Engineer

Зіркова схема

★ модель з однією таблицею фактів і таблицями вимірів навколо

Що це таке

Зіркова схема (англійською star schema) — спосіб розкласти таблиці для аналізу. Таблиця посередині записує бізнес-події, як-от продажі, відвантаження чи платежі, у вигляді чисел. Довкола неї стоять таблиці, що ці події описують: який товар, який магазин, який день, який клієнт. Якщо намалювати все це діаграмою й провести лінію від центральної таблиці до кожної з решти, вийде зірка.

Таблицю посередині називають таблицею фактів, а таблиці довкола — таблицями вимірів. Kimball Group визначає зіркову схему саме так: таблиці фактів, пов'язані з таблицями вимірів, які до них належать. Зв'язок тут — звичайний ключ. Кожна таблиця вимірів має єдиний первинний ключ, а кожен рядок фактів несе цей ключ як зовнішній, по одному на кожен вимір, що описує подію.

Такий поділ випливає з розрізнення, яке Kimball Group проводить між вимірюванням і контекстом. Вимірна модель будується довкола подій вимірювання в бізнес-процесі: товар відскановано на касі, замовлення відвантажено, страхове відшкодування виплачено. Числа, які дає подія, як-от кількість і сума, — це факти. Усе, що відповідає на запитання хто, що, де, коли, чому і як, — це контекст, і він живе у вимірах.

Два види таблиць мають протилежну форму. На найдетальнішому рівні таблиця фактів містить один рядок на подію: числові показники, зовнішній ключ до кожного виміру, а іноді ще й номер документа, як-от рахунку чи чека. Коли дата, клієнт і магазин документа вже стали вимірами, від нього лишається тільки цей номер, тож його зберігають просто в рядку фактів без жодної таблиці вимірів за ним. Kimball Group називає такий номер виродженим виміром (degenerate dimension). Таблиця вимірів, навпаки,

широка, пласка й навмисно денормалізована, з багатьма описовими текстовими стовпцями. Вимір товару повторює бренд, категорію і відділ у кожному рядку товару, замість того щоб винести їх в окремі таблиці.

Зіркова схема — реляційна форма вимірної моделі, складена зі звичайних таблиць у звичайній базі даних SQL. Інша форма, яку описує Kimball Group, — OLAP-куб: багатовимірна база з тими самими атрибутами й фактами, до якої звертаються мовами з більшими аналітичними можливостями, ніж SQL, наприклад MDX, а XMLA забезпечує протокол доступу. Куби часто будують із зіркової схеми як останній крок перед тим, як дані побачать люди, або як шар агрегатів. Правила моделювання для обох форм однакові, і Kimball Group називає їх незалежними від архітектури: вони так само підходять сховищу даних у стилі Kimball, як і корпоративній інформаційній фабриці (Corporate Information Factory) Білла Інмона, побудованій за схемою «хаб і спиці».

Застосуйте поняття до рішення

Застосуйте «Що таке зіркова схема?» до конкретного рішення. Ми можемо пов'язати поняття із системою, яку ви створюєте або купуєте. Дані продажів, операцій і фінансів зведені разом, щоб два відділи не приносили на одну нараду дві різні відповіді.

Переглянути наші послуги Зв'язатися з нами →

Перевірте припущення за терміном

Використайте пояснення та пов'язані роботи, щоб сформулювати запитання для проєкту. Ми можемо перевірити припущення разом до вибору інструмента. Прискорила пайплайн географічних даних у 50 разів — 50x швидші пайплайни географічних даних

Прискорила пайплайн географічних даних у 50 разів Зв'язатися з нами →

Чому це важливо

Така форма робить аналітичні запитання схожими одне на одне. Виберіть факти, які треба підсумувати, з’єднайте таблицю фактів із вимірами, що їх описують, відфільтруйте за одними стовпцями вимірів і згрупуйте за іншими. Kimball Group описує таблиці фактів як головну ціль обчислень і динамічних агрегацій, а атрибути вимірів — як головне джерело обмежень у запитах, групувань і підписів у звітах. Аналітик, який один раз засвоїв цей шаблон, напише наступний запит до нової таблиці фактів без сторонньої допомоги, а інструмент звітності складе такий запит зі списку полів, бо кожна зірка влаштована однаково.

