Engineer

Analysdatalager

en databas som samlar verksamhetens data för analys

Vad det är

Ett analysdatalager är en databas som byggs för att besvara frågor om en verksamhet, inte för att driva den. Systemen som tar emot beställningar, bokar leveranser, skickar fakturor och registrerar supportärenden har var och ett sina egna data, formade för sitt eget arbete. Analysdatalagret får kopior från dem alla, sparar kopiorna i flera år och ordnar dem så att en fråga som spänner över flera system blir en enda databasfråga i stället för en veckas export till kalkylark. Ofta säger man bara datalager, på engelska data warehouse, och arbetet med att bygga och driva ett sådant kallas data warehousing.

Skillnaden mot databaserna det kopierar från ligger i arbetsbelastningen. Ett ordersystem läser och skriver några få rader åt gången, tusentals gånger i minuten, och håller bara nuet: när en kund flyttar skrivs den gamla adressen över. Ett analysdatalager skrivs i stora omgångar, läses av frågor som går igenom miljontals rader på en gång och sparar det förflutna med avsikt, så att fjolårets siffror fortfarande kan brytas ned så som verksamheten såg ut i fjol.

Ralph Kimball förde in dimensionsmodellering i branschen 1996, och Kimball Group delar ett sådant system i två rum. Back room rymmer ETL-systemet, som de delar upp i 34 delsystem fördelade på fyra uppgifter: att hämta data ur källorna, att rensa och stämma av dem, att leverera dem till presentationsområdet och att styra hela processen. Front room rymmer presentationsområdet, det som användarna faktiskt ställer frågor mot. Det är organiserat efter affärsprocess, hållet på den mest detaljerade nivå som varje process registrerar, kompletterat med aggregerade sammanställningar för snabbhetens skull och uppbyggt av fakta- och dimensionstabeller. Runt det finns BI-applikationerna: ad hoc-frågor, standardrapporter, dashboards, styrkort (scorecards) och analysapplikationer. Kimball Group beskriver front room som den del av systemet som verksamheten möter.

Det är inte den enda uppbyggnaden. Kimball Groups egen arkitektursida beskriver Corporate Information Factory, en hub-and-spoke-design där atomära data ligger i ett normaliserat analysdatalager för hela företaget i tredje normalform och ett andra ETL-steg fyller dimensionella tabeller för rapportering, och en hybrid där ett befintligt normaliserat analysdatalager förser ett dimensionellt presentationsområde. Gemensamt för alla är åtskillnaden: systemen som driver verksamheten är inte platsen där den analyseras.

Ett analysdatalager är en roll, inte en produkt. Det kan vara en PostgreSQL- eller Microsoft SQL Server-databas på en server som teamet driver själv, eller en hanterad tjänst som Google BigQuery, som Google beskriver som en fullt hanterad dataplattform som kombinerar ett datalager i molnet med analysverktyg. BigQuery håller lagring och beräkning åtskilda, så att de kan skalas var för sig, lagrar data kolumnvis i stället för radvis och tar emot frågor i ANSI-standard-SQL.

Använd begreppet i ett beslut

Använd ”Vad är ett analysdatalager?” i ett konkret beslut. Vi kan knyta begreppet till systemet ni bygger eller köper. Siffror från försäljning, drift och ekonomi samlade på ett ställe, så att två avdelningar slutar komma med två olika svar till samma möte.

Se våra tjänster Hör av er →

Granska antagandena bakom ordet

Använd förklaringen och det länkade arbetet för att formulera frågor till projektet. Vi kan undersöka antagandena tillsammans innan ett verktyg väljs. Gjorde pipelinen för geodata 50 gånger snabbare — 50x snabbare datapipelines för geodata

Gjorde pipelinen för geodata 50 gånger snabbare Hör av er →

Varför det spelar roll

Ett företag som ställer samma fråga till två system får två svar. Försäljningen räknar en kund när en order läggs, ekonomi när fakturan är betald och support varje gång ett ärende kommer in. Ingen av dem har fel utifrån sitt eget syfte, och ingen av dem stämmer med de andra. Analysdatalagret är platsen där man enas om en enda definition och skriver ner den i laddningen, så att varje rapport som bygger på den räknar på samma sätt.

