ТЕХНОЛОГІЇ ПОБУДОВИ СХОВИЩ ДАНИХ

5.1. Базові технології побудови сховищ даних
5.2. Інтеграція даних
5.3. Архітектура даних
5.4. Вітрини та кіоски даних

Основні терміни теми: властивості сховищ даних, інтеграція, OLAP-технологія, методи інтеграції даних, консолідація даних, федералізація даних, розповсюдження даних, архітектура даних, моделі архітектури, вітрини даних, кіоски даних, корпоративні сховища, схема «зірка», схема «сніжинка».

5.1. Базові технології побудови сховищ даних

Ідея, покладена в основу технологій інформаційних сховищ, полягає в тому, що проводити оперативний аналіз безпосередньо на базі інформаційних систем неефективно. Натомість, всі необхідні для аналізу дані витягуються з декількох традиційних баз даних (в основному, реляційних), перетворюються і потім поміщаються в одне джерело даних – сховище даних.

В процесі занурення дані:
очищаються - усунення непотрібної інформації;
агрегуються - обчислення сум, середніх;
трансформуються - перетворення типів даних, реорганізація структур зберігання;
об'єднуються із зовнішніх і внутрішніх джерел - приведення до єдиних форматів;
синхронізуються - відповідність одному моменту часу.

Сьогодні, технології побудови сховищ даних є основою для створення повноцінних інтелектуальних система налізу даних, орієнтованих на рішення слабко структурованих задач прийняття рішень, оскільки вони містить дані, що володіють наступними властивостями:

Цілісністю і внутрішнім взаємозв'язком. Хоча дані занурюються з різних джерел, але вони об'єднані єдиними законами іменування, способами вимірювання атрибутів і т.д. Це має велике значення для корпоративних організацій, в яких одночасно можуть експлуатуватися різні по своїй архітектурі обчислювальні системи, що представляють однакові дані по-різному. Наприклад, можуть використовуватися декілька різних форматів представлення дат, або один і той же показник може називатися різним чином, наприклад, "вірогідність доведення інформації" і "вірогідність отримання інформації". В процесі занурення подібні невідповідності усуваються автоматично.

Предметною орієнтованістю. Локальні бази даних містять мегабайти інформації, абсолютно не потрібної для аналізу (адреси, поштові індекси, ідентифікатори записів і т.п.). Подібна інформація не заноситься в сховище, що обмежує спектр розглядаємих при ухваленні рішення даних до мінімуму.

Відсутністю часової прив'язки. Оперативні системи охоплюють невеликий інтервал часу, що досягається за рахунок періодичної архівації даних. Сховища даних, навпаки, містять історичні дані, накопичені за великий інтервал часу (роки, десятиліття).

Доступністю виключно для читання. Модифікація даних не проводиться, оскільки вона може привести до порушення цілісності сховища даних. Оскільки не потрібно мінімізувати час занурення, то структура сховища може бути оптимізована для обробки певних запитів, що досягається за рахунок денормалізації реляційної схеми, попередньої агрегації і побудови найбільш доречних індексів.

Інтегрованість означає, що дані задовольняють вимогам всього підприємства, а не одній функції бізнесу. Цим сховище даних гарантує, що однакові звіти, що згенерували для різних аналітиків, міститимуть однакові результати.

Незмінність означає, що, потрапивши один раз в сховищі, дані там зберігаються і не змінюються. Дані в сховищі можуть лише додаватися.

Всі дані, які містяться в сховищі, можна розділити на наступні категорії:

метадані (дані про даних);

агреговані дані;

детальні дані.

Важливою особливістю систем інтелектуального аналізу даних на основі сховищ даних є метадані. Це різного роду системні словники, що дозволяють контролювати склад і структуру інформації в сховищі, управляти процесами завантаження і розрахунків і т.п. Без прошарку управлінських даних сховище з часом загрожує перетворитися на велике електронне звалище.