Друга причина — читабельність. Операційну базу проєктують так, щоб кожен факт зберігався один раз і лишався узгодженим, тож категорія товару зазвичай лежить в окремій довідковій таблиці й приєднується назад за

ідентифікатором. Для системи, що приймає замовлення, це правильний дизайн, а для людини, яка шукає продажі за категоріями, — хибний: їй спершу доводиться з'ясовувати, які таблиці з'єднувати і в якому порядку. Пласкі виміри збирають усі описові стовпці в одному місці, і все в них записано словами. Kimball Group застерігає від альтернативи, тобто від нормалізації ієрархії виміру в додаткові таблиці, бо бізнес-користувачам такий результат важко зрозуміти й у ньому важко орієнтуватися.

Третя причина — добре побудована зірка може відповісти на запитання, яких ніхто не ставив, коли її проєктували. Гранулярність таблиці фактів, тобто точне значення одного рядка, оголошують раніше за будь-який вимір чи факт, і Kimball Group радить починати з атомарної гранулярності: найнижчого рівня, який фіксує бізнес-процес. Таблицю з одним рядком на позицію чека можна підсумувати за днями, магазинами, брендами чи будь-яким їхнім поєднанням. Таблицю, де зберігаються тільки денні підсумки за магазинами, вже ніколи не розкладеш за товарами, хоч як сформулюй запитання. За словами Kimball Group, атомарні дані витримують непередбачувані запити, а зведені дані наперед припускають, які запитання їм ставитимуть.

Ціна — навмисна надлишковість. Назва відділу повторюється в кожному рядку товару, і завантаження, яке будує зірку, мусить тримати ці повтори узгодженими й вирішувати, що робити, коли опис змінюється. Зірка також не замінює операційну модель: її будують з операційної моделі процесом завантаження, і вона актуальна рівно настільки, наскільки свіжий останній запуск цього завантаження.

Як це виглядає на практиці

Роздрібна мережа хоче бачити продажі за категоріями товарів і місяцями. Бізнес-процес тут — касовий чек, а гранулярність оголошено як один рядок на позицію чека. Вимір товару — одна пласка таблиця, а таблиця фактів містить ключ для кожного виміру, номер чека як вироджений вимір і два показники.

```
CREATE TABLE dim_product (
    product_key  bigint PRIMARY KEY,   -- сурогатний ключ, належить сховищу
    sku          text NOT NULL,        -- код у системі-джерелі
    product_name text NOT NULL,
    brand        text NOT NULL,
    category     text NOT NULL,
```

```
    department   text NOT NULL
);

CREATE TABLE fact_sales (
    date_key       integer NOT NULL REFERENCES dim_date,
    store_key      integer NOT NULL REFERENCES dim_store,
    product_key    bigint  NOT NULL REFERENCES dim_product,
    receipt_number text    NOT NULL,   -- вироджений вимір
    quantity       integer NOT NULL,
    sales_amount   numeric(12, 2) NOT NULL
);

SELECT d.year_month, p.category, sum(f.sales_amount) AS sales
FROM fact_sales AS f
JOIN dim_date AS d USING (date_key)
JOIN dim_product AS p USING (product_key)
WHERE p.department = 'Бакалія'
GROUP BY d.year_month, p.category
ORDER BY d.year_month, p.category;
```

Запит має ту саму форму, що й будь-який запит до зірки: таблиця фактів, по одному з’єднанню на вимір, фільтри й групування за текстом вимірів і сума за фактом. Щоб додати до звіту регіон магазину, потрібні ще одне з’єднання і ще один стовпець у `GROUP BY`; таблиця фактів при цьому не змінюється. Номер чека лишається в рядку фактів, тож позиції одного кошика все ще можна порахувати разом, але окремої таблиці вимірів він не отримує, бо більше про нього сказати нічого.

Зірка — одна з відповідей на обіцянку послуги Проєктування та моделювання баз даних: структура, що відповідає реальній роботі бізнесу, тож потрібні згодом запитання до даних усе ще можна буде поставити. Операційна модель, з якої зірку будують, — друга половина цієї роботи. У межах поточної роботи для NextMariner засновниця змоделювала морську предметну галузь у 348 нормалізованих таблицях у межах 34 схем PostgreSQL: фахівці, компанії, судна, вакансії, відгуки. Там дрібні стабільні переліки, як-от статуси, типи й категорії, стали довідковими таблицями SMALLINT, завдяки чому рядки лишаються компактними, а з’єднання дешевими, а справжні зовнішні ключі й обмеження роблять цілісність відповідальністю самої бази. Саме така форма потрібна системі, що записує рядки по одному і мусить лишатися коректною. Зірка для звітності поверх такої системи йде у зворотному напрямку: у пласких вимірах вона знову перетворює коди на слова, бо її читає людина, а не застосунок.

