

ЗАДАЧІ СХОВИЩ ДАНИХ

Лекція 2

Автор Голуб Б.Л.



ПРИЧИНИ ПОЯВИ СД

- аналітична робота з даними в СД (спеціалізованих БД) не повинна позначатися на продуктивності основних БД;
- аналітики та працівники управління повинні мати можливість повністю орієнтуватися на спеціалізовані сховища в режимі

" Що , якщо ..."



ПРИЧИНИ НЕМОЖЛИВОСТІ ВИКОРИСТАННЯ ТРАДИЦІЙНИХ БД

- недостовірність даних;
- низька продуктивність при нестандартних запитах;
- неможливість перетворення різнорідних даних, оскільки вони часто не мають позначок часу.

Проблеми при підготовці звітів виникають через те, що:

- ✓ важко зрозуміти , де знаходяться дані , необхідні для аналізу і ухвалення рішення ;
- ✓ більшість БД орієнтовані лише на стандартні запити;
- ✓ потрібно залучати програмістів для виконання нестандартних запитів.



ДОСВІД ВИКОРИСТАННЯ БД

- Незважаючи на велику кількість даних, можливостей їх збору та зберігання, організації досі відчують серйозний недолік в інформації, необхідної для прийняття рішень.
- Існуючі системи збору та обробки корпоративних даних в принципі не придатні для використання в ППР. Дані різнотипні і розподілені як всередині організації, так і за її межами.
- Особам, які приймають рішення, і аналітикам доводиться приймати рішення не тільки в умовах неповної, а й часто недостовірної та суперечливої інформації. До того ж не завжди вдається отримати необхідну інформацію своєчасно і в наочному вигляді.

У результаті - невдалі рішення.



ТАКИМ ЧИНОМ:

*Виникає необхідність в технологіях,
що дозволяють автоматично збирати дані з
різних баз даних , систем обробки даних,
погоджувати і об'єднувати в
предметно-орієнтований формат, який
потрібен аналітикам*



Вимоги до СД

СД має бути *предметно - орієнтованим, інтегрованим, призначеним для підтримки прийняття рішень* .

Сховище являє собою таке середовище накопичення даних, яке оптимізоване для виконання складних аналітичних запитів управлінського персоналу. Ці запити можуть бути досить індивідуальні для кожної організації, кожного підрозділу і навіть окремого аналітика.



ПРЕДМЕТНА ОРІЄНТОВАНІСТЬ

- **Додатки** завжди оперують функціями, такими, наприклад , як відкриття рахунку, кредитування , виписування накладної , зарахування на рахунок тощо.
- **Сховище** даних організовано навколо фактів і предметів , таких , як угода, сума кредиту, покупець , постачальник , продукт тощо.



ІНТЕГРОВАНІСТЬ

Інтегрованість проявляється у багатьох аспектах:

- в узгодженості імен,
- в узгодженості одиниць вимірювання змінних,
- в узгодженості структур даних,
- в узгодженості фізичних атрибутів даних тощо.



ЗАЛЕЖНІСТЬ ВІД ЧАСУ

- Всі дані у сховищі в певний момент часу сумісні.
- Інший прояв залежності сховища даних від часу полягає в його структурі. Кожна структура сховища включає - явно чи неявно - елемент часу.
- Третій прояв залежності сховища даних від часу складається у виконанні правила, що дані, одного разу коректно в сховище записані, не можуть бути поновлені.



ПОСТІЙНІСТЬ

- Модифікація даних не проводиться, оскільки може привести до порушення їх цілісності.
- Оскільки не потрібно мінімізувати час занурення, то структура сховища може бути оптимізована для обробки певних запитів, що досягається за рахунок денормалізації реляційної схеми, попереднього агрегування і побудови відповідних індексів.



ОСОБЛИВОСТІ СД

- Сховища даних містять інформацію, зібрану з декількох оперативних баз даних. Сховища, як правило, на порядок більше оперативних баз, часто маючи обсяг від сотень гігабайт до декількох терабайт.
- Як правило, сховище даних підтримується незалежно від оперативних баз даних організації, оскільки вимоги до функціональності і продуктивності аналітичних додатків відрізняються від вимог до транзакційних систем.
- Сховища даних створюються спеціально для додатків підтримки прийняття рішень і надають накопичені за певний час, зведені і консолідовані дані, які більш прийнятні для аналізу, ніж детальні індивідуальні записи. Робоче навантаження складається з нестандартних, складних запитів, які звертаються до мільйонів записів і виконують величезну кількість сканувань, з'єднання і агрегування. Час відповіді на запит в даному випадку важливіше, ніж пропускна здатність.



ЕТАПИ

- вилучення і перетворення;
- очищення даних;
- завантаження;
- оновлення;
- управління метаданими.



ЕТАП ВИЛУЧЕННЯ ТА ПЕРЕТВОРЕННЯ

- Мета етапу вилучення даних - перенести дані з різнорідних джерел в базу даних, де їх можна модифікувати і додати в сховище
- Мета етапу перетворення даних - усунути невідповідності в схемі і угодах щодо значень атрибутів. Набір правил і скриптів, як правило, виконує перетворення даних з вихідної схеми в підсумкову схему



ЕТАП ОЧИЩЕННЯ ДАНИХ

- Помилки при введенні даних і відмінності в схемах можуть привести до того, що таблиця вимірювань «Клієнт» матиме кілька відповідних кортежів для одного клієнта, що призводить до неточних відповідей на запити і некоректним моделям видобутку даних.
- Інструменти, які допомагають визначити і виправити аномалії даних, повинні мати високу віддачу.



ЕТАП ЗАВАНТАЖЕННЯ

- Після того, як дані витягнуті і перетворені, можливо, що їх ще необхідно додатково обробити перед тим, як додати в сховище
- Як правило, утиліти фонового завантаження підтримують такі функції, як перевірка обмежень цілісності; сортування; підсумовування, агрегування і виконання інших обчислень для створення похідних таблиць, що розміщуються в сховище; створення індексів і інших способів доступу
- Крім наповнення сховища, утиліта завантаження повинна дозволяти системним адміністраторам перевіряти статус; скасовувати, призупиняти та відновлювати завантаження; відновлювати роботу після помилки без втрати цілісності даних



КОЛИ ОНОВЛЮВАТИ І ЯК ОНОВЛЮВАТИ

- Зазвичай сховища даних оновлюються періодично відповідно до заздалегідь встановленому розкладу, наприклад, щодня або щотижня
- Адміністратори сховища даних визначають правила оновлення в залежності від вимог користувачів і трафіку. Розклад оновлень може бути різним для різних джерел даних.



МЕТАДАНІ

- Метадані - інформація будь-якого роду, яка потрібна для управління сховищем даних, а управління метаданими - істотний компонент архітектури зберігання
- Метадані сховища часто розміщуються в репозиторії, який дозволяє спільно використовувати метадані різних інструментів і процесів при проектуванні, установці, використанні, експлуатації та адмініструванні сховища
- *До адміністративних метаданих відноситься вся інформація, яка потрібна для налаштування і використання сховища даних*
- *Бізнес-метадані включають в себе бізнес-терміни і визначення, належність даних і правила оплати послуг сховища*
- *Оперативні метадані - це інформація, зібрана під час роботи сховища даних, така як походження перенесених і перетворених даних; статус використання даних (активні, заархівовані або віддалені); дані моніторингу, такі як статистика використання, повідомлення про помилки і результати аудиту*