Ключовою відмінністю інформації сховища є не тільки спосіб його наповнення (з різних зовнішніх і внутрішніх джерел), але і модель зберігання даних, особливо це стосується агрегованої інформації. У сховищі інформація розміщується в денормалізованому вигляді у формі класичної «сніжинки» або «зірки». Такий підхід дозволяє істотно понизити час відгуку бази даних при виконанні запитів. Не вдаючись до технологічних особливостей проектування, відзначимо, що відбувається це за рахунок деякої надмірності зберігання даних. Така форма зберігання інформації в корпоративних сховищах є загальносвітовою практикою. Разом з тим частина детальних даних цілком може зберігатися в нормалізованих таблицях. Вимога до спеціальних структур зберігання обумовлена деяким спеціальним способом використання даних, про що доцільно сказати докладніше.

5.2. Інтеграція даних

Використовувати інформацію, накопичену в сховищі, можна як за допомогою традиційних звітів, так і з використанням динамічних запитів до бази даних. Існують також абсолютно специфічні способи використання інформації, призначені спеціально для аналітичних завдань. До них відносяться так звані OLAP-технології (On-line Analytical Processing) і технології інтелектуального аналізу даних. З практичної точки зору рішучий крок, який робить OLAP-технологія, полягає в тому, щоб, відмовившись від зайвої спільності, зробити процес аналізу максимально швидким. В рамках цієї технології передбачається, що склад і структура показників для аналізу відомий наперед і міняється дуже рідко (для деяких видів систем - практично не міняється). Користувач може виконувати над даними в такому багатовимірному уявленні набір OLAP-операцій підйому (консолідації по деяких напрямах), спуску (деталізації по деякому напряму), повороту (зміни напряму сортування). Детальні дані в системах сховищ даних найчастіше є джерелом для інтелектуального аналізу.

Таким чином, дані, розміщені в інформаційне сховище даних, організовуючись в інтегровану цілісну структуру, володіють природними внутрішніми зв'язками, набувають нових властивостей, що дає їм можливість набути статус інформації.

Розглянемо характеристики інтеграції даних в сховищі даних. Як вже відомо, метою інтеграції даних є отримання єдиної та цільної картини бізнес-даних. Для її досягнення застосуємо модель, яка включає додатки, продукти, технології та методи:

додатки - це рішення, створені постачальниками відповідно до вимог клієнтів, які використовують один або більше продуктів інтеграції даних;

продукти - це готові комерційні рішення, що підтримують одну або більше технології інтеграції даних;

технології реалізують один або більше методів інтеграції даних;

методи - це підходи до інтеграції даних, незалежні від технологій.

Існує три основні методи інтеграції даних: консолідація, федералізація і розповсюдження.

Консолідація даних. При використанні цього методу дані збираються з декількох первинних систем і інтегруються в одне постійне місце зберігання. Таке місце зберігання може бути використане для підготовки звітності і проведення аналізу, як у випадку застосування

сховища даних, або як джерело даних для інших додатків, як у випадку впровадження операційного складу даних. При використанні цього методу зазвичай існує деяка затримка між моментом оновлення інформації в первинних системах і часом, коли ці зміни з'являються в кінцевому місці зберігання. Залежно від потреб бізнесу таке відставання може складати декілька секунд, годин або багато днів. Термін «*режим, наближений до реального часу»* часто використовується для опису кінцевих даних, оновлення яких відстає від джерела на декілька секунд, хвилин або годин. Дані, що не відстають від джерела, вважаються даними «*в режимі реального часу»*, але це важко досягнути при використанні методу консолідації даних.

Перевагою консолідації даних є те, що цей підхід дозволяє здійснювати трансформацію значних об'ємів даних (реструктуризацію, узгодження, очищення і/або агрегацію) в процесі їх передачі від первинних систем до кінцевих місць зберігання. Деякі складнощі, пов'язані з даним підходом, - це значні обчислювальні ресурси, які потрібні для підтримки процесу консолідації даних, а також істотні ресурси пам'яті, необхідні для підтримки кінцевого місця зберігання. Але з урахуванням постійного вдосконалення апаратних засобів це не проблема.