Де виникають проблеми

Виміри нормалізують, і зірка перетворюється на сніжинку. Винести бренд, категорію і відділ з виміру товару в окремі таблиці здається охайним рішенням кожному, хто звик до операційного проєктування. Kimball Group називає результат сніжинкою (snowflake) і радить так не робити: бізнес-користувачам його важко зрозуміти й у ньому важко орієнтуватися, а додаткові з'єднання можуть сповільнити запити, хоча пласкі виміри містять рівно ту саму інформацію. Тримайте кожен вимір однією широкою таблицею, а ієрархію нехай повторює в кожному рядку завантаження.

Гранулярність підганяють під перший звіт. Таблиця фактів із денними продажами за магазинами відповідає на той дашборд, для якого її будували, і ні на що детальніше; перше ж запитання про окремий товар чи кошик вимагає нового завантаження з джерела. Таблиця, що змішує денні підсумки з окремими позиціями, ще гірша, бо будь-яка сума за нею рахує той самий продаж двічі. Оголосіть гранулярність першою, почніть з атомарного рівня, який фіксує процес, і дайте кожній гранулярності власну таблицю фактів. Тоді для швидкості можна додати зведені таблиці, але саме як окремі таблиці.

Атрибути вимірів — це коди з системи-джерела. Операційні системи зберігають компактні коди, і якщо скопіювати їх просто у вимір, вийдуть звіти з підписами рядків, які розуміє лише система-джерело. Kimball Group допускає коди як атрибути, але вважає найкориснішими розгорнуті описи, бо саме вони стають підписами у звітах. Розшифровуйте кожен код у слова під час завантаження, а сам код тримайте в стовпці поруч з описом для тих, кому треба простежити рядок до джерела.

Таблицю фактів формують як звіт. Спокусливо спроєктувати її за електронною таблицею, якою бізнес уже користується, зі стовпцем на кожен місяць чи рядком на кожен пункт управлінського звіту. Kimball Group прямо каже, що дизайн таблиці фактів спирається на фізичну діяльність, яку вимірюють, а не на звіти, які з неї потім зроблять. Вигляд звітів змінюється, касовий чек — ні. Моделюйте подію, один рядок на кожен її випадок на оголошеній гранулярності, а потрібний вигляд нехай дають запити.

Що почитати далі

- Star Schemas and OLAP Cubes — визначення, факти й виміри як вимірювання і контекст, а також те, як куби пов'язані із зірками.

- Grain — що означає один рядок у таблиці фактів і чому першою йде атомарна гранулярність.
- Fact Table Structure — числові показники й зовнішні ключі в одному рядку на подію, спроєктовані від діяльності, а не від звіту.
- Dimension Table Structure — широкі, пласкі, денормалізовані таблиці, чиї описові атрибути визначають фільтри, групування й підписи.
- Snowflaked Dimensions — чому нормалізація ієрархії виміру робить його незручнішим у користуванні.
- Microsoft: доступ до даних багатовимірних моделей — MDX як мова запитів і обчислень та XMLA як інтерфейс, через який клієнт може надіслати інструкцію MDX.

Словник

Обирайте за фактичними умовами

Надайте репрезентативний набір даних, запитання до нього та поточне вузьке місце. Узгодимо вимірювання якості, пропускної здатності та відновлення.

Дивитися, скільки це коштує Зв'язатися з нами →

Зробіть термінологію практичною

Запитайте, як «Що таке зіркова схема?» стосується вашої організації. Конкретного прикладу з поточної системи достатньо для початку розмови.

Переглянути наші послуги

welcome@engineer.company +45 50 10 61 84 Копенгаген, Данія

Візьміть цю сторінку із собою

Візитівка

Дбайливо зроблено з любов'ю — D.

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

Авторське право © 2025 – дотепер · Engineer ApS · Усі права захищено.

Надруковано з https://engineer.company/uk/glossary/star-schema/