Engineer

Data Warehouse

eine Datenbank, die Geschäftsdaten zur Analyse bündelt

Was es ist

Ein Data Warehouse ist eine Datenbank, die Fragen über ein Unternehmen beantworten soll, statt dessen Betrieb abzuwickeln. Die Systeme für Bestellungen, Versandbuchungen, Rechnungen und Supportanrufe halten jeweils eigene Daten vor, passend zu ihrer Arbeit. Ein Warehouse empfängt Kopien aus all diesen Systemen, bewahrt sie über Jahre auf und ordnet sie so an, dass eine systemübergreifende Frage eine einzige Abfrage wird statt einer Woche voller Exporte. Aufbau und Betrieb heißen Data Warehousing.

Der Unterschied zu den Quelldatenbanken liegt in der Arbeitslast. Ein Bestellsystem liest und schreibt wenige Zeilen auf einmal, tausende Male pro Minute, und behält nur die Gegenwart: Zieht ein Kunde um, wird die alte Adresse überschrieben. Ein Warehouse wird schubweise beschrieben und mit Abfragen gelesen, die Millionen Zeilen gleichzeitig durchsuchen. Es erhält die Vergangenheit absichtlich, damit sich die Zahlen des Vorjahres noch so aufschlüsseln lassen, wie das Geschäft im Vorjahr aussah.

Ralph Kimball brachte 1996 die dimensionale Modellierung in die Warehouse-Branche. Die Kimball Group teilt ein solches System in zwei Räume. Im hinteren Raum liegt das ETL-System. Sie gliedert es in 34 Teilsysteme unter vier Aufgaben: Daten aus den Quellen extrahieren, bereinigen und vereinheitlichen, in den Präsentationsbereich liefern und den Gesamtprozess steuern. Im vorderen Raum liegt der Präsentationsbereich, den Menschen tatsächlich abfragen. Er ist nach Geschäftsprozessen organisiert und hält den höchsten von jedem Prozess erfassten Detailgrad vor. Aggregierte Zusammenfassungen kommen für mehr Geschwindigkeit hinzu. Seine Struktur besteht aus Fakten- und Dimensionstabellen. Darum herum liegen die BI-Anwendungen: Ad-hoc-Abfragen, Standardberichte, Dashboards, Scorecards und analytische Anwendungen. Die Kimball Group nennt den vorderen Raum „das öffentliche Gesicht des DW/BI-Systems“.

Das ist nicht die einzige Anordnung. Die eigene Architekturseite der Kimball Group beschreibt die Corporate Information Factory: eine Hub-and-Spoke-Architektur, die atomare Daten in einem normalisierten unternehmensweiten Warehouse in dritter Normalform hält und mit einem zweiten ETL-Schritt dimensionale Berichtstabellen befüllt. Daneben beschreibt sie eine Mischform, bei der ein vorhandenes normalisiertes Warehouse einen dimensionalen Präsentationsbereich versorgt. Gemeinsam ist allen die Trennung: Die Systeme, die den Geschäftsbetrieb abwickeln, sind nicht der Ort für seine Analyse.

Ein Warehouse ist eine Aufgabe, kein Produkt. Es kann eine PostgreSQL- oder Microsoft-SQL-Server-Datenbank auf einem selbst betriebenen Server sein oder ein verwalteter Dienst wie Google BigQuery. Google beschreibt BigQuery als vollständig verwaltete Datenplattform, die ein Cloud Data Warehouse mit Analysewerkzeugen verbindet. BigQuery trennt Speicherung und Berechnung, sodass beide unabhängig skalieren, speichert Daten spalten- statt zeilenweise und nimmt Abfragen in ANSI-Standard-SQL entgegen.

Das Konzept für eine Entscheidung nutzen

Nutzen Sie „Was ist ein Data Warehouse?“ für eine konkrete Entscheidung. Wir können helfen, das Konzept mit dem System in Verbindung zu bringen, das Sie entwickeln oder einkaufen. Vertriebs-, Betriebs- und Finanzdaten werden zusammengeführt, damit zwei Abteilungen nicht länger mit zwei Antworten in dieselbe Besprechung kommen.

Unsere Leistungen entdecken Kontakt aufnehmen →

Die Annahmen hinter dem Begriff prüfen

Nutzen Sie die Erklärung und die verlinkten Arbeiten, um Fragen für Ihr Projekt zu formulieren. Wir können die Annahmen gemeinsam mit Ihnen prüfen, bevor ein Werkzeug ausgewählt wird. Geografische Datenpipeline 50-mal schneller — 50x schnellere Geodaten-Pipelines

Geografische Datenpipeline 50-mal schneller Kontakt aufnehmen →

Warum es wichtig ist

Ein Unternehmen erhält auf dieselbe Frage an zwei Systeme zwei Antworten. Der Vertrieb zählt einen Kunden bei der Bestellung, das Finanzwesen bei der bezahlten Rechnung und der Support bei jedem eingehenden Ticket. Für ihren eigenen Zweck liegt keiner falsch. Zusammen passen die Zahlen aber nicht. Im Warehouse wird eine Definition vereinbart und im Ladeprozess festgeschrieben, damit jeder darauf aufbauende Bericht auf dieselbe Weise zählt.