Консолідація даних - це основний підхід, який використовується програмними додатками сховищ даних для побудови і підтримки оперативних складів даних і корпоративних сховищ. Консолідація даних також може знайти застосування для створення залежної вітрини даних, але в цьому випадку в процесі консолідації використовується тільки одне джерело даних (наприклад, корпоративне сховище). У середовищі сховищ даних однієї з найпоширеніших технологій підтримки консолідації є технологія ETL (витягання, перетворення і завантаження - extract, transform, and load). Ще одна поширена технологія консолідації даних - управління змістом корпорації (Enterprise Content Management - ECM). Більшість рішень ECM направлені на консолідацію і управління неструктурованими даними, такими як документи, звіти і web-сторінки.

Федералізація даних. Процес забезпечує єдину віртуальну картину одного або декількох первинних файлів даних. Якщо бізнес-додаток генерує запит до цієї віртуальної картини, то процесор федералізації даних витягує дані відповідних первинних складів даних, інтегрує їх так, щоб вони відповідали віртуальній картині і вимогам запиту, і відправляє результати бізнес-додатку від якого прийшов запит. За визначенням, процес федералізації даних завжди полягає у *витяганні* даних з первинних систем на підставі зовнішніх вимог.

Всі необхідні перетворення даних здійснюються при їх витяганні з первинних файлів. Інтеграція корпоративної інформації (Enterprise Information Integration - EII) - це приклад технології, яка підтримує федеральний підхід до інтеграції даних. Один з ключових елементів федеральної системи - це метадані, які використовуються процесором федералізації даних для доступу до первинних даних. В деяких випадках ці метадані можуть складатися виключно з визначень віртуальної картини, які ставляться у відповідність («меппіруются») первинним файлам. У більш передових рішеннях метадані також можуть містити детальну інформацію про кількість даних, що знаходяться в первинних системах, а також про шляхи доступу до них. Така розширена інформація може допомогти федеральному рішенню оптимізувати доступ до первинних систем.

Вважається, що основна перевага федерального підходу - той факт, що він забезпечує доступ до поточних даних і позбавляє від необхідності консолідувати первинні дані в новому сховищі даних. Але слід пам'ятати, що федералізація даних не дуже добре підходить для витягання і узгодження великих масивів даних або для тих додатків, де існують серйозні проблеми з якістю даних в первинних системах. Ще один істотний чинник - потенційний вплив на продуктивність і додаткові витрати на доступ до численних джерел даних під час виконання програми.

Федеральна архітектура дуже корисна для крупних транснаціональних корпорацій і є вельми зручним підходом для підтримки балансу між необхідністю автономії місцевих підрозділів компанії і їх гнучкості, з одного боку, і стандартизації і централізованого контролю, які здійснює центральний офіс, - з іншою. При цьому федеральним сховищем може бути як єдине фізичне федеральне сховище, так і федерація дрібніших спеціалізованих сховищ даних.

Розповсюдження даних. Додатки розповсюдження даних здійснюють копіювання даних з одного місця в інше. Ці додатки зазвичай працюють в оперативному режимі і проводять переміщення даних до місць призначення, тобто залежать від певних подій. Оновлення в первинній системі можуть передаватися в кінцеву систему синхронно або асинхронно. Синхронна передача вимагає, щоб оновлення в обох системах відбувалися під час однієї і тієї ж фізичної транзакції. Незалежно від використовуваного типу синхронізації, метод розповсюдження гарантує доставку даних в систему призначення. Така гарантія - це ключова відмітна ознака розповсюдження даних. Більшість технологій синхронного розповсюдження даних підтримують двосторонній обмін даними між первинними і кінцевими системами. Прикладами технологій, що підтримують розповсюдження даних, є інтеграція корпоративних додатків (Enterprise Application Ntegration - EAI) і тиражування корпоративних даних (Enterprise Data Replication - EDR).

