos para proporcionar también a los clientes servicios de aumento de datos, como modelado de datos, enriquecimiento de datos y análisis de datos.
6. **Intercambio de datos:** la compañía busca formas de mejorar sus productos de datos mediante el suministro de datos externos y ofreciendo sus productos de datos a un público más amplio, generalmente, a través de un mercado de datos o mediante el intercambio de datos.

Monetizar los datos

La mayoría de las organizaciones ya comparten datos o planean hacerlo, pero podrían pasar por alto cómo monetizar sus datos. Existe un mercado inmenso que está creciendo con gran rapidez para monetizar los datos. En el informe «Predicciones de 2019 para la transformación digital» de IDC, la consultora predijo que el 80 % de las empresas crearía capacidades de gestión y monetización de datos para 2020, y que para 2023, el 95 % de las entidades habrá incorporado nuevos conjuntos digitales de indicadores clave de rendimiento (KPI).



CONSEJO

Con la arquitectura correcta para compartir datos, puedes analizar fácilmente más datos para descubrir nuevos productos, servicios y oportunidades de mercado.

MAXIMIZAR LAS OPORTUNIDADES DE INGRESOS



ESTUDIO DE CASOS

Environics Analytics es una de las principales compañías de análisis de datos de los Estados Unidos. Para facilitar información basada en datos a más de 3000 clientes, Environics ingiere y analiza grandes cantidades de datos demográficos, de ubicación y de consumidores.

Hace poco, Environics trasladó estas actividades analíticas a un data warehouse construido en la nube que puede manejar cualquier cantidad de datos y cualquier cantidad de cargas de trabajo. Un servicio de intercambio de datos integrado permite a los clientes descubrir y obtener nuevos datos al instante. Según Sean Howard, vicepresidente senior de desarrollo de productos de Environics, contar con un servicio seguro para compartir datos ofrece un mecanismo de entrega de datos muy cómodo y genera inmensas oportunidades para aumentar los ingresos. La plataforma en la nube se amplía o reduce rápidamente para satisfacer las necesidades analíticas de cada usuario, sin la ayuda del equipo de TI.

Anteriormente, los científicos de datos de Environics almacenaban los conjuntos de datos en sus ordenadores y compartían productos terminados con clientes a través de un servidor FTP, lo que creaba confusión interna e inhibía el crecimiento. La exploración de enormes conjuntos de datos que contienen miles de millones de filas de eventos exigía el apoyo constante del equipo de TI para instalar hardware, crear entornos de servidor SQL, optimizar el rendimiento de las consultas y monitorizar el uso de los recursos de almacenamiento y procesamiento.

Ahora, al tener un entorno analítico que se amplía según sea necesario, los científicos de datos pueden crear prototipos de grandes conjuntos de datos procedentes de cualquier industria, fuente o tipo de archivo. Pueden convertir miles de millones de puntos de datos sin procesar en productos de datos viables. El servicio seguro para compartir datos aumenta la lealtad del cliente, reduce los costes de cumplimiento y elimina las transferencias de archivos innecesarias, al tiempo que simplifica drásticamente la administración de las versiones.

Las empresas minoristas, bancos, cooperativas de crédito, empresas inmobiliarias, organizaciones sin fines de lucro y agencias gubernamentales utilizan el intercambio de datos para tomar decisiones informadas sobre consumidores y mercados con mayor facilidad. Environics ahora está haciendo pruebas con los datos de Internet of Things (IoT) y otras fuentes de big data, gracias a un servicio continuo de ingestión de datos que agiliza la carga de datos y permite llevar a cabo análisis casi en tiempo real. «La participación en el intercambio de datos impulsará el crecimiento comercial real y nos ayudará a poner nuestros datos delante de más clientes potenciales», dijo Howard.

- » Reforzar la recuperación ante desastres y la continuidad del negocio
- » Habilitar la portabilidad entre nubes, sin limitarse a un solo proveedor
- » Usar iniciativas de expansión global
- » Simplificar la seguridad y la administración en entornos multinube

Capítulo 7

Maximizar las opciones con una estrategia multinube

Contar con un data warehouse que pueda abarcar varias regiones y varias nubes ofrece enormes ventajas para el intercambio de datos, la continuidad del negocio y la penetración geográfica. Según el informe «Estado de la nube en 2019» de Flexera, el 84 % de las organizaciones tienen una estrategia multinube, lo que refleja las realidades del mercado. Ya sea Amazon Web Services, Microsoft Azure o Google Cloud Platform, cada servicio en la nube responde a necesidades ligeramente diferentes.

Para las organizaciones que desean tener un alcance global con su data warehouse, tiene sentido aplicar una estrategia a través de las distintas nubes: permite el movimiento libre y seguro de los datos a cualquier parte del mundo al mismo tiempo que te permite seleccionar los proveedores de almacenamiento en la nube que mejor se ajustan a tus necesidades. Por ejemplo, cada departamento de tu organización puede tener necesidades distintas en relación con la nube. En lugar de exigir que todas las unidades de negocio usen el mismo proveedor, una estrategia multinube permite que cada unidad use la nube que funcione mejor para esa unidad. Si esta flexibilidad te parece importante, busca un proveedor que admita múltiples entornos de nube y que ofrezca soporte a través de las distintas nubes.

