**Лекція №11. Розподілена обробка даних. Системи аналітичної обробки даних *OLAP.* Сховища даних та *OLAP***

Класична реляційна модель передбачає неподільність даних, що зберігаються в полях записів таблиць. Це означає, що інформація в таблиці представляється в першій нормальній формі. Існує ряд випадків, коли це обмеження заважає ефективній реалізації додатків. Постреляціонних модель даних (ПРМД) являє собою розширену реляційну модель, яка знімає обмеження неподільності даних, що зберігаються в записах таблиць. ПРМД допускає багатозначні поля - поля, значення яких складаються з підзначень. Набір значень багатозначних полів вважається самостійною таблицею, вбудованої в основну таблицю.

Крім забезпечення вкладеності полів ПРМД підтримує асоційовані багатозначні поля (множинні групи). Сукупність асоційованих полів називається асоціацією. При цьому в рядку перше значення одного стовпця асоціації відповідає першими значенням всіх інших стовпців асоціації. Аналогічним чином пов'язані всі другі значення стовпців і т.д. На довжину полів і кількість полів в записах таблиці не накладає вимога сталості. Це означає, що структура даних і таблиць має велику гнучкість.

Оскільки ПРМД допускає зберігання в таблицях ненормалізованих даних, виникає проблема забезпечення цілісності і несуперечності даних. Ця проблема вирішується включенням в СУБД механізмів, подібних збереженим процедурам в клієнт-серверних системах. Для опису функцій контролю значень в полях є можливість створювати процедури (коди конверсії та коди кореляції), автоматично викликаються до або після звернення до даних. Коди кореляції виконуються відразу після читання даних, перед їх обробкою. Коди конверсії, навпаки, виконуються після обробки даних.

Перевагою ПРМД є можливість представлення сукупності пов'язаних реляційних таблиць однієї постреляціонной таблицею. Це забезпечує високу наочність представлення інформації і підвищення ефективності її обробки.

Недоліком ПРМД є складність вирішення проблеми забезпечення цілісності і несуперечності збережених даних.

***Багатомірний підхід до представлення даних***

Багатомірний підхід до представлення даних в базі з'явився практично одночасно з реляційних, але реально працюючих багатовимірних СУБД (БСУБД) до середини 1990-х рр. було дуже мало.

Поштовхом до використання багатовимірної моделі даних (БМД) послужила опублікована в 1993 р програмна стаття одного з основоположників реляційного підходу Е. Кодда. У ній сформульовані 12 основних вимог до систем класу OLAP (OnLine Analytical Processing - оперативна аналітична обробка), найважливіші з яких пов'язані з можливостями концептуального подання та обробки багатовимірних даних. Багатовимірні системи дозволяють оперативно обробляти інформацію для проведення аналізу та прийняття рішення.

У розвитку концепцій OLAP можна виділити наступні два напрямки:

* системи оперативної (транзакционной) обробки;
* системи аналітичної обробки (системи підтримки прийняття рішень).

Реляційні СУБД призначалися для інформаційних систем оперативної обробки інформації і в цій області є дуже ефективнтими. У системах аналітичної обробки вони показали себе кілька неповороткими і недостатньо гнучкими. Більш ефективними тут виявляються багатовимірні СУБД (БСУБД), які є вузькоспеціалізованими СУБД, призначеними для інтерактивної аналітичної обробки інформації. З ними пов’язане поняття сховищ даних. Розкриємо основні поняття, використовувані в цих СУБД.

***Сховище даних*** – це предметно-орієнтований, інтегрований, прив'язаний до часу незмінний набір даних, призначений для підтримки та прийняття рішень.

Характеристики сховища даних:

* *предметна орієнтованість* – властивість, яка відображає необхідність збереження даних, призначених для підтримки прийняття рішень, а не звичайних оперативно-прикладних даних;
* *інтегрованість* – властивість, яка відображає необхідність створення інтегрованого джерела, яке забезпечує узгодженість інформації, що зберігається, задля надання користувачеві єдиного узагальненого представлення даних.
* *прив'язка до часу* – властивість, яка відображає необхідність прив’язки даних до деякого моменту чи проміжку часу задля їх коректності та точності.
* *незмінюваність* – властивість, яка відображає регулярне поповнення даних за рахунок інформації з оперативних систем обробки, а не їх оновлення в оперативному режимі.