Великою перевагою методу розповсюдження даних є те, що він може бути використаний для переміщення даних в режимі реального часу або близькому до нього. Інші достоїнства включають гарантовану доставку даних і двостороннє розповсюдження даних. Метод розповсюдження даних може також використовуватися для урівноваження робочого навантаження, створення резервних копій і відновлення даних, зокрема у разі надзвичайних ситуацій. Практичне застосування цього методу відрізняється чималою різноманітністю як в плані продуктивності, так і відносно можливостей реструктуризації і очищення даних. Деякі корпоративні продукти розповсюдження даних можуть підтримувати переміщення і реструктуризацію крупних масивів даних, тоді як продукти EAI часто мають обмежені можливості пересування великої кількості даних і їх реструктуризації. Одна з причин подібної відмінності - той факт, що в центрі архітектури тиражування корпоративних даних лежать дані, а в центрі технології EAI - повідомлення або транзакції.

Гібридний підхід. Методи, які використовуються додатками інтеграції даних, залежать як від потреб бізнесу, так і від технологічних вимог. Достатньо часто додаток інтеграції даних використовує так званий гібридний підхід, який включає декілька методів інтеграції. Хороший приклад такого підходу - інтеграція даних про клієнтів (Customer Data Ntegration - CDI), метою якої є забезпечення узгодженої картини інформації про клієнтів.

Найпростіший підхід до CDI - це створення консолідованого сховища даних про клієнтів, який містить дані, отримані з первинних систем. Відставання інформації в консолідованому сховищі залежатиме від режиму консолідації даних (оперативний або пакетний) і від частоти оновлення цієї інформації. Інший підхід до CDI - це федералізація даних, коли визначаються віртуальні бізнес-представлення даних про клієнтів в первинних системах. Ці уявлення використовуються бізнес-додатками для доступу до поточної інформації про клієнтів в первинних системах. При федеральному підході також може використовуватися довідковий файл метаданих для зв'язку інформації про клієнтів на основі загальних ключових елементів.

Гібридний підхід, що використовує як консолідацію, так і федералізацію даних, також може мати місце. Загальні дані про клієнтів (ім'я, адреса і т.д.) можуть бути консолідовані в одному сховищі, а дані, які відносяться до певного первинного додатку (наприклад, замовлення), можуть бути федералізовані. Такий гібридний підхід може бути розширений за рахунок розповсюдження даних. Якщо клієнт оновлює своє ім'я і адресу під час транзакції в Інтернет-магазині, то ці зміни можуть бути відправлені до консолідованого сховища даних, а звідти поширені в інші первинні системи, такі як база даних про клієнтів роздрібного магазина.

5.3. Архітектура даних

На сьогоднішній день існує два основні підходи до архітектури сховищ даних. Це так звана корпоративна інформаційна фабрика (Corporate Information Factory - CIF) Білла Інмона і

сховище даних з архітектурою шини (Data Warehouse Bus - BUS) Ральфа Кимболла. Розглянемо кожний з них докладніше.

Corporate Information Factory. Колись цей підхід був відомий під назвою корпоративного сховища даних (Enterprise Data Warehouse - EDW). Робота такого сховища починається з скоординованого витягання даних джерел. Після цього завантажується реляційна база даних з третьою нормальною формою, що містить атомарні дані. Отримане нормалізоване сховище використовується для того, щоб наповнити інформацією додаткові репозиторії презентаційних даних, тобто даних, підготовлених для аналізу. Ці репозиторії, зокрема, включають спеціалізовані сховища для вивчення і «здобичі» даних (Data Mining), а також вітрини даних.

При такому сценарії кінцеві вітрини даних створюються для обслуговування бізнес-відділів або для реалізації бізнес-функцій і використовують просторові моделі для структуризації сумарних даних. Атомарні дані залишаються доступними через нормалізоване сховище даних. Очевидно, що структура атомарних і сумарних даних при такому підході істотно розрізняється. В якості відмітних характеристик підходу Білл Інмона до архітектури сховищ даних можна назвати наступні:

використання реляційної моделі організації атомарних даних і просторовою - для організації сумарних даних;

використання ітеративного або «спірального» підходу при створенні великих сховищ даних, тобто «будівництво» сховища не відразу, а по частинах. Це дозволяє при необхідності вносити зміни в невеликі блоки даних або програмних кодів і позбавляє від необхідності перепрограмувати значні об'єми даних в сховищі. Те ж саме можна сказати і про потенційні помилки: вони також будуть локалізовані в межах порівняльного невеликого масиву без ризику зіпсувати все сховище;

використання третьої нормальної форми для організації атомарних даних, що забезпечує високий ступінь детальної інтегрованих даних і, відповідно, надає корпораціям широкі можливості для маніпулювання ними і зміни формату і способу представлення даних в міру необхідності;

сховище даних - це проект корпоративного масштабу, що охоплює всі відділи і обслуговує потреби всіх користувачів корпорації.