Das Werkzeug der Kimball Group dafür ist die abgestimmte Dimension: ein Kunde, ein Produkt, ein Kalender, zentral im ETL-System gepflegt und von jeder passenden Faktentabelle gemeinsam genutzt. Eine Vertriebs- und eine Supportfaktentabelle, die auf dieselbe Kundendimension verweisen, lassen sich

Zeile für Zeile vergleichen. Ihre Busarchitektur stellt Geschäftsprozesse und gemeinsam genutzte Dimensionen in einer Matrix gegenüber, die sie als architektonischen Bauplan bezeichnet. So lässt sich das Warehouse Geschäftsprozess für Geschäftsprozess aufbauen. Laut Kimball verkürzt das die Markteinführungszeit. Ein Team kann in diesem Quartal Bestellungen und im nächsten Retouren liefern, ohne das bereits Ausgelieferte neu zu entwerfen, solange beide dieselben Dimensionen verwenden.

Der zweite Grund ist Historie. Operative Systeme halten den aktuellen Zustand, ein Warehouse kann jeden Zustand aufbewahren. Die Kimball-Technik für ein sich änderndes Attribut heißt Typ 2: Sie ergänzt eine neue Dimensionszeile mit den neuen Werten und einem neuen Surrogatschlüssel. Jeder ab diesem Zeitpunkt geladene Fakt verweist auf die neue Zeile. Umsätze vor einem Kundenumzug bleiben der alten Region zugeordnet, spätere der neuen. Mindestens drei Zusatzspalten, Gültigkeitsbeginn, Gültigkeitsende und ein Kennzeichen für die aktuelle Zeile, ermöglichen die Frage, was an einem beliebigen Datum galt.

Der dritte Grund ist schiere Last. Ein Bericht über drei Jahre Bestellungen gehört nicht auf die Datenbank, die heutige Bestellungen annimmt. Dort konkurriert er mit dem Checkout um denselben Speicher, Datenträger und dieselben Sperren. Durch die Verlagerung der Analyse in eine eigene Datenbank lässt sich jede Seite auf ihre jeweilige Arbeit abstimmen.

Der Preis ist, dass ein Warehouse eine Kopie ist. Es ist so aktuell wie sein letzter Ladevorgang und so korrekt wie die Regeln dieses Vorgangs. Jemand muss beides verantworten: den Zeitplan für den Ladeprozess und die geschäftlich vereinbarten Definitionen von Verkauf, Kunde und Monat.

Wie es in der Praxis aussieht

Das Herz eines dimensionalen Warehouses ist eine Faktentabelle, deren Granularität zuerst festgelegt wird, hier eine Zeile pro Bestellposition. Die beschreibenden Dimensionen umgeben sie. Die Kundendimension trägt die Typ-2-Spalten. Ein umziehender Kunde erhält daher eine zweite Zeile statt einer überschriebenen.

```
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;
```

Jede Bestellposition wurde mit der damals aktuellen Kundenzeile geladen. Deshalb meldet die Abfrage Umsatz nach der Region, in der der Kunde beim Kauf wohnte. Ein Join über `customer_id` mit Filter auf `is_current` ordnet dagegen die gesamte Historie den heutigen Regionen zu. Beides sind berechtigte Fragen. Das Modell beantwortet beide ohne eine zweite Datenkopie.

Das Versprechen der Leistung zur Konzeption von Data Warehouses entspricht genau dem ersten Grund: Zahlen aus Vertrieb, Betrieb und Finanzwesen an einem Ort zusammenführen, damit zwei Abteilungen nicht mehr zwei Antworten in dieselbe Besprechung bringen. Im größeren Maßstab entwarf und betrieb die Gründerin 100 Data-Warehouse-Datenbanken zwischen 2022 und 2024 auf PostgreSQL, Microsoft SQL Server und Google BigQuery. Sie enthielten vor allem GIS- und Zeitreihendaten wie räumliche Koordinaten und Sensormesswerte. Die Arbeit verband normalisierte und denormalisierte Tabellenstrukturen mit räumlicher Indizierung (PostGIS auf PostgreSQL) und Zeitreihenpartitionierung auf den ersten beiden Datenbanksystemen sowie zeitpartitionierte, geclusterte Tabellen mit Kompression und spaltenorientierter Speicherung auf BigQuery. Die Antwortzeiten von GIS- und Zeitreihenabfragen sanken um 40 bis 60 %. Die Plattform ermöglichte der Organisation, die diese Datenbanken nutzte, neue Datenprodukte wie Echtzeit-Analysedashboards einzuführen.

