Engineer

Entrepôt de données

base rassemblant les données d'une entreprise pour l'analyse

Ce que c’est

Un entrepôt de données est une base construite pour répondre aux questions sur une entreprise plutôt que pour la faire fonctionner. Les systèmes qui prennent les commandes, réservent les expéditions, émettent les factures ou enregistrent les appels d’assistance gardent chacun leurs données, adaptées à leur travail. L’entrepôt en reçoit des copies, les conserve des années et les organise pour qu’une question portant sur plusieurs systèmes demande une requête plutôt qu’une semaine d’exports. Construire et exploiter un tel ensemble s’appelle l’entreposage de données.

La différence avec les bases sources est la charge de travail. Un système de commandes lit et écrit quelques lignes à la fois, des milliers de fois par minute, et ne garde que le présent : lorsqu’un client déménage, son ancienne adresse est écrasée. L’entrepôt s’alimente par lots, se lit par des requêtes parcourant des millions de lignes à la fois et garde volontairement le passé. Les chiffres de l’an dernier peuvent ainsi encore être ventilés selon la structure de l’entreprise de l’an dernier.

Ralph Kimball a apporté la modélisation dimensionnelle au secteur des entrepôts en 1996, et le Kimball Group divise le système en deux pièces. L’arrière-boutique contient le système ETL, décomposé en 34 sous-systèmes regroupés en quatre fonctions : extraire les données des sources, les nettoyer et les harmoniser, les livrer à l’espace de présentation et gérer l’ensemble. La salle d’accueil contient l’espace de présentation, celui que les gens interrogent. Il est organisé par processus métier, au détail le plus fin enregistré par chacun, avec des résumés agrégés pour accélérer la lecture et une structure en tables de faits et de dimensions. Autour se trouvent les applications BI : requêtes ponctuelles, rapports standard, tableaux de bord, tableaux de suivi des objectifs et applications analytiques. Le Kimball Group appelle cette salle « la face publique du système DW/BI ».

Ce n’est pas la seule organisation. La page d’architecture du Kimball Group décrit la Corporate Information Factory : une architecture à centre et branches (hub-and-spoke) qui garde les données atomiques dans un entrepôt d’entreprise normalisé en troisième forme normale, puis utilise un second ETL pour alimenter les tables dimensionnelles des rapports. Elle décrit aussi un hybride où un entrepôt normalisé existant fournit l’espace de présentation dimensionnel. Le point commun est la séparation : les systèmes qui font fonctionner l’entreprise ne sont pas le lieu de son analyse.

Un entrepôt est un rôle, pas un produit. Il peut s’agir d’une base PostgreSQL ou Microsoft SQL Server sur un serveur géré par l’équipe, ou d’un service géré comme Google BigQuery. Google le décrit comme une plateforme de données entièrement gérée associant entrepôt dans le cloud et outils analytiques. BigQuery sépare stockage et calcul pour les faire évoluer indépendamment, stocke par colonne plutôt que par ligne et accepte les requêtes en SQL standard ANSI.

Utilisez le concept pour prendre une décision

Utilisez « Qu'est-ce qu'un entrepôt de données ? » pour éclairer une décision réelle. Nous pouvons vous aider à relier ce concept au système que vous construisez ou achetez. Réunir les chiffres des ventes, des opérations et des finances pour que deux services ne présentent plus deux réponses à la même réunion.

Découvrir nos prestations Contactez-nous →

Vérifiez les hypothèses derrière le terme

Appuyez-vous sur l'explication et les réalisations liées pour formuler les questions de votre projet. Nous pouvons examiner les hypothèses avec vous avant le choix d'un outil. Pipeline géographique accéléré de 50 fois — 50x d’accélération des pipelines de données géographiques

Pipeline géographique accéléré de 50 fois Contactez-nous →

Pourquoi c’est important

Une entreprise posant la même question à deux systèmes obtient deux réponses. Les ventes comptent un client à la commande, la finance au paiement d’une facture, l’assistance à l’arrivée d’un ticket. Chaque comptage répond à son propre besoin, mais les chiffres ne concordent pas entre eux. L’entrepôt est le lieu où une définition est convenue et inscrite dans le chargement, pour que chaque rapport compte de la même façon.