сховище даних - це не механічна колекція вітрин даних, а фізично цілісний об'єкт.

Data Warehouse Bus. У цій моделі первинні дані перетворяться в інформацію, придатну для використання, на етапі підготовки даних. При цьому обов'язково приймаються до уваги вимоги до швидкості обробки інформації і якості даних. Як і в моделі Білла Інмона, підготовка даних починається з скоординованого витягання даних з джерел. Ряд операцій здійснюється централізовано, наприклад, підтримка і зберігання загальних довідкових даних, інші дії можуть бути розподіленими. Область уявлення просторово структурована, при цьому вона може бути централізованою або розподіленою.

Просторова модель сховища даних містить ту ж атомарну інформацію, що і нормалізована модель, але інформація структурована по-іншому, щоб полегшити її використання і виконання запитів. Ця модель включає як атомарні дані, так і загальну інформацію (агрегати в зв'язаних таблицях або багатовимірних кубах) відповідно до вимог продуктивності або просторового розподілу даних. Запити в процесі виконання звертаються до все більш низького рівня деталізації без додаткового перепрограмування з боку користувачів або розробників додатку.

На відміну від підходу Білла Інмона, просторові моделі будуються для обслуговування бізнес-процесів (які, у свою чергу, пов'язані з бізнес-показниками), а не бізнес-відділа. Наприклад, дані про замовлення, які повинні бути доступні для загально корпоративного використання, вносяться до просторового сховища даних тільки один раз, на відміну від CIF-підходу, в якому їх довелося б тричі копіювати у вітрини даних відділів маркетингу, продажів і фінансів. Після того, як в сховищі з'являється інформація про основні бізнес-процеси, консолідовані просторові моделі можуть видавати їх перехресні характеристики. Матриця

корпоративного сховища даних з архітектурою шини виявляє і підсилює зв'язки між показниками бізнес-процесів (фактами) і описовими атрибутами (вимірюваннями).

Підсумовуючи все вищесказане, можна відзначити типові риси підходу Ральфа Кимболла:

використання просторової моделі організації даних з архітектурою «зірка» (star scheme);

використання дворівневої архітектури, яка включає стадію підготовки даних, недоступну для кінцевих користувачів, і сховище даних з архітектурою шини як таке. До складу останнього входять декілька вітрин атомарних даних, декілька вітрин агрегованих даних і персональна вітрина даних, але воно не містить одного фізично цілісного або централізованого сховища даних;

сховище даних з архітектурою шини володіє наступними характеристиками: воно просторове, воно включає як дані про транзакції, так і сумарні дані, воно включає вітрини даних, присвячені тільки одній предметній області або що мають тільки одну таблицю фактів (fact table), воно може містити безліч вітрин даних в межах однієї бази даних;

сховище даних не є єдиним фізичним репозиторієм (на відміну від підходу Білла Інмона). Це «віртуальне» сховище. Це колекція вітрин даних, кожна з яких має архітектуру типу «зірка».

5.4. Вітрини та кіоски даних

У найбільш загальному виді сховища даних можуть бути розбиті на два типи: корпоративні сховища даних (Enterprise Data Warehouses) і кіоски або вітрини даних (Data Marts).