¿Qué significa «a través de las nubes»?

«*Multinube*» significa que puedes almacenar tus datos en varias nubes diferentes. «*A través de las nubes*» significa que puedes acceder a los datos de todas esas nubes al mismo tiempo, migrar sin problemas las operaciones analíticas de una nube a otra y compartir datos entre las distintas nubes. Este es el santo grial del data warehouse en la nube porque no tienes que limitarte a un solo proveedor de la nube. ¿Por qué es esto importante?

- » Es una ventaja estratégica para las compañías globales porque no todos los proveedores de la nube operan en todas las regiones.
- » Es útil si adquieres una empresa que se ha estandarizado en una nube diferente de la que tú estás utilizando.
- » Si tienes pensado compartir o monetizar tus datos, expandirás tu mercado objetivo si tienes una plataforma de gestión de datos unificada que abarque distintas regiones y nubes.

En los apartados siguientes, revisamos las tecnologías que hacen posible un data warehouse a través de las nubes.



CONSEJO

Trabaja con un proveedor de data warehouse que haya hecho la dura labor de resolver las diferencias entre las configuraciones de la nube y que haya creado su solución en una base de código común que abarque todas las nubes.

Aprovechar la replicación global

La *replicación de datos* es el proceso de almacenar datos en más de una ubicación para garantizar su disponibilidad cuando se produce una interrupción a nivel regional. También es la tecnología fundamental que te permite compartir datos entre regiones y nubes. Los almacenes de datos necesitan tecnología avanzada de replicación de datos para maximizar las opciones de implementación regional, permitir la continuidad del negocio y expandir las operaciones en todo el mundo.

Tu plataforma de data warehouse debe hacer posible la replicación a través de distintas regiones y distintas nubes, sin reducir el rendimiento de las operaciones con tus datos primarios.

Minimizar la interrupción del servicio

La replicación del data warehouse a través de las nubes es importante en situaciones de recuperación ante desastres que resultan críticas para la empresa. En caso de una interrupción, te garantiza poder reanudar al

instante las actividades de procesamiento de datos sin que se produzca ningún tiempo de inactividad (consulta la figura 7-1). Sin embargo, sin la tecnología de replicación de datos adecuada, la restauración de las copias de seguridad geográficas de grandes almacenes de datos puede llevar horas o días. ¿Se cumplirán tus objetivos en cuanto al tiempo de recuperación?

Replicación a través de regiones y nubes

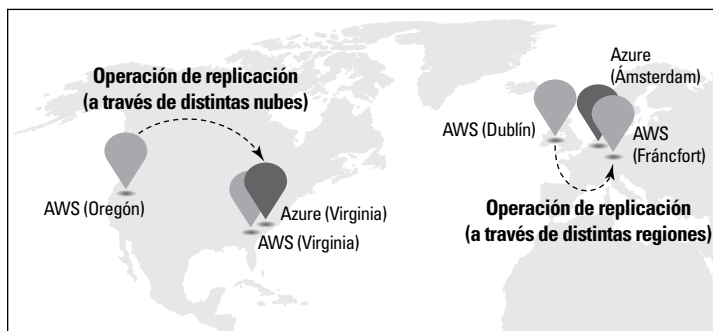


FIGURA 7-1: la replicación de datos global garantiza la continuidad del negocio durante las interrupciones.

Pregunta a tu proveedor de data warehouse si cuenta con un acceso y recuperación instantáneos para bases de datos de cualquier tamaño, en cualquier nube y en cualquier región. Si se produce un desastre en algún lugar concreto del mundo, debes poder acceder de inmediato a los datos replicados en una región o servicio en la nube diferente. Averigua si tu proveedor de data warehouse replica las bases de datos y las mantiene sincronizadas a través de las distintas plataformas y regiones de la nube.

Compatibilidad con varias nubes

La portabilidad de datos es una dificultad generalizada para todas las organizaciones que tienen grandes cantidades de datos. Cada proveedor de nube pública tiene niveles diferentes de penetración regional. Mover datos y cargas de trabajo entre regiones geográficas y nubes es más fácil con una arquitectura creada para funcionar través de distintas nubes.

La portabilidad de los datos simplifica el cumplimiento normativo si tu industria requiere que tus datos permanezcan dentro de un determinado país o región. En este caso, también es más fácil la fusión o adquisición de otra compañía que pueda tener un proveedor de nube diferente.

Cumplir las obligaciones sobre soberanía de datos

A medida que tu empresa crece, es posible que desees ubicar tus operaciones de procesamiento de datos en las regiones en las que prestas servicios. Tener una estrategia multinube te brinda la flexibilidad de seleccionar la nube más sólida en cada región, de forma que puedas configurar una arquitectura que minimice la latencia, cumpla los requisitos de residencia geográfica y cumpla las obligaciones sobre soberanía de datos. Podrás expandir tus operaciones a regiones remotas sin sacrificar el acceso a los datos, y descubrirás el valor de una única fuente de verdad para toda tu organización.

