Engineer

Almacén de datos

base de datos que reúne datos de un negocio para analizarlos

Qué es

Un almacén de datos es una base de datos construida para responder preguntas sobre un negocio, en lugar de hacerlo funcionar. Los sistemas que reciben pedidos, reservan envíos, emiten facturas y registran llamadas de soporte mantienen sus propios datos, adaptados a su trabajo. Un almacén recibe copias de todos ellos, las conserva durante años y las organiza para que una pregunta que abarque varios sistemas se resuelva con una consulta, en lugar de una semana de exportaciones. Construir y gestionar uno es lo que se denomina data warehousing.

La diferencia respecto a las bases de datos de las que copia está en la carga de trabajo. Un sistema de pedidos lee y escribe unas pocas filas cada vez, miles de veces por minuto, y conserva solo el presente: cuando un cliente se muda, se sobrescribe su dirección anterior. Un almacén se escribe por lotes, se lee mediante consultas que recorren millones de filas a la vez y conserva deliberadamente el pasado, para que las cifras del año anterior todavía puedan desglosarse según cómo era el negocio entonces.

Ralph Kimball introdujo el modelado dimensional en el sector de los almacenes de datos en 1996, y Kimball Group divide un sistema así en dos salas. La sala de preparación contiene el sistema ETL, que divide en 34 subsistemas agrupados en cuatro tareas: extraer datos de las fuentes, limpiarlos y armonizarlos, entregarlos al área de presentación y gestionar todo el proceso. La sala de consulta contiene el área de presentación, que es lo que la gente consulta realmente. Se organiza por proceso de negocio, se mantiene al nivel más detallado que registra cada proceso, añade resúmenes agregados para ganar velocidad y se estructura en tablas de hechos y dimensiones. A su alrededor están las aplicaciones de BI: consultas puntuales, informes estándar, paneles de control, cuadros de mando y aplicaciones analíticas. Kimball Group llama a la sala de consulta «la cara pública del sistema DW/BI».

Esa no es la única organización posible. La propia página de arquitectura de Kimball Group describe Corporate Information Factory, un diseño radial que conserva los datos al máximo nivel de detalle en un almacén empresarial normalizado en tercera forma normal y ejecuta una segunda etapa ETL para cargar tablas dimensionales destinadas a informes, y un modelo híbrido en el que un almacén normalizado existente alimenta un área de presentación dimensional. Todas estas organizaciones comparten la separación: los sistemas que hacen funcionar el negocio no son el lugar donde se analiza.

Un almacén de datos es una función, no un producto. Puede ser una base de datos PostgreSQL o Microsoft SQL Server en un servidor que gestiona el equipo, o un servicio gestionado como Google BigQuery, que Google describe como una plataforma de datos totalmente gestionada que combina un almacén de datos en la nube con herramientas analíticas. BigQuery separa almacenamiento y cómputo para que cada uno escale por su cuenta, guarda los datos por columnas en lugar de por filas y acepta consultas en SQL conforme al estándar ANSI.

Utilice el concepto en una decisión

Aplique «¿Qué es un almacén de datos?» a una decisión real. Podemos ayudarle a relacionar el concepto con el sistema que está construyendo o comprando. Cifras de ventas, operaciones y finanzas reunidas para que dos departamentos no lleven dos respuestas distintas a la misma reunión.

Explore nuestros servicios Póngase en contacto →

Compruebe los supuestos detrás del término

Utilice la explicación y los trabajos enlazados para formular preguntas sobre su proyecto. Podemos examinar los supuestos con usted antes de elegir una herramienta. Flujo de datos geográficos 50 veces más rápido — 50x más rapidez en los flujos de procesamiento de datos geográficos

Flujo de datos geográficos 50 veces más rápido Póngase en contacto →

Por qué importa

Un negocio que hace la misma pregunta a dos sistemas obtiene dos respuestas. Ventas cuenta un cliente cuando se hace un pedido, finanzas cuando se paga una factura y soporte cada vez que llega una incidencia. Ninguno se equivoca para su propio fin y las cifras de unos no cuadran con las de los otros. El almacén es donde se acuerda una definición y se deja escrita en el proceso de carga, para que todos los informes construidos sobre él cuenten de la misma manera.