Корпоративні сховища даних містять інформацію, що відноситься до всієї корпорації і зібрану з безлічі оперативних джерел для консолідованого аналізу. Зазвичай такі сховища охоплюють цілий ряд аспектів діяльності корпорації і використовуються для ухвалення як тактичних, так і стратегічних рішень. Корпоративне сховище містить детальну і узагальнену інформацію, його об'єм може досягати від 50 Гбайт до одного або декількох терабайт. Вартість створення і підтримки корпоративних сховищ може бути дуже високою. Зазвичай їх створенням займаються централізовані відділи інформаційних технологій, причому створюються вони зверху вниз, тобто спочатку проектується загальна схема, і тільки тоді починається заповнення даними. Такий процес може займати декілька років.

Кіоски або вітрини даних містять підмножину корпоративних даних і будуються для відділів або підрозділів усередині організації. Кіоски даних часто будуються силами самого відділу і охоплюють конкретний аспект, що цікавить співробітників даного відділу. Кіоск даних може отримувати дані з корпоративного сховища (залежний кіоск) або, що поширеніше, дані можуть поступати безпосередньо з оперативних джерел (незалежний кіоск). Кіоски і сховища даних будуються за схожими принципами і використовують практично одні і ті ж технології.

Багато компаній, що усвідомлюють необхідність розробки корпоративного сховища даних, все ж таки не в силах справитися зі всіма завданнями виділення, стандартизації і об'єднання терабайт даних. Натомість вони вважають за краще будувати кіоски (або вітрини) даних (Data Marts) - спеціалізовані сховища даних, присвячені тільки одному напряму діяльності організації. Кіоск (вітрина) даних - це, найчастіше, найбільш керований різновид сховища даних. Його безперечний недолік полягає в тому, що без сховища даних, яке охоплювало б інформацію всього підприємства, неможливо порівнювати і аналізувати дані по всіх відділах і процесах. У багатьох компаніях вже зрозуміли, що кіоски (вітрини) даних можуть послужити хорошу службу і навіть стати єдино можливим рішенням для виконання термінових аналітичних завдань, але створення спеціалізованих кіосків без попередньої розробки корпоративної інфраструктури сховища даних може згодом привести до великих утруднень.

За класичним визначенням, вітрина (кіоск) даних (Data Mart) є підмножиною сховища даних, що відображає специфіку підрозділу (бізнес-об'єкт) і що забезпечує підвищену продуктивність. Таким чином, вітрина є ланкою, на якій базується конкретна аналітична система для вирішення свого кола завдань. Проте можлива ситуація, коли деяка область діяльності підприємства практично не корелює з іншими, і можливо побудувати відповідну вітрину даних автономно, без прив'язки до корпоративного сховища. Тоді вітрина поповнюватиметься даними безпосередньо з оперативних систем обробки транзакцій. Такі вітрини даних отримали назву незалежних, на відміну від класичних залежних від сховища даних і поповнюваних з нього вітрин.

У ряді випадків представляється доцільним розвернути вітрину (кіоск) даних замість повністю сформованого сховища. Вітрини даних накладають менші зобов'язання, вони дешевше і простіше в побудові і базуються на дешевших серверах, а не на мультипроцесорних комплексах. При такому підході немає необхідності задіювати цілу інформаційну систему корпорації і підтримувати складні процедури синхронного оновлення вітрини даних при оновленні сховища. В той же час необхідний розуміти, що при такому підході вітрини даних можуть розмножитися в цілі комплекси незалежних інформаційних баз даних, і природно буде поставлено завдання управління індивідуальними стратегіями пошуку, обслуговування і відновлення. З іншого боку, будувати єдине корпоративне сховище на основі множини незалежних вітрин даних значно вигідніше, ніж спираючись на розсіяні по системах обробки транзакцій дані.

Так що ж доцільно застосовувати: єдине сховище, самостійні вітрини (кіоски) даних, сховище із залежними вітринами або інші варіанти? Універсальної відповіді на питання про необхідність застосування того або іншого варіанту не існує. В кожному випадку оптимальний варіант визначається вимогами бізнесу, інтенсивністю запитів, мережевою архітектурою, необхідною швидкістю реакції і іншими умовами.

Кіоски добре підходять компаніям, які вимушені підтримувати системи підтримки ухвалення рішення і не володіють достатнім досвідом для розробки повномасштабного сховища даних.