La replicación de datos también facilita el intercambio y la monetización de los datos, así como la incorporación de socios en el intercambio, todo ello al tiempo que se mantiene el principio fundamental del uso compartido de los datos: los datos existen a nivel local en una fuente de datos única, desde la cual se accede a ellos sin que tengan que moverse.

Simplificar la seguridad

Al trabajar con varias nubes, ¿cómo te aseguras de que se apliquen las mismas configuraciones y técnicas de seguridad a todos tus proveedores de la nube? ¿Tendrás que resolver las diferencias en las pistas de auditoría y registros de eventos? ¿Tendrán que lidiar tus expertos en ciberseguridad con diferentes conjuntos de reglas o ajustar los múltiples sistemas de administración de claves para cifrar los datos? Una base de código unificado que abarque todas las plataformas en la nube simplifica todas estas operaciones. No necesitarás contratar a personas con habilidades únicas ni que tengan que familiarizarse con los matices de las distintas nubes.



RECUERDA

La tecnología de replicación avanzada te permite compartir datos fácilmente entre muchas regiones y entre nubes de distintos proveedores, sin configurar canalizaciones de datos, copiar datos ni resolver diferencias de seguridad. Esto amplía tus mercados, facilita la participación de socios y te brinda un ecosistema sólido para analizar y compartir datos.

- » Establecer una seguridad de los datos integral
- » Cumplir las normas de privacidad
- » Verificar atestaciones y certificaciones
- » Mejorar la retención, protección y disponibilidad de los datos

Capítulo 8

Asegurar tus datos

Datos sobre la seguridad en la nube: En la mayoría de los casos, tus datos están más seguros en la nube que en tu propio centro de datos. Según una encuesta de Deloitte para ejecutivos de TI de 2019, creada por un equipo en el que se encontraban Tom Davenport, Ashish Verma y David Linthicum, más del 90 % de las organizaciones mantiene sus datos principalmente en plataformas en la nube. Otro de los hallazgos de la encuesta fue que la seguridad y el gobierno de los datos fueron los principales impulsores para que las organizaciones migraran sus datos a la nube.

Los proveedores de la nube SaaS prestan servicio a miles o incluso millones de clientes. Pueden permitirse los recursos necesarios para proporcionar seguridad de extremo a extremo de nivel industrial para los datos. Sin embargo, no todos los proveedores de servicios en la nube se esfuerzan por proteger tus datos. Si te fijas, verás que las capacidades de seguridad varían considerablemente.

Analizar los aspectos clave

Proteger tus datos y cumplir con las regulaciones pertinentes debe ser un aspecto fundamental de la arquitectura, implementación y funcionamiento de un servicio de data warehouse en la nube. Todos los aspectos del servicio deben centrarse en proteger tus datos como parte de una estrategia de seguridad multicapa que tenga en cuenta las amenazas de seguridad actuales y en desarrollo. Esta estrategia debe abordar las interfaces externas, el control de acceso, el almacenamiento de los datos

y la infraestructura física además de la monitorización integral, las alertas y las prácticas de ciberseguridad verificables.

Cifrado de datos predeterminado

Cifrar los datos significa aplicar un algoritmo de cifrado para traducir el texto claro en texto cifrado. Esto es fundamental para la seguridad. Cifra los datos desde el momento en que salgan de tus instalaciones, durante su transferencia por Internet y hasta su llegada al almacén: cuando se guarden en disco, cuando se trasladen a una ubicación provisional, cuando se coloquen en un objeto de base de datos y cuando se almacenen en caché en un data warehouse virtual. Los resultados de las consultas también deben cifrarse. Todo esto debería estar incorporado y no ser opcional.

El proveedor también debe proteger las claves de descifrado que decodifican tus datos. Los mejores proveedores de servicios emplean el cifrado de 256 bits del AES con un modelo de clave jerárquica. Este método cifra las claves de cifrado e propicia la rotación de claves que limita el tiempo durante el que se puede utilizar cualquier clave individual.



RECUERDA

Tus datos se encuentran probablemente en muchos lugares. Debes proteger y controlar el flujo de datos en cada punto. Todos los datos deben cifrarse de extremo a extremo y automáticamente, en tránsito y en reposo.

Aplicar el control de acceso

Asegurar los datos es solo un aspecto de la seguridad integral. Los fallos en la seguridad de los datos se suelen producir porque los usuarios eligen contraseñas poco seguras y procedimientos de autenticación rudimentarios. Un servicio de data warehouse en la nube siempre debe autorizar a los usuarios, autenticar las credenciales y otorgar acceso solamente a los datos a los que los usuarios están autorizados a acceder.

El punto de partida es el *control de acceso basado en roles*, que garantiza que los usuarios solo puedan acceder a los datos que se les permite ver. El control de acceso debe aplicarse a todos los objetos de la base de datos, incluidas las tablas, esquemas y cualquier extensión virtual del data warehouse. Para lograr la máxima comodidad y seguridad, tu data warehouse en la nube también debe proporcionar una *autenticación multifactor*, lo que requiere una verificación secundaria, como un código de seguridad enviado al teléfono móvil de un usuario que solo puede usarse una vez.

Los procedimientos de inicio de sesión único y la *autenticación federada* facilitan que las personas inicien sesión en el servicio de data warehouse

directamente desde otras aplicaciones aprobadas. La autenticación federada centraliza la gestión de identidades y los procedimientos de control de acceso, lo que facilita que tu equipo administre los privilegios de acceso de los usuarios.