L’outil du Kimball Group est la dimension conforme : un client, un produit, un calendrier, maintenus de façon centralisée dans l’ETL et partagés par chaque table de faits qui en a besoin. Une table de faits de vente et une d’assistance pointant vers la même dimension client se comparent ligne par ligne.

L’architecture en bus met en matrice les processus métier et leurs dimensions communes, dans ce qu’ils appellent le plan d’architecture. Elle permet de construire l’entrepôt un processus à la fois, ce qui, selon eux, raccourcit le délai de mise sur le marché. Une équipe peut livrer le volet commandes ce trimestre et le volet retours au suivant sans revoir ce qui est déjà livré, tant que les dimensions sont communes.

L’historique est la deuxième raison. Les systèmes opérationnels gardent l’état actuel ; un entrepôt peut garder tous les états. La technique Kimball Type 2 ajoute, pour un attribut changeant, une nouvelle ligne de dimension avec les nouvelles valeurs et une nouvelle clé de substitution. Chaque fait chargé à partir de ce moment pointe vers elle. Les ventes d’avant un déménagement gardent l’ancienne région ; celles d’après rejoignent la nouvelle. Au moins trois colonnes supplémentaires, date d’effet, date d’expiration et indicateur de ligne en cours de validité, permettent de demander ce qui était vrai à une date donnée.

La troisième raison est la charge. Un rapport parcourant trois années de commandes n’a pas sa place dans la base qui reçoit celles d’aujourd’hui, où il entre en concurrence avec le passage en caisse pour la mémoire, le disque et les verrous. Placer l’analyse dans sa propre base permet d’optimiser chacun des deux côtés pour son travail.

La contrepartie est que l’entrepôt est une copie. Sa fraîcheur est celle du dernier chargement, son exactitude celle des règles appliquées, et quelqu’un doit répondre des deux : le calendrier d’exécution et les définitions d’une vente, d’un client et d’un mois convenues dans l’entreprise.

À quoi cela ressemble en pratique

Le cœur d’un entrepôt dimensionnel est une table de faits dont on déclare d’abord la granularité : ici une ligne par ligne de commande, entourée des dimensions qui la décrivent. La dimension client porte les colonnes Type 2 ; on ajoute donc une seconde ligne pour un client qui déménage, au lieu d’écraser l’ancienne.

```
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;
```

Chaque ligne de commande ayant été chargée avec la ligne client qui était valide à ce moment-là, la requête donne le chiffre d'affaires par région de résidence au moment de l'achat. Joindre plutôt sur `customer_id` et filtrer sur `is_current` réexprime tout l'historique selon les régions actuelles. Les deux questions sont légitimes, et le modèle répond aux deux sans seconde copie des données.

La promesse de la prestation de conception d'entrepôts de données correspond précisément à la première raison : réunir les chiffres des ventes, des opérations et de la finance pour que deux départements cessent d'apporter deux réponses à la même réunion. À plus grande échelle, la fondatrice a conçu et exploité 100 bases d'entrepôts de données sous PostgreSQL, Microsoft SQL Server et Google BigQuery entre 2022 et 2024. Elles contenaient surtout des données SIG et des séries temporelles, comme des coordonnées spatiales et des relevés de capteurs. Le travail associait des tables normalisées et dénormalisées à l'indexation spatiale, avec PostGIS sous PostgreSQL, et au partitionnement temporel sur les deux premiers moteurs. BigQuery utilisait des tables partitionnées dans le temps et organisées par clustering, avec compression et stockage en colonnes. Les temps de réponse des requêtes sur les données SIG et les séries temporelles ont diminué de 40 à 60 %, et la plateforme a permis à l'organisation qui utilisait ces bases de lancer de nouveaux produits de données, comme des tableaux de bord d'analyse en temps réel.