***Технологія сховищ даних*** – це технологія управління даними і їх аналізу.

Переваги використання технології сховищ даних:

* потенційно висока віддача від інвестицій;
* підвищення конкурентоспроможності;
* підвищення ефективності праці осіб, відповідальних за прийняття рішень.

Проблеми сховищ даних:

* недооцінювання ресурсів, необхідних для завантаження даних;
* приховані проблеми джерел даних;
* відсутність необхідних даних у наявних архівах;
* підвищення вимог кінцевих користувачів;
* гомогенизація даних;
* високі вимоги ресурсів;
* володіння даними;
* складний супровід;
* довготривалий характер проектів;
* складнощі інтеграції.

**Архітектура сховищ даних**

Основні компоненти сховища даних:

1. ***Оперативні дані*** – дані, які розміщуються в сховищі.
2. ***Менеджер завантаження*** – це зовнішній компонент, який виконує всі операції, пов'язані з отриманням і завантаженням даних в сховище.
3. ***Менеджер сховища*** – це компонент, який виконує всі операції, пов'язані з управлінням інформацією, розміщеною в сховищі даних. Операції менеджера сховища:

* аналіз несуперечності даних;
* перетворення і переміщення вихідних даних з тимчасового сховища в основні таблиці сховища даних;
* створення індексів і представлень для базових таблиць;
* денормалізація даних (в разі необхідності);
* узагальнення даних (в разі необхідності);
* резервне копіювання та архівування даних.

1. ***Менеджер запитів*** – це внутрішній компонент, який виконує всі операції, пов'язані з управлінням призначеними для користувача запитами.
2. ***Детальні дані*** – дані, що описані в схемі бази даних.
3. ***Частково та глибоко узагальнені дані*** – всі дані, що попередньо оброблені менеджером сховища з метою їх часткового або глибокого узагальнення.
4. ***Архівні та резервні копії*** – компонент сховища, який відповідає за підготовку детальної і узагальненої інформації для розміщення в резервні та архівні копії.
5. ***Метадані*** – компонент сховища, в якому зберігаються всі метадані, що використовуються будь-якими процесами сховища.
6. ***Засоби доступу до даних кінцевого користувача***. Основні групи інструментів користувачів для доступу до даних:

* інструменти створення звітів і запитів;
* інструменти розробки додатків;
* інструменти інформаційної системи керівника;
* інструменти оперативної аналітичної обробки (ОLAР-інструменти);
* інструменти розробки даних.

***OLAP-інструменти*** – це інструменти оперативної аналітичної обробки даних, які створюються на основі концепції багатовимірної бази даних та дозволяють кваліфікованим користувачам аналізувати дані за допомогою складних багатовимірних представлень.

***Розробка даних*** – це процес відкриття нових усвідомлених кореляцій, розподілів і тенденцій шляхом переробки величезної кількості інформації, витягнутої із сховища або магазину даних, з використанням статистичних і математичних методів, а також методів штучного інтелекту.

**Інформаційні потоки в сховищі**

***Вхідний потік*** – це процеси, пов'язані з отриманням, очищенням та завантаженням інформації з джерел даних у сховище даних.

***Висхідний потік*** – це процеси, пов'язані з підвищенням цінності даних, що зберігаю у сховищі, за допомогою узагальнення, упаковки та розподілення вихідних даних.

***Низхідний потік*** – це процеси, пов'язані з архівуванням та резервним копіюванням інформації у сховищі даних.

***Вихідний потік*** – це процеси, пов'язані з наданням даних користувачам.

***Метапотік*** – це процеси, пов'язані з управлінням метаданими.

В технології сховищ даних головна увага приділяється керуванню п‘ятьма основними інформаційними потоками: вхідним, висхідним, низхідним, вихідним та метапотоком. Місце кожного з цих потоків в інформаційній ситемі в [[[1]](#footnote-1), с.1241-1245] визначається схемою, наведеною на рис.1, яка була адаптована під макроструктуру інтегрованої статистичної інформаційної системи (ІСІС).