CONSEJO

El proveedor de tu data warehouse en la nube no debe tener acceso a datos de clientes sin cifrar, a menos que le otorgues ese acceso explícitamente.

Monitorización de parches, actualizaciones y redes

Los parches de software y las actualizaciones de seguridad deben instalarse en todos los componentes de software pertinentes en cuanto dichas actualizaciones estén disponibles. El proveedor también debe implementar pruebas de seguridad periódicas (también conocidas como pruebas de penetración) que lleve a cabo una empresa de seguridad independiente para verificar las vulnerabilidades de manera proactiva.

Las medidas de seguridad física en el centro de datos deben incluir controles de acceso biométricos, vigilantes armados y videovigilancia para garantizar que nadie acceda de forma no autorizada. Es preciso controlar todavía más todos los equipos físicos y virtuales mediante rigurosos procedimientos de software para auditorías, monitorización y alertas. Como protección adicional, las herramientas de monitorización de la integridad de archivos (FIM) aseguran que no se manipulen los archivos críticos del sistema, y las listas blancas de direcciones IP te permiten restringir el acceso al data warehouse solo a redes de confianza. Una lista blanca es una lista de direcciones de correo electrónico o nombres de dominio desde los que un programa de bloqueo de correo electrónico permitirá recibir mensajes.

Los «eventos» de seguridad, generados por los sistemas de monitorización de ciberseguridad que vigilan la red, deben registrarse automáticamente en un sistema de gestión de eventos e información de seguridad (SIEM) resistente a la manipulación. Se deben enviar alertas automáticas al personal de seguridad cuando se detecte actividad sospechosa.

Asegurar la protección, retención y redundancia de los datos

En caso de un accidente, debes poder restaurar o consultar al momento versiones anteriores de tus datos en una tabla o base de datos en un periodo de retención especificado, según lo establezca el acuerdo de nivel de servicio (ANS) que hayas contratado con el proveedor del data warehouse en la nube. Una estrategia de retención de datos completa

debe ir más allá de la duplicación de los datos dentro de la misma región o zona de la nube: debe replicar esos datos entre múltiples zonas de disponibilidad para ofrecer redundancia geográfica. Opcionalmente, la conmutación por error automática a estas otras zonas puede garantizar operaciones comerciales sin interrupción.

Exigir el aislamiento de inquilinos

Si tu proveedor de data warehouse utiliza un entorno multiinquilino en la nube, en el que muchos clientes comparten la misma infraestructura física, asegúrate de que cada cliente tenga un data warehouse virtual aislado del resto de los almacenes de datos. Para el almacenamiento, este aislamiento debe extenderse hasta la capa de máquina virtual: el entorno de data warehouse de cada cliente debe estar aislado del entorno del resto de los clientes, gobernado por directorios independientes y claves de cifrado únicas. Algunos proveedores ofrecen también redes privadas virtuales (VPN) y puentes dedicados desde el sistema de un cliente al data warehouse en la nube. Estos servicios dedicados garantizan que los componentes más sensibles de tu data warehouse estén completamente separados de los de otros clientes.

Garantizar el gobierno y cumplimiento

El gobierno de datos garantiza que se pueda acceder a los datos corporativos y estos se utilicen de la forma adecuada, y que las prácticas diarias de administración de datos cumplan con todos los requisitos reglamentarios pertinentes. Las políticas de gobierno establecen reglas y procedimientos para controlar la propiedad y la accesibilidad de tus datos. Los tipos de información que comúnmente se incluyen en estas directrices incluyen información de tarjetas de crédito, números de la seguridad social, fechas de nacimiento, información de la red IP y coordenadas de geolocalización.

Exigir atestaciones y certificaciones

El cumplimiento no se ciñe simplemente a prácticas sólidas de ciberseguridad. También tiene que ver con garantizar que tu proveedor de data warehouse pueda demostrar que cuenta con los procedimientos de seguridad requeridos. Poner remedio a los fallos de seguridad de los datos puede costar millones de dólares, y estos dañan permanentemente las relaciones con tus clientes.

Los informes de atestación estándar de la industria verifican que los proveedores de la nube usan los controles de seguridad apropiados. Por ejemplo, un proveedor de data warehouse en la nube necesita demostrar que monitoriza y responde adecuadamente a las amenazas e incidentes

de seguridad y que cuenta con suficientes procedimientos de respuesta ante incidentes (consulta la figura 8-1).

Seguridad en el Data warehouse acorde a los estándares de la Industria

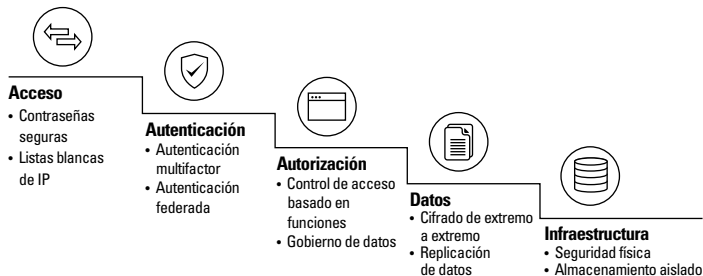


FIGURA 8-1: verifica que todo el tráfico de datos esté cifrado y sea seguro, y que tus proveedores de la nube tengan todas las certificaciones relevantes.