Прийоми моделювання кіосків (вітрин) даних відрізняються від прийомів моделювання сховищ даних через різні вимоги до структур даних. Якщо основною задачею сховища даних є зберігання консолідованої історичної інформації, то вітрина даних будується з урахуванням вимог по доступу до даних і представлення інформації. Як правило, для моделювання вітрин (кіосків) даних використовуються типи моделі під назвою: схема «зірка» і схема «сніжинка». Зупинимося докладніше на кожному з цих типів моделей.

Схема «зірка» - популярний тип моделі даних для вітрин даних. Дана модель характеризується наявністю таблиці фактів, оточеної пов'язаними з нею таблицями розмірностей. Запити до такої структури включають прості об'єднання таблиці фактів з кожною з таблиць розмірностей. Характеризується високою продуктивністю запитів. Проектується для виконання аналітичних запитів. Характеризується невеликою надмірністю даних і високою в порівнянні з нормалізованими структурами продуктивністю. Деякі промислові СУБД і інструменти класу OLAP/Reporting уміють використовувати переваги схеми «зірка» для скорочення часу виконання запитів.

Розглянемо компоненти схеми «зірка».

Розмірності. У технології багатовимірного моделювання розмірність - це аспект, в розрізі якого можна отримувати, фільтрувати, групувати і відображати інформацію про факти. Типові розмірності, що зустрічаються практично в будь-якій моделі:

- Клієнт
- Продукт
- Час
- Географія
- Співробітник

Розмірності, як правило, мають багаторівневу ієрархічну структуру. Наприклад, розмірність ЧАС може мати наступну структуру: РІК – КВАРТАЛ - МІСЯЦЬ - ДЕНЬ.

Факти. Факти - це зазвичай числові величини, що зберігаються в таблиці фактів і є предметом аналізу. Приклади фактів: об'єм операцій, кількість проданих одиниць товару і так далі. Факти мають ряд властивостей, на яких ми коротко зупинимося.

Адитивні факти. Адитивність визначає можливість підсумування факту уздовж певної розмірності. Такі факти можна підсумовувати і групувати уздовж всієї розмірності на будь-яких рівнях ієрархії.

Напівадитивні факти. Напівадитивний факт — це факт, який можна підсумовувати уздовж певної розмірності, і не можна — уздовж інших. Прикладом може служити залишок на рахунку (або залишок товару на складі). Дану величину не можна підсумовувати уздовж розмірності ЧАС. Проте сума залишків по рахунках уздовж розмірності є предметом для аналізу. Фахівці рекомендують моделювати напівадитивні факти так, щоб зробити їх більш адитивними. Наприклад, представити відсоток складовими його величинами.

Неадитивні факти. Неадитивні факти взагалі не можна підсумовувати. Приклад неадитивного факту — відношення (наприклад, виражене у відсотках).

Таблиці покриття. Таблиці покриття використовуються з метою моделювання поєднання розмірностей, для яких відсутні факти. Наприклад, потрібно знайти кількість категорій продуктів, які сьогодні жодного разу не продавалися. Таблиця фактів продажів не може відповісти на дане питання, оскільки вона реєструє лише факти продажів. Для того, щоб модель дозволяла відповідати на подібні питання, потрібна додаткова таблиця фактів (яка, по суті справи, не містить фактів).

Схема «сніжинка» використовується для нормалізації схеми «зірка». Вона декілька скорочує надмірність в таблицях розмірності. Одним з достоїнств є швидше виконання запитів про структуру розмірності (запити вигляду «Вибрати всі рядки з таблиці розмірності на певному рівні»), які дуже часто виконуються при аналізі даних, і можуть затримувати хід аналізу. Проте основною відмінністю схеми !сніжинка! є не економія дискового простору, а можливість мати таблиці фактів з різним рівнем деталізації. Наприклад, фактичні дані на рівні дня, а планові — на рівні місяця.