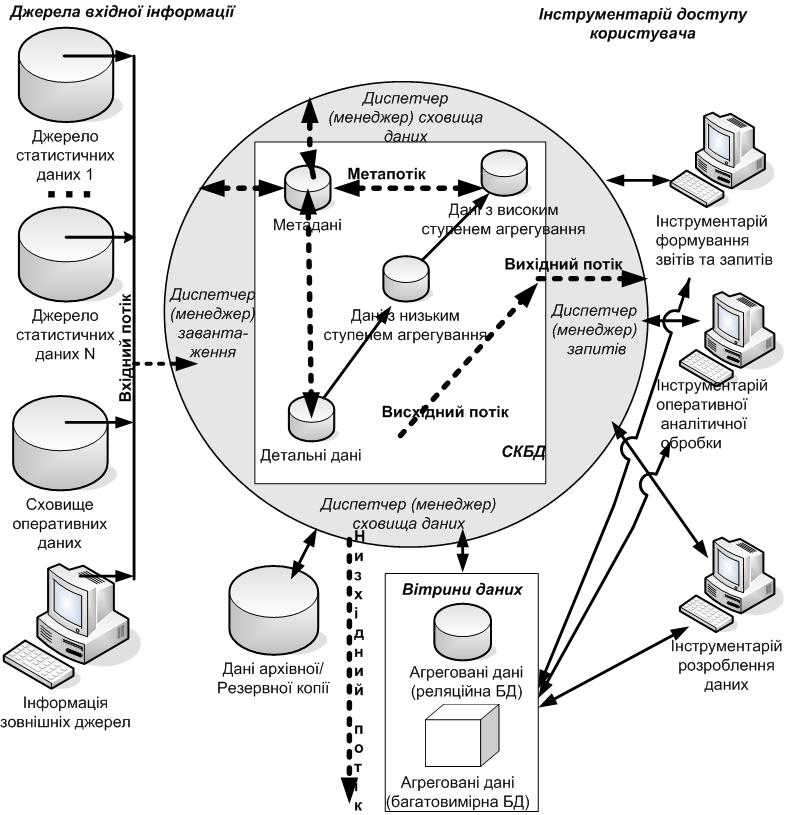


Рис.1. Типова архітектура сховища даних і вітрини даних

На схемі інформаційні потоки зв‘язані з конкретними інформаційними складовими і компонентами сховища даних, які спрощено позначено на схемі як "диспетчери (менеджери)". Компоненти сховища даних можуть складатися з однієї або більше БД, збудованих з використанням реляційної СКБД, багатовимірної СКБД (OLAP-системи - засоби оперативної аналітичної обробки) або обох типів. Можливість такого поділу на різні БД окреслена на схемі.

Джерелами вхідної інформації є оперативні та архівні дані з різних статистичних задач, що вирішуються "власником" ІСІС, і зовнішні джерела, наприклад, статистична інформація з Інтернет-сайтів, аналітичні таблиці науково-дослідних установ. Вхідний потік пов‘язаний з вибіркою інформації із джерел даних з метою подальшого завантаження до сховища даних.

Диспетчер завантаження забезпечує гнучке використання операцій виділення, фільтрування, перевірки достовірності, вибирання даних із зовнішніх джерел і їх перебудови. Він повинен налаштовуватися на певну схему опрацювання вхідних даних згідно з метаописами репозитарія метаданих, який для спрощення викладення позначений на схемі як "метадані", маючи на увазі і наявність відповідного інструментарію. Диспетчер завантаження може бути зовнішнім компонентом сховища даних, а не тільки вбудованим, як показано на схемі.

Висхідний потік відноситься до інформації, яка формується зі збережених в сховищі даних шляхом підсумовування, документування і розподілу початкових даних.

Підтримку висхідного потоку забезпечує диспетчер сховища даних - внутрішній функціональний компонент сховища даних, який виконує операції аналізу та керування даними, використовуючи метаописи. До цих операцій відносяться аналіз узгодженості та відсутності суперечливості даних, перетворення та переміщення даних (наприклад, з локального сховища до основних таблиць сховища даних), створення індексів, денормалізація даних за необхідністю, часткове чи глибоке узагальнення даних, резервне копіювання та архівування даних.