Además de las certificaciones de tecnología estándar de la industria, como ISO/IEC 27001 y SOC 1/SOC 2 tipo II, comprueba que tu proveedor de la nube también cumple con todas las regulaciones gubernamentales y de la industria aplicables. En función de tu negocio, esto puede incluir las certificaciones PCI, HIPAA/Health Information Trust Alliance (HITRUST) y FedRAMP.

Solicita pruebas y asegúrate de que tus proveedores te entregan una copia del informe completo para cada estándar pertinente, no solo las cartas introductorias. Por ejemplo, el informe SOC 2 tipo II verifica que se han realizado controles técnicos y administrativos apropiados de manera constante durante los últimos 12 meses. La atestación de cumplimiento de PCI-DSS indica si tu proveedor almacena y procesa adecuadamente la información de las tarjetas de crédito. Si manejas información de salud protegida, exige que tus proveedores cumplan con las directrices de la HIPAA.



CONSEJO

El cumplimiento y las atestaciones demuestran que tu proveedor de data warehouse es serio y transparente en asuntos relacionados con la seguridad.

Los proveedores de la nube también deben proporcionar pruebas de que los proveedores de software externos con los que trabajan cumplen con las normativas y llevan a cabo auditorías de seguridad periódicas. Tus datos son tan seguros como el eslabón más débil de la cadena de tecnología, por lo tanto, asegúrate de que todos los agentes cuenten con controles de seguridad sólidos y cumplan con las prácticas de seguridad estándar de la industria. Si falta alguna prueba de cumplimiento, obtén la documentación justificativa.



RECUERDA

Trabaja solo con proveedores de la nube que demuestren que respetan las prácticas de seguridad de extremo a extremo aprobadas por la industria, confirmadas por auditores independientes. Estos aspectos del cumplimiento deben abarcar tus requisitos mínimos para este repositorio de datos tan importante.

Insistir en una actitud integral frente a la seguridad

Hacer bien todo lo relacionado con la seguridad es costoso y requiere conocimientos especializados. Los fallos en los equipos, las vulnerabilidades en la red y los percances de mantenimiento pueden provocar la pérdida de datos e introducir incoherencias en tus datos. Una práctica de seguridad integral abarca muchos aspectos. Tu proveedor de data warehouse en la nube debe tener procedimientos que protejan de la destrucción accidental o intencional. Algunos proveedores ofrecen capacidades de seguridad rudimentarias, que dejan el cifrado, el control de acceso y la monitorización de la seguridad en manos del cliente. La seguridad debe ser una base del servicio de data warehouse; tú no debes tener que hacer nada extra para proteger tus datos.



CONSEJO

Es mucho más probable que un proveedor que ofrezca transparencia sobre sus certificaciones de seguridad cuente con un programa de seguridad robusto.

- » Crear un entorno de almacenamiento rentable
- » Obtener el mejor valor y rendimiento a través de la arquitectura y los precios

Capítulo 9

Minimizar los costes del data warehouse

En este capítulo, examinamos cómo ejecutar un data warehouse construido en la nube y cómo tu proveedor de data warehouse puede ayudarte a minimizar los costes a largo plazo.

Minimizar el coste de almacenamiento

Cuantos más datos puedas almacenar, más información podrás obtener. Afortunadamente, el almacenamiento en la nube de Amazon, Microsoft y Google es ahora relativamente económico, por lo que no debes sentirte limitado en cuanto a la cantidad y los tipos de datos que almacenes. Analiza bien las condiciones para asegurarte de que tu proveedor de data warehouse en la nube no esté añadiendo un recargo adicional a los costes de almacenamiento en bruto. El proveedor debe aplicarte directamente los precios de venta. Tu proveedor de almacén puede proporcionar un valor adicional al *comprimir* tus datos entre tres y cinco veces. La compresión triple significa que tienes un tercio de la cantidad de datos para almacenar, a un tercio del coste.

Analiza las condiciones del acuerdo de uso: debes pagar solo el almacenamiento que utilices y no el exceso o la capacidad de almacenamiento «reservada». Tampoco debes pagar para clonar bases de datos dentro de tu data warehouse para actividades de desarrollo y prueba. Debes poder consultar, no copiar, tus datos varias veces y, por lo tanto, no tener que pagar más por el almacenamiento.

Tu data warehouse en la nube también debe permitirte almacenar y consultar datos estructurados y semiestructurados, como JSON. Por último, busca un proveedor que ofrezca capacidades *multinube*, porque eso puede ahorrar costes futuros si migras tu data warehouse a otro entorno de almacenamiento en la nube.

Maximizar la eficiencia informática

Los recursos informáticos son más caros que los recursos de almacenamiento, por lo que el servicio de data warehouse debe permitirte ampliar cada recurso de forma independiente y facilitar la conexión exacta de los recursos informáticos que necesitas según un modelo de precios basado en el uso. El proveedor debe facturar solo los recursos que utilices, por segundos, y cancelar automáticamente los recursos informáticos cuando dejas de usarlos, para evitar costes descontrolados. El precio basado en el uso frente a la suscripción te permite elegir cómo consumes los recursos.