La herramienta de Kimball Group para conseguirlo es la dimensión conformada: un cliente, un producto, un calendario, mantenidos de forma central en el sistema ETL y compartidos por todas las tablas de hechos que los necesitan. Una tabla de hechos de ventas y otra de soporte que apunten a la misma dimensión de cliente pueden compararse fila a fila. Su arquitectura de bus cruza los

procesos de negocio con las dimensiones que comparten en una matriz que llama plano de arquitectura, y permite construir el almacén proceso a proceso, lo que, según el grupo, acorta el tiempo de llegada al mercado. Un equipo puede entregar pedidos este trimestre y devoluciones el siguiente sin rediseñar lo ya entregado, siempre que ambos utilicen las mismas dimensiones.

El historial es la segunda razón. Los sistemas que gestionan las operaciones del negocio conservan el estado actual; un almacén puede conservar todos los estados. La técnica de Kimball para un atributo que cambia, llamada tipo 2, añade una fila nueva a la dimensión con los valores nuevos y una clave sustituta nueva, y cada hecho cargado a partir de ese momento apunta a la nueva fila. Las ventas anteriores a la mudanza de un cliente siguen asociadas a la región antigua; las posteriores se asignan a la nueva. Al menos tres columnas adicionales, una fecha de inicio de vigencia, una fecha de fin de vigencia y un indicador de fila actual, permiten preguntar qué era cierto en cualquier fecha.

La tercera razón es la propia carga. Un informe que recorre tres años de pedidos no corresponde a la base de datos que recibe los pedidos de hoy, donde compite por la misma memoria, disco y bloqueos que el proceso de compra. Trasladar el análisis a su propia base de datos permite ajustar cada parte a su trabajo.

El precio es que un almacén es una copia. Su actualidad depende de la última carga y su corrección de las reglas de esa carga, y alguien tiene que responder por ambas: la programación que la ejecuta y las definiciones de venta, cliente y mes que ha acordado el negocio.

Cómo se aplica en la práctica

El núcleo de un almacén dimensional es una tabla de hechos cuyo grano, es decir, el nivel de detalle de cada fila, se declara primero; aquí es una fila por línea de pedido, rodeada por las dimensiones que la describen. La dimensión de cliente contiene las columnas de tipo 2, por lo que un cliente que se muda recibe una segunda fila en lugar de que se sobrescriba la anterior.

```
CREATE TABLE dim_customer (
    customer_key bigint PRIMARY KEY,   -- surrogate key, owned by the warehouse
    customer_id  text NOT NULL,        -- the key in the source system
    region       text NOT NULL,
    valid_from   date NOT NULL,
    valid_to     date,                 -- empty while the row is current
    is_current   boolean NOT NULL
```

```
);

CREATE TABLE fact_order_line (
    date_key     integer NOT NULL REFERENCES dim_date,
    customer_key bigint  NOT NULL REFERENCES dim_customer,
    product_key  bigint  NOT NULL REFERENCES dim_product,
    quantity     integer NOT NULL,
    net_amount   numeric(12, 2) NOT NULL
);

SELECT d.year, c.region, sum(f.net_amount) AS revenue
FROM fact_order_line AS f
JOIN dim_date AS d USING (date_key)
JOIN dim_customer AS c USING (customer_key)
GROUP BY d.year, c.region
ORDER BY d.year, c.region;
```

Como cada línea de pedido se cargó con la fila de cliente vigente en aquel momento, la consulta muestra los ingresos por la región donde vivía el cliente cuando compró. En cambio, unir por `customer_id` y filtrar por `is_current` recalcula todo el historial según las regiones de hoy. Ambas son preguntas legítimas y el modelo responde a cualquiera de las dos sin una segunda copia de los datos.

La promesa del servicio de diseño de almacenes de datos es exactamente esa primera razón: reunir las cifras de ventas, operaciones y finanzas en un solo lugar, para que dos departamentos dejen de llevar dos respuestas a la misma reunión. A mayor escala, la fundadora diseñó la arquitectura y gestionó 100 bases de datos de almacenes de datos en PostgreSQL, Microsoft SQL Server y Google BigQuery entre 2022 y 2024, que contenían principalmente datos GIS y series temporales, como coordenadas espaciales y lecturas de sensores. El trabajo combinó estructuras de tablas normalizadas y desnormalizadas con indexación espacial (PostGIS en PostgreSQL) y particionamiento temporal en los dos primeros motores, y tablas con particionamiento temporal y agrupación (clustering), compresión y almacenamiento columnar en BigQuery. Los tiempos de respuesta de las consultas sobre datos GIS y series temporales se redujeron entre un 40 y un 60 %, y la plataforma permitió que la organización que utilizaba esas bases de datos lanzara nuevos productos de datos, como paneles de análisis en tiempo real.