На рис. 1 показано процес перетворення висхідного потоку від детальних даних до даних з високим ступенем агрегування. Сукупність даних, отримана на кожному етапі перетворення, може мати певні ознаки, за яким виконується логічний поділ сховища даних на певні інформаційні розділи. В загальних рисах склад інформаційних розділів є таким.

Детальні дані – це розділ, де зберігаються усі детальні дані: і первинні найнижчого рівня деталізації (звітні форми респондентів), і узагальнені до певного рівня агрегування (макромасиви). Дані з низьким і високим ступенем агрегування – це розділи, де містяться дані, за допомогою диспетчеру сховища попередньо оброблені з метою їх часткового чи глибокого узагальнення. В цих розділах зберігаються у певний спосіб відсортовані та згруповані дані, необхідні для виконання обчислень. Розділ, до якого відносяться похідні показники, має визначатися сукупністю їх якісних ознак, наприклад, якщо до даних з низьким ступенем агрегування відносяться абсолютні показники по регіонах, то в результаті виконання розрахунку середнього по країні отримуємо показник, що відноситься до даних з високим ступенем агрегування. Склад розділів з агрегованими даними може змінюватися в залежності від запитів користувачів, оскільки узагальнені дані поновлюються у міру надходження нових даних до системи.

Низхідний потік пов‘язаний з операціями архівування та резервного копіювання інформації в сховищі. Архівування застарілих даних відіграє важливу роль для забезпечення високої ефективності і продуктивності сховища даних завдяки перенесенню застарілих даних на архівний носій (оптичні диски, магнітні стрічки тощо). Резервне копіювання передбачає процедури, які забезпечують можливість відновлення поточного стану сховища у випадку втрати даних із-за збоїв в програмному та апаратному забезпеченні.

Вихідний потік – це інформація, яка надається користувачу за допомогою диспетчеру запитів – внутрішнього елементу сховища даних, що виконує операції пов‘язані з керуванням запитами користувачів.

Метапотік (метадані) – це інформація репозитарія метаданих, яка обслуговується операціями, пов‘язаними з підтримкою метаданих, що стосуються інших потоків. Головним призначенням метапотоку є опис потоків та процесів, які перебігають в ІСІС (наприклад, описи джерел даних, процеси перебудови вхідних даних, інтеграції та узагальнення тощо). Метапотік повинен постійно поновлюватися у відповідності до якісних змін, що відбуваються в інших потоках.

Зі сховищами даних тісно пов’язана ще одна інформаційна структура - вітрини даних, які можуть містити і детальні дані, і агреговані дані, і відповідну до них метаінформацію. Вітрини даних є спеціально сформованими фрагментами сховища даних відповідно до потреб конкретних груп користувачів.

**Інструменти та технології сховищ**

Інструменти і технології, які використовуються при створенні і супроводі сховищ даних:

* інструменти отримання, очищення та перетворення даних;
* СУБД для сховища даних;
* метадані сховища даних;
* інструменти управління та адміністрування.

Інтегровані рішення для отримання, очищення та перетворення даних з наступним завантаженням у конкретну систему:

* генератори коду;
* інструменти реплікації інформації бази даних;
* механізми динамічного перетворення.

Вимоги до СУБД для сховища даних:

* висока продуктивність завантаження даних;
* можливість обробки даних під час завантаження;
* наявність засобів управління якістю даних;
* висока продуктивність запитів;
* масштабованість за розміром;
* масштабованість за кількістю користувачів;
* можливість організації мережі сховищ даних;
* наявність засобів адміністрування сховища;
* підтримка інтегрованого багатовимірного аналізу;
* розширений набір функціональних засобів запитів.

Основне призначення метаданих – це збереження інформації про виникнення даних з метою надання можливості адміністраторам сховища знати історію будь-якого елементу даних, який розміщений у сховищі.

Метадані, пов'язані з перетворенням та завантаженням даних, повинні описувати джерело даних та будь-які зміни, внесені в ці дані.

Метадані, пов'язані з управлінням даними, повинні описувати спосіб збереження даних у сховищі.