Los términos flexibles también deben permitirte «ajustar el tamaño» de los clústeres de procesamiento a cada carga de trabajo. Si estás ejecutando un trabajo de extracción, transferencia y carga con requisitos de procesamiento bajos, puedes usar un pequeño clúster con esa carga de trabajo en lugar de incurrir en el coste de un clúster sobreadimensionado. Si necesitas probar nuevos módulos de aprendizaje automático, puedes utilizar un clúster grande. Esto te brinda una escalabilidad precisa para cada carga de trabajo y minimiza los costes de uso. Ejecutar tu almacén costará menos que trabajar con almacenes locales y con las distintas versiones de almacenes en la nube que les siguen, que utilizan recursos enormes y producen resultados limitados. Las cargas de trabajo no se ralentizarán ni se detendrán gracias a los clústeres de procesamiento dedicados a cada carga de trabajo.

EN ESTE CAPÍTULO

- » Hacer una lista de tus necesidades de data warehouse y criterios de éxito
- » Tener en cuenta todos los factores en el coste total de propiedad
- » Probar el data warehouse antes de comprar

Capítulo 10

Seis pasos para comenzar a usar el data warehouse en la nube

En este capítulo, te guiamos en seis pasos clave para elegir un data warehouse en la nube para tu organización. El proceso comienza con la evaluación de las necesidades de tu data warehouse y finaliza con la prueba de tu mejor opción. Al final, tendrás un plan que te ayudará a elegir tu solución con plena confianza.

Paso 1: evalúa tus necesidades

El data warehouse adecuado para ti debe satisfacer tus necesidades actuales y poder satisfacer tus necesidades futuras. Por lo tanto, ten en cuenta la naturaleza de tus datos, las habilidades y herramientas ya existentes, tus necesidades de uso, los planes futuros para tu negocio y cómo un data warehouse puede hacer avanzar tu negocio mucho más de lo que imaginabas:

- » **Datos:** ¿Qué tipos de datos debe contener el data warehouse? ¿A qué velocidad se crean nuevos datos? ¿Con qué frecuencia se moverán los datos al almacén? ¿A qué datos cruciales no puedes acceder hoy?

- » **Se ajusta a las habilidades, herramientas y procesos existentes:** ¿Qué herramientas y habilidades de tu equipo servirán en función de las diversas opciones de almacenes de datos en la nube? ¿Qué procesos se verán afectados por un data warehouse en la nube?
- » **Uso:** ¿Qué usuarios y aplicaciones accederán al data warehouse? ¿Qué tipo de consultas ejecutarás? ¿A cuántos datos necesitarán acceder los usuarios y con qué rapidez? ¿Cómo variarán las cargas de trabajo con el tiempo? ¿Qué rendimiento requieren tus usuarios y aplicaciones? ¿Cuántos usuarios deben acceder al data warehouse, pero no lo hacen actualmente debido a las limitaciones de recursos?
- » **Intercambio de datos:** ¿Planeas compartir datos de manera segura en toda tu organización y con clientes o socios? Si es así, ¿qué tipos de datos compartirás?, ¿crearás un mercado o intercambio de datos para monetizar también los datos? ¿Permitirás que estos consumidores de datos accedan a datos sin procesar? o ¿enriquecerás esos datos ofreciendo también servicios de datos, como el análisis?
- » **Acceso global:** ¿Planeas almacenar datos en un almacén de objetos públicos, como Amazon S3, Microsoft Azure o Google Cloud Platform? ¿Tienes requisitos específicos de soberanía funcional, regional o de datos que te exijan mantener estas relaciones? ¿Necesitas una arquitectura a través de distintas nubes para maximizar las opciones de implementación regional, para impulsar la recuperación ante desastres o para garantizar la continuidad del negocio global?
- » **Recursos:** ¿Qué recursos humanos están disponibles para administrar el data warehouse? ¿Qué inversión deseas hacer para monitorizar y administrar la disponibilidad, el rendimiento y la seguridad? ¿Cuentas con experiencia orientada al desarrollo y pruebas del data warehouse, o un equipo de DevOps que simplifique estos procesos?

Paso 2: migrar o comenzar desde cero

Cada proyecto de data warehouse en la nube debe comenzar evaluando qué cantidad del entorno existente debe migrar al nuevo sistema y qué debe construirse desde cero para el data warehouse en la nube. Estas decisiones pueden abarcar todo, desde el diseño de los procesos de extracción, transformación y carga hasta métodos para el ciclo de vida de desarrollo del software y modelos de datos. Piensa en lo siguiente:

- » **¿Se trata de un proyecto totalmente nuevo?** Si es así, a menudo tiene sentido diseñar el proyecto para aprovechar al máximo las capacidades de un data warehouse en la nube en lugar de llevar a cabo una implementación existente con restricciones.