Методика побудови вітрин (кіосків) даних з простої теоретичної дисципліни поступово перетворюється на складну науку, повну варіацій і напрямів. Якщо раніше було відомо лише про EDW (Enterprise Data Warehouse), то тепер з’явилися поступово розвиваємі вітрини даних (Incremental Architected Data Mart, ADM), розподілені вітрини (кіоски) даних (Distributed Data Mart, DDM), об'єднані вітрини даних (Federated Data Mart, FDM). Розглянемо деякі з цих нових напрямків.

Системи об'єднаних вітрин даних. У багатьох організаціях склалася практика реалізації багаточисельних сховищ даних. Хоча, за визначенням, існує лише одне сховище даних, а всі останні об'єкти є його підмножиною або вітринами (кіосками) даних, що поступово розвиваються, але не всі організації дотримуються цього правила. Таким чином, в багатьох компаніях існує два, три, десяток і навіть більше систем сховищ даних. Поширення сховищ даних привело до розвитку архітектури сховища даних підприємства, а саме: до появи об'єднаних систем сховищ даних або вітрин (кіосків) даних.

Система об'єднаних вітрин даних характеризується спільним використанням загальних інформаційних ресурсів, усуваючи, таким чином, надмірність і гарантуючи достовірність інформації по всій організації.

Позитивними рисами об'єднаних вітрин даних є: загальна семантика бізнес-правила; один набір процесів витягання і бізнес-правил; децентралізовані ресурси і управління; паралельна розробка.

Недоліками такого архітектурного рішення є: необхідність в координуванні робіт; складнощі в подоланні «політичних» моментів і вирішенні питань авторських прав; потрібна узгодженість серед різних відділів по питаннях архітектури, бізнес-правил і семантики; складне технічне середовище; наявність багаточисельних репозиторіїв метаданих.

Непроектуємі вітрини даних. Поява непроектуємих вітрин даних (Non-Architected Data Marts) пояснюється, перш за все, складнощами, пов'язаними з реалізацією систем EDW і FDW. Брудні і швидко отримувані набори даних не піддаються очищенню і, отже, не можуть використовуватися для подальшої інтеграції з будь-якими іншими джерелами даних систем сховищ даних. Дуже швидко вони перетворюються на застарілі системи, які лише додають проблем, не вирішують їх. Для цих систем характерні багаточисельні процеси витягання, безліч бізнес-правил, невірогідність інформації.

Позитивними рисами непроектуємих вітрин даних є: висока продуктивність; низька вартість. Недоліками таких систем є: недостовірна інформація; багаточисельні процеси витягання; багаточисельні бізнес-правила; підвищена складність при інтеграції.

Система вітрин (кіосків) даних, що поступово розвиваються. Ця архітектура є альтернативою сховища даних підприємства. Для наповнення таких вітрин зазвичай використовується інструментальний засіб класу підприємства, що реалізовує стратегію «витягаєш один раз, наповнюєш багато».

Достоїнствами таких вітрин даних є: єдиний набір процесів витягання; здійснимий масштаб. Недоліки: найбільш ефективні при використанні інструментального засобу класу підприємства; необхідність в архітектурі вітрин даних підприємства (Enterprise Data Mart Architecture, EDMA).

Методика побудови вітрин (кіосків) даних виявилася напрямом ринку проектів інтелектуального аналізу даних, що нестримно розвивається, швидко змінюється. Якщо раніше не було механізмів їх ефективного проектування, і був лише один способів їх створення, в даний час можна знайти незчисленне число таких інструментів і ряд технологій життєздатної архітектури таких систем. За умови вибору відповідної архітектури і належного підходу до проекту можна побудувати систему сховища та вітрин даних, яка забезпечить не лише високе повернення інвестицій, але і значно підвищить ефективність функціонування всього підприємства.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. Охарактеризуйте основні властивості сховищ даних.
2. Суть інтеграцій при створення сховищ даних.
3. Методи інтеграції даних: консолідація даних.
4. Методи інтеграції даних: федералізація даних.
5. Методи інтеграції даних: розповсюдження даних.
6. Основні типи архітектури даних.
7. Корпоративні сховища і їх різновиди.
8. Основні елементи сховища даних типу «зірка».
9. Основні елементи сховища даних типу «сніжинка».