Операції, які повинні дозволяти виконувати інструменти управління і адміністрування сховищ даних:

* моніторинг завантаження даних з декількох джерел;
* перевірка якості і цілісності даних;
* управління та оновлення метаданих;
* моніторинг поточної продуктивності бази даних з метою забезпечення швидкої реакції на запит та ефективного використання ресурсів;
* аудит процесів використання сховища даних з метою збору інформації про вартість виконаної кожним користувачем роботи;
* реплікація, розбиття та розподіл даних;
* підтримка ефективного управління сховищем даних;
* очищення даних;
* архівування і резервне копіювання даних;
* засоби відновлення після збою;
* управління засобами захисту.

**Проектування сховищ**

Основними задачами проектування сховищ даних є визначення кандидатів на змінні, факти та виміри. Для цього застосовуються кілька підходів, кожен з яких характеризується певною послідовністю ідентифікації основних елементів моделі. Згідно з послідовністю ідентифікації елементів сховища даних можна визначити такі підходи до визначення основних елементів сховища даних:

**Підхід «від запиту».** Передусім визначаються змінні, потім виміри, пов'язані зі змінними, а надалі формуються факти. Цей підхід називається «від запиту», оскільки він орієнтований насамперед на аналітичні запити до сховища даних. Тобто визначення елементів сховища даних виконується на основі аналізу запитів користувачів-аналітиків.

**Підхід, орієнтований на бізнес.** Визначаються факти, потім виміри, а на завершення змінні. Цей підхід називається орієнтованим на бізнес, оскільки спочатку аналізується предметна область,

При проектуванні СД найбільш поширеними є схеми "зірка" та "сніжинка" (розглядали на ЛР), а також "зірка-сніжинка".

***Схема "зірка"*** – це логічна структура, в центрі якої знаходиться таблиця фактів (з детальними даними), оточена таблицями розмірності (з даними-посиланнями). ***Схема "сніжинка"*** – це варіант схеми "зірка", в якому кожна розмірність може мати свої власні розмірності.

***Схема "зірка-сніжинка"*** – це гібридна структура, яка включає комбінацію денормалізованої схеми "зірка" і нормалізованої схеми "сніжинка".

Чинники, необхідні для оптимального проектування бази даних:

* визначення необхідного часу відгуку для кожного додатку підтримки прийняття рішення;
* пошук компромісу між необхідністю використання статистичних вибірок підмножин даних та необхідністю обробки детальних відомостей;
* визначення стовпців, що видаляються;
* скорочення розміру стовпців таблиці фактів;
* визначення найкращого способу застосування зовнішніх ключів, які налаштовуються та не налаштовуються;
* визначення оптимального підходу для введення в таблицю фактів розмірності "час";
* секціонування таблиць фактів для поліпшення їх керованості.

**Інтерактивна аналітична обробка даних (OLAP)**

Якийсь час назад OLAP-продукти підтримували або реляційне, або багатовимірне зберігання. Сьогодні, як правило, один і той же продукт забезпечує обидва ці види зберігання, а також третій вигляд - змішаний. Застосовуються наступні терміни:

MOLAP (Multidimensional OLAP) - і детальні дані, і агрегати зберігаються в багатовимірній БД. В цьому випадку виходить найбільша надмірність, оскільки багатовимірні дані повністю містять реляційні.

ROLAP (Relational OLAP) - детальні дані залишаються там, де вони "жили" спочатку - в реляційній БД; агрегати зберігаються в тій же БД в спеціально створених службових таблицях.

HOLAP (Hybrid OLAP) - детальні дані залишаються на місці (в реляційній БД), а агрегати зберігаються в багатовимірній БД.

Кожний з цих способів має свої переваги і недоліки і повинен застосовуватися залежно від умов - об'єму даних, потужності реляційної СУБД і т.д. При зберіганні даних в багатовимірних структурах виникає потенційна проблема "розбухання" за рахунок зберігання порожніх значень. Адже якщо в багатовимірному масиві зарезервовано місце під все можливі комбінації міток вимірювань, а реально заповнена лише мала частина (наприклад, ряд продуктів продається тільки в невеликому числі регіонів), то велика частина куба порожнітиме, хоча місце буде зайнято. Сучасні OLAP-продукти уміють справлятися з цією проблемою.