- » **¿Qué partes de tu sistema actual ocasionan más quebraderos de cabeza?** Una migración bien planificada podría centrarse en mover las cargas de trabajo más problemáticas al data warehouse en la nube en primer lugar. O bien, quizás prefieras migrar las cargas de trabajo más sencillas para obtener beneficios rápidamente.
- » **¿Qué aspectos de tu sistema actual tienen en cuenta las restricciones que ya no están presentes en un data warehouse en la nube?** Las herramientas y los procesos diseñados para solucionar las limitaciones de recursos, para evitar el esfuerzo disruptivo que se necesita para añadir capacidad o para optimizar costes pueden ser innecesarios con la solución de la nube correcta.
- » **¿Cómo acceden los usuarios y las aplicaciones actuales al data warehouse?** Los usuarios y aplicaciones que dependen de interfaces estándar de la industria, como SQL, y que usan herramientas estándar de extracción, transferencia y carga, e inteligencia empresarial experimentarán menos cambios al adaptarse a un nuevo enfoque.
- » **¿Qué probabilidad hay de que cambien tus requisitos en relación con los datos y análisis en el futuro?** Es probable que una solución desarrollada para evolucionar dure más de lo esperado y haga surgir nuevas oportunidades que saquen provecho a las capacidades avanzadas, como el intercambio seguro de datos y el acceso global a los datos.



RECUERDA

Si tienes un data warehouse tradicional grande y complejo, migra una pequeña parte del sistema para empezar a sentirte cómodo con el uso de un data warehouse en la nube. Después, puedes expandir sistemáticamente tu huella en la nube.

Paso 3: establece los criterios de éxito

¿Cómo vas a medir el éxito del cambio a un nuevo data warehouse en la nube? Elige requisitos empresariales y técnicos importantes. Los criterios deben centrarse en el rendimiento, la concurrencia, la simplicidad y el coste total de propiedad.



RECUERDA

Si tu nuevo data warehouse en la nube tiene capacidades que no estaban disponibles en tu sistema anterior, y esas capacidades son relevantes para evaluar el éxito empresarial y técnico de tu nueva solución, asegúrate de incluirlas.

A medida que estableces los criterios de éxito de tu nueva solución, piensa en cómo medirás ese éxito decidiendo qué criterios son cuantificables y cuáles son cualitativos, cómo medirás los criterios cuantificables y cómo evaluarás los criterios cualitativos.



CASE STUDY

SOLUCIÓN A LOS PROBLEMAS DE LATENCIA

White Ops es un proveedor líder de servicios de seguridad cibernética. A diferencia de los enfoques tradicionales que emplean análisis estadísticos, White Ops lucha contra la actividad delictiva diferenciando entre la interacción robótica y humana, trabajando para descubrir y caracterizar nuevos patrones de fraude. Este proceso constante requiere almacenar y procesar grandes cantidades de datos.

White Ops había confiado previamente en sistemas NoSQL para almacenar y procesar esos datos. Sin embargo, la latencia hasta obtener resultados era de al menos 24 horas, en función de la carga de trabajo. Cuantas más solicitudes había, mayores eran los retrasos.

Para aumentar la productividad y el rendimiento, White Ops implementó un data warehouse en la nube con SQL como lenguaje principal y lo ofreció como un servicio. El data warehouse permite a White Ops tener todos los datos en un solo lugar, ampliar el sistema con elasticidad, consultar diversos datos con lenguaje SQL estándar y acelerar la evolución de sus ofertas de prevención del fraude.

White Ops ahora puede consolidar y ampliar cantidades masivas de datos, permitir el acceso a los datos sin depender de especialistas con profundos conocimientos de programación y ayudar a sus clientes a evitar los efectos devastadores del fraude en línea.

Paso 4: evalúa las soluciones

Una vez que has definido cuáles son las necesidades de tu data warehouse y los criterios de éxito, podrás comenzar a evaluar las soluciones. A lo largo de este libro, explicamos las diferencias entre las opciones disponibles (consulta los capítulos 3, 4 y 5). Cuando compares las distintas opciones, asegúrate de que cumplan los criterios siguientes:

- »» Están preparadas para las necesidades actuales y futuras
- »» Integran datos estructurados y semiestructurados, los almacenan en un solo lugar y evitan crear silos de datos
- »» Son compatibles con las habilidades, herramientas y experiencia con las que cuentas
- »» Protegen contra la pérdida de datos y permiten una fácil recuperación de datos
- »» Aseguran tus datos con protección por contraseña y cifrado estándar de la industria

- »» Garantizan que los datos y análisis estén siempre disponibles
- »» Agilizan la canalización de datos para que los nuevos datos estén disponibles para su análisis en el menor tiempo posible
- »» Optimizan el tiempo de generación de valor, de forma que puedas obtener los beneficios de tu nuevo data warehouse lo antes posible
- »» Dedicar recursos a cargas de trabajo aisladas
- »» Comparten datos sin tener que copiar ni mover datos activos y conectan fácilmente a proveedores de datos y consumidores
- »» Replican bases de datos y las mantienen sincronizadas entre cuentas, plataformas en la nube y regiones para mejorar la continuidad del negocio, y agilizar la expansión
- »» Proporcionan clonación de base de datos de copia cero para el desarrollo y pruebas, y para admitir múltiples casos de uso, como la creación de informes, exploración de datos y análisis predictivos
- »» Facilitan la recuperación de datos perdidos debido a errores o ataques al volver a versiones anteriores de datos
- »» Amplían el procesamiento y el almacenamiento de forma independiente y automática, y aumentan la concurrencia sin ralentizar el rendimiento