Où les problèmes surviennent

Une table de faits mélange deux granularités, et chaque total compte deux fois une partie des sommes. Une commande contient des lignes, mais aussi des frais de livraison qui appartiennent à l'ensemble. Les répéter sur chaque ligne les fait compter une fois par ligne dans la somme. Le processus de conception Kimball en quatre étapes place la déclaration de granularité en deuxième, juste après le choix du processus métier et avant toute dimension ou fait. Sa règle interdit de réunir deux granularités dans une même table de faits. Gardez une table distincte à la granularité commande pour les frais globaux, ou répartissez-les entre lignes selon une règle convenue dans l'entreprise.

Un client déménage et le rapport de l'an dernier change. Si la région de la ligne client est simplement écrasée, toutes les ventes passées suivent le client dans la nouvelle région, et un rapport approuvé en janvier raconte autre chose en juin. Traitez les attributs dont l'historique compte avec Type 2 : fermez la ligne en cours de validité et ajoutez-en une avec une nouvelle clé de substitution, dans une transaction unique, pour qu'aucun chargement ne voie ces deux lignes comme valides en même temps.

```
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;
```

Deux équipes créent deux dimensions client et l'entrepôt reproduit le désaccord qu'il devait résoudre. Un nouveau domaine métier est plus rapide à mettre à disposition avec sa copie des listes de clients ou de produits, mais chaque copie dérive : l'une retire les comptes fermés, l'autre fusionne différemment les doublons. Construisez les dimensions partagées une fois dans l'ETL et imposez-les à chaque nouvelle table de faits. La matrice en bus est la liste de contrôle : avant l'ajout d'un processus, elle montre les dimensions existantes vers lesquelles la table doit pointer.

Les rapports interrogent les tables de transit. Les copies brutes arrivent d'abord dans l'arrière-boutique ; une fois chargées, il est tentant de les interroger directement. Elles portent les codes, doublons et lacunes de leurs sources et

changent avec elles. Chaque rapport hérite donc du problème que l'entrepôt devait résoudre. Réservez l'arrière-boutique à l'ETL et ne donnez aux rapports, tableaux de bord et analystes que l'espace de présentation, où dimensions conformes et granularité déclarée font concorder les chiffres.

Pour aller plus loin

- Architecture technique d'un système DW/BI Kimball — arrière-boutique et salle d'accueil, sous-systèmes ETL, métadonnées et variantes à centre et branches ou hybrides.
- Architecture en bus de l'entrepôt d'entreprise — dimensions conformes, matrice en bus et construction un processus métier à la fois.
- Conception dimensionnelle en quatre étapes — processus métier, granularité, dimensions et faits, dans cet ordre.
- Type 2 : ajouter une ligne — garder l'historique d'un attribut changeant avec une nouvelle ligne, une nouvelle clé de substitution et trois colonnes supplémentaires.
- Présentation de BigQuery — l'entrepôt géré de Google, séparant stockage et calcul et stockant les données par colonne.

Glossaire

Choisissez selon vos contraintes réelles

Présentez un jeu de données représentatif, les questions auxquelles il doit répondre et le goulot d'étranglement actuel. Nous pouvons convenir de la manière de mesurer la qualité, le débit et la reprise.

Consulter les tarifs Contactez-nous →

Rendez la terminologie concrète

Demandez comment « Qu'est-ce qu'un entrepôt de données ? » s'applique à votre organisation. Un exemple concret tiré de votre système actuel suffit pour engager la discussion.

Découvrir nos prestations

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

Emportez cette page avec vous

Fiche de contact

Créé avec soin et amour par D.

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

Droit d'auteur © 2025 – aujourd'hui · Engineer ApS · Tous droits réservés.

Imprimé depuis https://engineer.company/fr/glossary/data-warehouse/