***Оперативна аналітична обробка*** (OLAP) – це динамічний синтез, аналіз та консолідація великих об'ємів багатовимірних даних. Основні аналітичні операції серверів багатовимірних баз даних на основі OLAP:

*Консолідація*. Операція включає такі узагальнюючі операції, як просте підсумовування значень ("згортка") або розрахунок використанням складних виразів, які включають інші пов'язані дані.

*Низхідний аналіз* ("drill-down"). Операція, зворотна консолідації, яка включає відображення докладних відомостей для розглянутих консолідованих даних.

*Розбиття з поворотом* (slicing and dicing) або зведена таблиця. Операція дозволяє отримати представлення даних з різних точок зору.

Правила для ОLАР-систем:

1. Багатовимірне концептуальне представлення даних.

2. Прозорість.

3. Доступність.

4. Незмінна продуктивність підготовки звітів.

5. Архітектура "клієнт / сервер".

6. Універсальність вимірювань.

7. Динамічне управління розрідженістю матриць.

8. Розрахована на багато користувачів підтримка.

9. Необмежені перехресні операції між розмірностями.

10. Підтримка інтуїтивно зрозумілого маніпулювання даними.

11. Гнучкість засобів формування звітів.

12. Необмежена кількість вимірювань та рівнів узагальнення.

Основні категорії ОLAР-інструментів:

багатовимірні ОLAР-інструменти;

реляційні ОLAР-інструменти;

кероване середовище запитів.

**Технологія розробки даних**

***Розробка даних –*** процес вилучення з великих баз даних достовірної, попередньо невідомої, комплексної та значимої інформації і використання її для прийняття відповідальних бізнес-рішень. Основні операції методів розробки даних:

прогнозуюче моделювання;

сегментування баз даних;

аналіз зв'язків;

виявлення відхилень.

Методи прогнозуючого моделювання:

класифікація;

прогнозування значення.

***Класифікація*** використовується для встановлення спеціального зумовленого класу для кожного запису в базі даних на основі обмеженого набору можливих значень класу. Спеціалізації методу класифікації:

деревовидна індукція;

нейронна індукція.

Метою ***сегментування бази*** даних є її розбиття на деяку, заздалегідь невідому кількість сегментів, або кластерів, що складаються з подібних записів, тобто записів, які мають деякі загальні властивості, що дозволяє вважати їх однорідними. Метод сегментування бази даних також пов'язаний з методами демографічної та нейронної кластеризації, які відрізняються допустимими вхідними даними, методами розрахунку відстані між записами і представленням результуючих сегментів.

Метою ***аналізу зв'язків*** є встановлення зв'язків, або асоціацій, між окремими записами або наборами записів в базі даних. Спеціалізації методу аналізу зв’язків:

пошук асоціацій;

пошук послідовних закономірностей;

пошук аналогічних часових послідовностей.

Метод ***виявлення відхилень*** передбачає аналіз викидів, які висловлюють ступінь відхилення від деяких попередньо відомих значень математичного очікування і нормального середнього.

Важливі характеристики інструментів розробки даних:

наявність інструментів підготовки даних;

можливість вибору операцій розробки даних (алгоритмів);

масштабованість продукту і показники його продуктивності;

наявність інструментів візуалізації даних.

**Контрольні запитання**

1. Що таке OLAP-інструменти?

2. Що таке розробка даних?

3. Назвіть правила та поясніть для ОLАР-систем.

4. Назвіть основні категорії ОLAР-інструментів.

5. Які основні операції методів розробки даних?

6. Які основні категорії OLAP-інструментів вам відомі?

7. Які важливі характеристики інструментів розробки даних вам відомі?

8. Що таке OLAP-інструменти?

9. Що таке розробка даних?

10. Назвіть правила та поясніть для ОLАР-систем.

11. Назвіть основні категорії ОLAР-інструментів.

12. Які основні операції методів розробки даних?

13. Які основні категорії OLAP-інструментів вам відомі?

14. Які важливі характеристики інструментів розробки даних вам відомі?

1. Конолли Томас, Бегг Каролин. Базы данных. Проектирование, реализация и сопровождение. Теория и практика. 3-е издание: Пер.с англ. – М., Издательский дом "Вильямс",2003.-1440с.- стор.1241-1245 [↑](#footnote-ref-1)