Paso 5: calcula el coste total de propiedad

Si eliges un data warehouse en la nube basado en el precio, piensa en el coste total de propiedad de un data warehouse convencional, que incluye el coste de la licencia (generalmente, basado en el número de usuarios); hardware (servidores, dispositivos de almacenamiento, redes); centro de datos (espacio de oficina, electricidad, administración, mantenimiento y gestión continua); seguridad de los datos (protección por contraseña y cifrado); soluciones para garantizar la disponibilidad y la resistencia; soporte para ampliación y concurrencia; y creación de entornos de desarrollo y ensayos.

En el caso de algunas soluciones, es posible que debas pensar en los costes adicionales, como en la construcción y mantenimiento de varios *data marts*, tener múltiples copias de datos en *data marts* diferentes, formar al personal y tener múltiples sistemas (por ejemplo, SQL y NoSQL) para manejar los distintos tipos de datos, entre otras cosas.

Calcular los costes de las opciones de almacenes de datos en la nube suele ser más fácil, pero varía según los servicios que ofrezca el proveedor. Suponiendo que dejas todo en manos del proveedor eligiendo un data warehouse como servicio (DWaaS), puedes calcular el coste total de propiedad en función de la tarifa de suscripción mensual. Si optas por

una solución de infraestructura como servicio (IaaS) o plataforma como servicio (PaaS) (consulta el capítulo 5), debes añadir los costes de software, administración y servicios que no incluya la solución.



CONSEJO

Las organizaciones suelen calcular el coste total de propiedad durante la vida útil prevista del data warehouse, que generalmente es de uno a tres años. Una advertencia importante: la gente a menudo piensa que un sistema en la nube funciona las 24 horas del día, los 7 días de la semana y a gran capacidad, pasando por alto los ahorros que se pueden conseguir cuando la solución en la nube se amplía y reduce dinámicamente en respuesta a los cambios en la demanda, y solo se cobra por segundo.

Paso 6: establece una prueba de concepto

Después de analizar las diferentes opciones de almacenes de datos en la nube, ver demostraciones, hacer preguntas y reunirte con el equipo de cada proveedor, haz una prueba de concepto antes de elegir. Una prueba de concepto lleva a cabo un ensayo de una solución para determinar lo bien que satisface tus necesidades y cumple con tus criterios de éxito. Piensa en ello como si probaras un coche antes de comprarlo. Por lo general, dura uno o dos días, pero se puede llevar a cabo en el transcurso de varias semanas. Solicitas una prueba de concepto a un posible proveedor a sabiendas de que si la solución funciona satisfactoriamente, comprarás el producto. O, en el caso del data warehouse en la nube, a sabiendas de que te suscribirás al servicio.



CONSEJO

Al configurar la prueba de concepto, enumera todos los requisitos y criterios de éxito, no solo los problemas que estás tratando de resolver; incluye todo lo que es posible con una solución en la nube.

Desarrolla una lista de verificación exhaustiva de las necesidades del data warehouse y tus criterios de éxito como punto de partida. Asegúrate de que el nuevo data warehouse hace todo lo que hace tu data warehouse actual, pero mejor, y que supera los inconvenientes del sistema actual. Si llevas a cabo una prueba de concepto con varios proveedores, usa la misma lista de verificación con cada uno de ellos.

Logra una ventaja competitiva con el poder del data warehouse en la nube

Las organizaciones modernas ahora tienen acceso a cantidades de datos exponencialmente mayores que pueden analizar para obtener la información más exhaustiva posible. Además, las organizaciones desean compartir datos de manera segura —y obtener datos compartidos— a través de sus unidades de negocio, dentro de sus ecosistemas empresariales y más allá mediante el uso de intercambios de datos para monetizar esos datos. Pero acceder a esos datos plantea problemas aún mayores que siguen afectando a las plataformas tradicionales de análisis de datos. Las empresas modernas están observando actualmente que el data warehouse en la nube es la forma más eficaz y rentable de almacenar y analizar todos sus datos para todos sus usuarios empresariales. En este libro te contamos qué opciones hay en el mercado y cómo puede beneficiarse tu organización de esta nueva y emocionante tecnología.

Contenido...

- Por qué surgió el data warehouse en la nube
- Comparativa de almacenes de datos en la nube
- Cómo evaluar los distintos almacenes de datos
- Por qué son importantes la seguridad y el gobierno de los datos
- Los beneficios de una solución a través de distintas nubes
- Cómo el intercambio moderno de datos ofrece información más exhaustiva
- Estudios de caso reales



Joe Kraynak es un experto escritor de Dummies, y autor o coautor de docenas de libros sobre distintos temas. **David Baum** es un escritor independiente de temas económicos especializado en ciencia y tecnología.

Visita **Dummies.com**®

para ver vídeos, fotografías paso a paso, artículos con instrucciones o para comprar.

ISBN: 978-1-119-71455-2
Prohibida la reventa



para
dummies®



También disponible
en formato electrónico

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.