Kimball Groups verktyg för det här är den gemensamma, avstämda dimensionen, på engelska conformed dimension: en kund, en produkt, en kalender, förvaltade centralt i ETL-systemet och delade av varje faktatabell som behöver dem. En faktatabell för försäljning och en för supportärenden som båda pekar på samma kunddimension kan jämföras rad för rad. Deras bussarkitektur ställer affärsprocesserna mot de dimensioner de delar, i en matris de kallar den

arkitektoniska ritningen, och den gör det möjligt att bygga analysdatalagret en affärsprocess i taget, vilket enligt Kimball Group kortar tiden till marknaden. Ett team kan leverera beställningar det här kvartalet och returer nästa utan att bygga om det som redan är levererat, så länge båda använder samma dimensioner.

Historiken är det andra skälet. Driftsystem håller det aktuella tillståndet; ett analysdatalager kan hålla varje tillstånd. Kimballs teknik för en egenskap som ändras, kallad typ 2, lägger till en ny rad i dimensionen med de nya värdena och en ny surrogatnyckel, och varje faktarad som laddas från det ögonblicket pekar på den nya raden. Försäljning från före flytten stannar i den gamla regionen; försäljning efteråt hamnar i den nya. Minst tre extra kolumner, ett datum då raden börjar gälla, ett datum då den slutar gälla och en markering för aktuell rad, gör det möjligt att fråga vad som var sant ett visst datum.

Det tredje skälet är ren belastning. En rapport som går igenom tre års beställningar hör inte hemma på databasen som tar emot dagens beställningar, där den konkurrerar med kassan om samma minne, samma diskar och samma lås. När analysen flyttar till en egen databas kan varje sida trimmas för sitt eget arbete.

Priset är att ett analysdatalager är en kopia. Det är så färskt som den senaste laddningen och så korrekt som reglerna i den, och någon måste äga båda delarna: när laddningen körs, och de definitioner av en försäljning, en kund och en månad som verksamheten har enats om.

Så ser det ut i praktiken

Kärnan i ett dimensionellt analysdatalager är en faktatabell vars granularitet, på engelska grain, bestäms först, här en rad per orderrad, omgiven av de dimensioner som beskriver den. Kunddimensionen har kolumnerna för typ 2, så att en kund som flyttar får en ny rad i stället för en överskriven.

```
CREATE TABLE dim_customer (
    customer_key bigint PRIMARY KEY,   -- surrogatnyckel, ägs av lagret
    customer_id  text NOT NULL,        -- nyckeln i källsystemet
    region       text NOT NULL,
    valid_from   date NOT NULL,
    valid_to     date,                 -- tom så länge raden är aktuell
    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;
```

Eftersom varje orderrad laddades med den kundrad som var aktuell då, visar frågan intäkterna fördelade på den region kunden bodde i när köpet gjordes. Om man i stället joinar på `customer_id` och filtrerar på `is_current` räknas hela historiken om efter dagens regioner. Båda frågorna är legitima, och modellen besvarar båda utan en extra kopia av datan.

Tjänsten Design av datalager lovar just det första skälet: siffror från försäljning, drift och ekonomi samlade på ett ställe, så att två avdelningar slutar komma med två olika svar till samma möte. I större skala utformade och förvaltade grundaren 100 datalagerdatabaser i PostgreSQL, Microsoft SQL Server och Google BigQuery mellan 2022 och 2024, främst med GIS- och tidsseriedata som spatiala koordinater och sensoravläsningar. Arbetet kombinerade normaliserade och denormaliserade tabellayouter med spatial indexering (PostGIS i PostgreSQL) och tidsseriepartitionering på de två första plattformarna, och tidspartitionerade, klustrade tabeller med komprimering och kolumnlagring i BigQuery. Svarstiderna för databasfrågor mot GIS- och tidsseriedata minskade med 40–60 %, och den optimerade infrastrukturen gjorde det möjligt för organisationen som använde dessa databaser att lansera nya datadrivna produkter som dashboards för realtidsanalys.