Wo es schiefgeht

Eine Faktentabelle hält zwei Granularitäten, und jede Summe zählt einen Teil des Geldes doppelt. Eine Bestellung hat Positionen und eine Liefergebühr, die zur ganzen Bestellung gehört. Steht die Gebühr auf jeder Position, zählt eine Summe über die Positionen sie pro Position einmal. Im vierstufigen Entwurfsprozess der Kimball Group steht das Festlegen der Granularität an zweiter Stelle, direkt nach der Wahl des Geschäftsprozesses und bevor eine Dimension oder ein Fakt benannt wird. Ihre Regel lautet, dass verschiedene Granularitäten nicht dieselbe Faktentabelle teilen dürfen. Halten Sie Gebühren für die ganze Bestellung in einer getrennten Faktentabelle auf Bestellebene oder teilen Sie die Gebühr nach einer geschäftlich abgestimmten Regel auf die Positionen auf.

Ein Kunde zieht um, und der Vorjahresbericht ändert sich. Wird die Region auf der Kundenzeile einfach überschrieben, folgt jeder frühere Verkauf dem Kunden in die neue Region. Ein im Januar freigegebener Bericht erzählt im Juni etwas anderes. Behandeln Sie Attribute mit wichtiger Historie nach Typ 2: Schließen Sie die aktuelle Zeile und ergänzen Sie eine neue mit neuem Surrogatschlüssel, in einer Datenbanktransaktion, damit kein Ladeprozess je beide Zeilen als aktuell sieht.

```
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;
```

Zwei Teams bauen zwei Kundendimensionen, und das Warehouse wiederholt den Streit, den es beenden sollte. Jeder neue Themenbereich ist mit einer eigenen Kunden- oder Produktliste schneller fertig. Jede Kopie driftet ab: Eine entfernt geschlossene Konten, die andere führt Dubletten anders zusammen. Bauen Sie die gemeinsamen Dimensionen einmal im ETL-System und lassen Sie jede neue Faktentabelle sie verwenden. Die Busmatrix dient als Checkliste: Vor dem Hinzufügen eines Geschäftsprozesses zeigt sie, auf welche vorhandenen Dimensionen die neue Faktentabelle verweisen muss.

Berichte greifen auf die Staging-Tabellen zu. Rohkopien der Quellen landen zuerst im hinteren Raum. Sobald sie geladen sind, ist es verlockend, sie direkt abzufragen. Diese Tabellen tragen aber die eigenen Codes, Dubletten und Lücken jeder Quelle und ändern sich mit ihr. Jeder darauf aufgebaute Bericht erbt so das Problem, das das Warehouse lösen sollte. Lassen Sie den hinteren Raum allein dem ETL-System. Geben Sie Berichten, Dashboards und Analysten nur den Präsentationsbereich, in dem abgestimmte Dimensionen und festgelegte Granularität die Zahlen zur Übereinstimmung bringen.

Weiterführende Lektüre

- Kimballs technische DW/BI-Systemarchitektur — hinterer und vorderer Raum, ETL-Teilsysteme, Metadaten sowie Hub-and-Spoke- und Hybridalternativen.
- Busarchitektur für unternehmensweite Data Warehouses — abgestimmte Dimensionen, Busmatrix und der Aufbau eines Geschäftsprozesses nach dem anderen.
- Vierstufiger dimensionaler Entwurfsprozess — Geschäftsprozess, Granularität, Dimensionen und Fakten, in dieser Reihenfolge.
- Typ 2: neue Zeile hinzufügen — die Historie eines sich ändernden Attributs mit neuer Zeile, neuem Surrogatschlüssel und drei Zusatzspalten erhalten.
- BigQuery-Überblick — Googles verwaltetes Warehouse mit getrennter Speicherung und Berechnung und spaltenweise gespeicherten Daten.

Glossar

Für Ihre tatsächlichen Rahmenbedingungen auswählen

Bringen Sie einen repräsentativen Datensatz, die Fragen, die er beantworten muss, und den aktuellen Engpass mit. Wir können vereinbaren, wie Qualität, Durchsatz und Wiederherstellung gemessen werden.

Preise ansehen Kontakt aufnehmen →

Die Begriffe praktisch nutzbar machen

Fragen Sie, wie „Was ist ein Data Warehouse?“ auf Ihre Organisation zutrifft. Ein konkretes Beispiel aus Ihrem bestehenden System reicht aus, um das Gespräch zu beginnen.

Unsere Leistungen entdecken

welcome@engineer.company +45 50 10 61 84 Kopenhagen, Dänemark

Nehmen Sie diese Seite mit

Kontaktkarte

Mit Sorgfalt und Liebe gestaltet von D.

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

Copyright © 2025 – heute · Engineer ApS · Alle Rechte vorbehalten.

Gedruckt von https://engineer.company/de/glossary/data-warehouse/