Dónde surgen los problemas

Una tabla de hechos contiene dos granos y todos los totales cuentan parte del dinero dos veces. Un pedido tiene líneas y también un coste de envío que corresponde al pedido completo. Si se pone el coste en cada línea, una suma de las líneas lo cuenta una vez por línea. El proceso de diseño en cuatro pasos de Kimball Group sitúa la declaración del grano en segundo lugar, justo después de elegir el proceso de negocio y antes de nombrar ninguna dimensión ni hecho, y su regla es que distintos granos no deben compartir una tabla de hechos. Mantenga una tabla de hechos separada a nivel de pedido para los cargos que corresponden al pedido completo, o reparta el cargo entre las líneas según una regla acordada por el negocio.

Un cliente se muda y cambia el informe del año anterior. Cuando simplemente se sobrescribe la región en la fila del cliente, todas las ventas pasadas siguen al cliente a la nueva región, y un informe aprobado en enero cuenta una historia distinta en junio. Trate los atributos cuyo historial importa con el tipo 2: cierre la fila actual y añada otra con una clave sustituta nueva, dentro de una sola transacción de base de datos para que ninguna carga vea ambas filas como actuales.

```
BEGIN;
UPDATE dim_customer
SET valid_to = current_date, is_current = false
WHERE customer_id = 'C-1042' AND is_current;
INSERT INTO dim_customer
    (customer_key, customer_id, region, valid_from, valid_to, is_current)
VALUES (90731, 'C-1042', 'North', current_date, NULL, true);
COMMIT;
```

Dos equipos construyen dos dimensiones de cliente y el almacén repite la discusión que debía resolver. Cada área temática nueva se entrega más rápido con su propia copia de la lista de clientes o productos, y cada copia se desvía: una excluye las cuentas cerradas y la otra fusiona duplicados de otra forma. Construya las dimensiones compartidas una sola vez en el sistema ETL y haga que cada tabla de hechos nueva las utilice. La matriz de bus sirve de lista de comprobación: antes de añadir un proceso de negocio, muestra a qué dimensiones existentes debe apuntar la tabla de hechos nueva.

Los informes consultan las tablas de preparación. Las copias sin procesar de las fuentes llegan primero a la sala de preparación, y una vez allí resulta tentador consultarlas directamente porque ya están cargadas. Esas tablas contienen los códigos, duplicados y lagunas propios de cada fuente, y cambian cada vez que lo hace una fuente, por lo que todos los informes construidos sobre ellas heredan el problema que el almacén debía resolver. Reserve la sala de preparación solo para el sistema ETL y dé a informes, paneles de control y analistas acceso únicamente al área de presentación, donde las dimensiones conformadas y el grano declarado hacen que las cifras coincidan.

Lecturas adicionales

- Arquitectura técnica del sistema DW/BI de Kimball — las salas de preparación y consulta, los subsistemas ETL, los metadatos y las alternativas radial e híbrida.
- Arquitectura de bus del almacén de datos empresarial — dimensiones conformadas, matriz de bus y construcción proceso a proceso.
- Proceso de diseño dimensional en cuatro pasos — proceso de negocio, grano, dimensiones y hechos, en ese orden.
- Tipo 2: añadir una fila nueva — conservar el historial de un atributo que cambia mediante una fila nueva, una clave sustituta nueva y tres columnas adicionales.
- Descripción general de BigQuery — el almacén gestionado de Google, con almacenamiento y cómputo separados y datos guardados por columnas.

Glosario

Elija según sus restricciones reales

Traiga un conjunto de datos representativo, las preguntas que debe responder y el cuello de botella actual. Podemos acordar cómo medir la calidad, la capacidad de procesamiento y la recuperación.

Vea cuánto cuesta Póngase en contacto →

Haga práctica la terminología

Pregunte cómo se aplica «¿Qué es un almacén de datos?» a su organización. Basta un ejemplo concreto de su sistema actual para iniciar la conversación.

Explore nuestros servicios

welcome@engineer.company +45 50 10 61 84 Copenhague, Dinamarca

Llévese esta página

Tarjeta de contacto

Creado con esmero y cariño por D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Derechos de autor © 2025 – actualidad · Engineer ApS · Todos los derechos reservados.

Impreso desde https://engineer.company/es/glossary/data-warehouse/