Var det går fel

En faktatabell blandar två granulariteter, och varje summa räknar vissa pengar två gånger. En order har rader, och den har också en fraktavgift som hör till ordern som helhet. Läggs avgiften på varje rad räknar en summa över raderna den en gång per rad. I Kimball Groups designprocess i fyra steg kommer fastställandet av granulariteten som nummer två, direkt efter valet av

affärsprocess och innan någon dimension eller något faktum namnges, och deras regel är att olika granulariteter inte får dela en faktatabell. Håll en separat faktatabell på ordernivå för belopp som hör till hela ordern, eller fördela avgiften på raderna enligt en regel som verksamheten godkänner.

En kund flyttar, och fjolårets rapport ändras. När regionen på kundraden bara skrivs över följer all tidigare försäljning med kunden till den nya regionen, och en rapport som godkändes i januari berättar en annan historia i juni. Hantera egenskaper där historiken spelar roll med typ 2: stäng den aktuella raden och lägg till en ny med en ny surrogatnyckel, i en och samma databastransaktion, så att ingen laddning någonsin ser två aktuella rader.

```
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', 'Norr', current_date, NULL, true);
COMMIT;
```

Två team bygger var sin kunddimension, och diskussionen som analysdatalagret skulle avsluta börjar om. Ett nytt ämnesområde går snabbare att leverera med en egen kopia av kund- eller produktlistan, och varje kopia glider isär från de andra: en utelämnar avslutade konton, en annan slår ihop dubbletter på ett annat sätt. Bygg de gemensamma dimensionerna en gång i ETL-systemet och låt varje ny faktatabell använda dem. Bussmatrisen är checklistan: innan en affärsprocess läggs till visar den vilka befintliga dimensioner den nya faktatabellen ska peka på.

Rapporterna pekar på staging-tabellerna. Råa kopior av källorna landar först i back room, och när de väl ligger där är det frestande att fråga direkt mot dem, eftersom de redan är laddade. De tabellerna bär varje källas egna koder, dubbletter och luckor och ändras varje gång en källa gör det, så varje rapport som bygger på dem ärver precis det problem som analysdatalagret skulle lösa. Håll back room för ETL-systemet ensamt, och ge rapporter, dashboards och analytiker bara presentationsområdet, där de gemensamma dimensionerna och den fastställda granulariteten är det som får siffrorna att stämma.

Läs mer

- Kimball Technical DW/BI System Architecture — back room och front room, ETL-delsystemen, metadata och alternativen hub-and-spoke och hybrid.
- Enterprise Data Warehouse Bus Architecture — conformed dimensions, bussmatrisen och uppbyggnad en affärsprocess i taget.
- Four-Step Dimensional Design Process — affärsprocess, granularitet, dimensioner och fakta, i den ordningen.
- Type 2: Add New Row — historik för en egenskap som ändras, med en ny rad, en ny surrogatnyckel och tre extra kolumner.
- BigQuery overview — Googles hanterade analysdatalager, med lagring och beräkning åtskilda och data lagrade kolumnvis.

Ordlista

Välj utifrån era faktiska förutsättningar

Ta med ett representativt dataset, frågorna det ska besvara och dagens flaskhals. Vi kan avtala mått på kvalitet, genomströmning och återställning.

Se vad det kostar Hör av er →

Gör begreppen praktiska

Fråga hur ”Vad är ett analysdatalager?” gäller i er organisation. Ett konkret exempel från ert nuvarande system räcker för att börja samtalet.

Se våra tjänster

welcome@engineer.company +45 50 10 61 84 Köpenhamn, Danmark

Ta med dig den här sidan

Visitkort

Omsorgsfullt skapad med kärlek av D.

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

Upphovsrätt © 2025 – nu · Engineer ApS · Alla rättigheter förbehållna.

Utskrivet från https://engineer.company/sv/glossary/data-warehouse/