

Тема 12: Основні способи отримання та представлення даних

Області застосування баз даних

Історично системи управління базами даних орієнтувалися на вирішення завдань, пов'язаних у першу чергу з транзакційною обробкою структурованої інформації. Безумовно, найкращим, перевіреним часом рішенням тут була і залишається реляційна модель СУБД. Однак в останні роки область застосування баз даних незмінно розширювалася. З одного боку, потрібно керувати більш широким набором форматів даних, переходячи до вирішення спільних проблем управління корпоративною інформацією. З іншого - саме СУБД беруть на себе основні функції інтеграції даних і додатків корпоративних систем. (За даними Gartner Group, інформаційні відділи підприємств витрачають до 40% свого бюджету на вирішення завдань інтеграції діючих компонентів баз даних.) Саме цим пояснюється активний інтерес до обговорення архітектурних принципів і можливостей реалізації баз даних різних моделей - постреляційних, об'єктно-реляційних, XML.

Якщо постаратися класифікувати існуючі області застосування баз даних, а так само оцінити перспективи їхнього розвитку в даний час, то можна отримати **приблизний список найбільш поширених класів**:

- документографічні й документальні застосовуються у всіх базах органів влади та управління;
- бази даних з промислової, будівельної та сільськогосподарської продукції;
- бази даних з економічної та кон'юнктурної інформації (статистична, кредитно-фінансова, зовнішньоторговельна);
- фактографічні бази соціальних даних, які включають відомості про населення і про соціальні середовища;
- бази даних транспортних систем;
- довідкові дані для населення та установ (енциклопедії та довідники, розклади літаків і поїздів, адреси та телефони громадян і організацій);
- ресурсні бази даних, що включають фактографічну інформацію про природні ресурси (земля, вода, надра, біоресурси, гідрометеорологія, вторинні ресурси і відходи, екологічна обстановка);
- фактографічні бази і банки наукових даних, щоб забезпечити фундаментальні наукові дослідження;
- фактографічні бази даних у галузі культури і мистецтва;
- лінгвістичні бази даних, тобто машинні словники різного типу і призначення.

Останнім часом утворилися нові важливі області застосування баз даних, і кожна з них представляє принципово нове середовище, до якого необхідно адаптувати технології СУБД. Ці області отримали на ринку назви **інтелектуального аналізу даних (data mining), сховищ даних (data warehousing), репозитаріїв даних (data repository)**.

Інтелектуальний аналіз даних. Ідея інтелектуального аналізу даних (data mining), тобто добування інформації з величезних масивів даних, накопичених зовсім для інших цілей, викликає сьогодні підвищений ентузіазм. Наприклад, авіакомпанії домагаються оптимального заповнення рейсів за рахунок аналізу накопичених раніше даних про

резервування квитків. Можна навести одну історію про те, як була виявлена несподівана кореляція між покупками пива і покупками серветок у післяобідній період. Власник магазину наблизив один до одного відділи, які торгують пивом і серветками, а між ними помістив ще прилавки з картопляними чіпсами. У результаті збільшилися продажі усіх трьох видів товару.

Із запитамі, характерними для систем інтелектуального аналізу даних, пов'язана низка незвичайних проблем.

1. Вони включають, як правило, агрегацію величезних обсягів даних.
2. Вони мають нерегламентований характер;
3. їх формують особи, відповідальні за прийняття рішень, коли їм необхідно виявити будь-які неочевидні взаємозв'язки.
4. У додатках, пов'язаних, наприклад, з торгівлею цінними паперами, дуже важливий малий час відповіді. Суть проблеми полягає тут у тому, щоб скоротити загальний час, необхідний для написання, налагодження та виконання запиту.
5. Досить часто користувач не в змозі точно сформулювати запит - йому просто потрібно виявити "що-небудь цікаве".

Таким чином, з видобутком даних пов'язані такі дослідницькі напрями:

- методи оптимізації складних запитів, які включають, наприклад, агрегацію та групування;
- методи підтримки "багатовимірних" запитів, що відносяться до даних, організованих у вигляді "куба", в осередках якого знаходяться потрібні дані;
- методи оптимізації використання третинної пам'яті;
- мови запитів дуже високого рівня, а також інтерфейси для підтримки користувачів, які не є експертами, і яким потрібні відповіді на нерегламентовані запити.

Сховища даних.

У сховищі даних накопичуються дані з однієї або більше баз даних. Існує безліч потенційних застосувань, а також підходів до організації сховищ даних. Наприклад, великий магазин може підтримувати сховище даних на основі транзакційних даних про касові операції для цілей видобутку даних. У сховищі даних може зберігатися інформація з багатьох баз даних для використання в надзвичайних ситуаціях. Наприклад, в єдиному сховищі даних підтримуються відомості про цивільну інфраструктуру (дороги, мости, трубопроводи тощо), оскільки, наприклад, після землетрусу навряд чи вдасться отримати цю інформацію з міст, що знаходяться поблизу епіцентру. Ще один приклад - використання сховища даних як "матеріалізованого уявлення" інтегрованої інформації. Альтернативою медіаторних систем, які дають цілісне уявлення даних, витягнутих з безлічі джерел, можуть служити сховища даних, що забезпечують фізичне зберігання інтегрованих даних. На відміну від сховищ, медіатори надають інформацію, розсилаючи запити декільком джерелам, подібно до того, як це відбувається при реалізації уявлень.

Деякі дослідницькі проблеми, що стосуються сховищ даних, збігаються з тими, які характерні для інтеграції даних у цілому, але є і деякі **специфічні проблеми**.

1. Інструменти для створення **насосів даних (data pump)**, тобто модулів, що функціонують за середовищем джерел даних і поставляють у сховище ті зміни, які

істотні з точки зору сховища; при цьому дані мають транслюватися у відповідності з глобальною моделлю і схемою сховища.

2. Методи "**чистки даних**" (**data scrubbing**), які забезпечують узгодження даних, видалення елементів, що відповідають різним уявленням одного й того ж об'єкта (наприклад "Sally Tones" і "SA Tones"), а також видалення неправдоподібних значень.

3. **Засоби для створення і підтримки метасловника**, який інформує користувачів про способи отримання даних.

Репозитарії.

Програми, що відносяться до категорії **репозитаріїв**, характеризуються тим, що вони призначені для зберігання і управління як даними, так і метаданими, тобто інформацією про структуру даних. **Приклади репозитаріїв** - бази даних для підтримки комп'ютерного проектування, включаючи CASE (системи проектування програмного забезпечення), а також системи управління документами. Відмінна риса цих систем - часті зміни метаданих, характерні для будь-якого середовища проектування.

У репозитарії необхідно підтримувати безліч уявлень однієї й тієї ж або схожої інформації. Наприклад, програмний модуль має уявлення у вигляді вихідного коду, об'єктного коду, проміжного коду, готової програми, таблиць використань/визначень, документації. Зв'язки між усіма цими уявленнями повинні відстежуватися репозитарієм так, щоб зміни в одному з них автоматично поширювалися на інші уявлення того ж об'єкта.

Репозитарії повинні підтримувати поняття версій (моментальних знімків елементів даних, що змінюються в часі) і конфігурацій (версійних колекцій версій). Наприклад, різні релізи програмної системи будуть зазвичай формуватися як конфігурації з певних версій файлів вихідного коду.

Сховище повинне підтримувати еволюцію структури інформації та її метаданих таким чином, щоб при додаванні нових властивостей даних або нових зв'язків не була потрібна повна перекомпіляція.