УДК 004.658

Концепція електронного сховища даних

Георгій Асєєв,
доктор технічних наук,
професор, завідувач кафедри
інформаційних
технологій ХДАК

У зв'язку з виникненням величезної кількості оперативної і довідкової інформації погіршується якість управління електронними документами. Виходячи з цих умов, для своєчасного прийняття непередбачених та нестандартних рішень і прискорення доступу до даних, подається концепція оптимально організованого сховища даних, що забезпечує максимально швидкий і комфортний доступ до необхідної інформації

Сьогодні спостерігається небувале піднесення комплексної комп'ютеризації корпоративних організацій і цілих галузей, причому першорядна роль відводиться побудові автоматизованих систем документообігу і діловодства. Однак з часом виявилося, що, впровадивши могутні системи управління електронними документами, багато корпорацій не отримали очікуваного ефекту. У чому причина? Величезна кількість, до 80%, оперативної та довідкової інформації, як і раніше, залишається на паперових носіях і зберігається в бібліотеках або архівах. Ручне опрацювання, що застосовується у таких паперових сховищах, стає вузькою ланкою функціонування багатьох великих корпорацій (http://www.bizcom.ru/rus/bt/1995/nr8/08.htm):

- американські ділові й урядові агентства щодня створюють 900 млн сторінок інформації, в тому числі 76 млн аркушів і 21 млн інших документів;
- у США зберігається близько 1,3 трильйона документів на паперових носіях, але, як відзначають деякі компанії, їхня можливість одержати доступ до інформації, що зберігається на папері, обмежується лише 10%;
- на створення і переміщення інформації на паперових носіях витрачається значна частина валового доходу компаній, у деяких випадках — до 10%.

Вихід із такого становища вбачається у застосуванні технології побудови корпоративного електронного архіву (сховища даних), яка поки ще досить нова, а її реалізація вимагає певної сміливості від замовника і ставить непрості завдання перед системними інтеграторами [1—8].

Прийняті рішення. Одним з основних факторів успіху в управлінні, бізнесі та й у повсякденному житті є швидкість прийняття рішень та їхня якість. Не дивно, що спроби формалізувати або автоматизувати процес прийняття рішень почалися практично відразу з появою обчислювальних машин і продовжуються дотепер. Історія обчислювальної науки повна прикладів блискучих ідей і нездійснених надій. Багатьом спочатку здавалося, що ключ до створення універсального вирішувача проблем лежить у побудові системи логічного висновку на підставі формальної логіки. Згадаємо численні експертні системи, логічні й інші ігрові програми, що демонстрували "майже" задовольняючі результати. Ще б трохи-трохи швидше, трохи-трохи ближче до реальних умов, і, здавалося, проблема буде вирішена раз і назавжди. На жаль, перебороти це "майже" так і не вдалося.

Розвиток комп'ютерних технологій, тим часом, йшов своєю чергою. Маленькі настільні комп'ютери розповсюдилися по усьому світі, витиснувши старі незграбні мейнфрейми. Обчислювальні мережі обплели всю планету. Швидкість передачі інформації досягла одного мегабіта на секунду, а можливість її одержання — практично необмежена. Теза про взаємозв'язок і взаємозалежність усіх видів людської діяльності стала відчутною реальністю. Кількість інформації, що звалюється на користувача комп'ютера, перевищила швидкість її опрацювання. Наслідки прийнятих рішень виявляються не через роки опісля, а буквально наступного дня, а часу на роздуми залишається усе менше і менше.

Як же діяв донедавна державний службовець, підприємець, банкір, одним словом, той, хто повинний приймати відповідальні рішення за службовим обов'язком? Звичайно, до його послуг є база даних (БД), поповнювана в міру надходження нової інформації. Безсумнівно, є кваліфіковані фахівці, які вміють цю інформацію витягати і представляти в зручному для сприйняття вигляді. І, зрозуміло, є експерти, здатні цю інформацію інтерпретувати. Досить було спочатку пояснити програмістам, яка саме інформація потрібна, трохи почекати, поки вона буде готова, потім проконсультуватися з експертами, і рішення готове! Правда, програмісти не відразу зрозуміли, чого від них вимагають, і надали інформацію лише з третьої спроби, та й сам керівник тільки після цього побачив, що зажадав не зовсім йому потрібне, до того ж експерти запросили додаткові відомості, довелося знову звертатися до програмістів і так далі.

Що таке сховище даних? Спробуємо розібратися, що ж відбулося в наведеному випадку. Треба сказати, що з подібними ситуаціями люди зіткнулися давно і почасти навчилися уникати подібних колізій. По-перше, стандартизована значна частина інформації, що надається керівникові. Для цього розроблена ціла низка стандартних форм звітності. По-друге, керівникам різних рівнів надається інформація необхідного ступеня деталізації: від запитуваних відомостей, допустимо, про денну реалізацію друкованої продукції для менеджера по продажах до зведених квартальних звітів для вищого керівництва. По-третє, усе це відбувається не постійно, а за фіксованими термінами: наприкінці дня, місяця, кварталу, року.

На жаль, регламентованість схеми обертається втратою гнучкості, коли яке-небудь непередбачене звертання за інформацією перетворюється на проблему. Але ж складність прогнозування сучасного ринку, величезні швидкості розповсюдження інформації роблять особливо актуальним своєчасне прийняття оперативних і нестандартних рішень [9].

Вихід з цієї ситуації викреслився відносно недавно — наприкінці 80-х — початку 90-х років минулого століття. Хід міркувань при цьому був приблизно таким. Безсумнівно, база даних як джерело інформації необхідна. Однак звичайна база даних обслуговує не тільки керівників, які приймають рішення, а й інших користувачів, які вводять та модифікують інформацію, вилучають дані, що втратили актуальність тощо. Неминуче виникає уповільнення доступу до даних через те, що різні транзакції доводиться опрацьовувати послідовно. До того ж, було помічено, що частота

запитів до БД залежить від ступеня деталізації потрібних даних, а саме, чим дані більш агреговані, тим частіше по них звертаються. Звідси напрошується висновок: для прискорення доступу до даних потрібна окрема БД, що працює тільки в режимі читання і зберігає агреговані дані [9].

Далі, оскільки інформація, необхідна при прийнятті рішень, як правило, носить загальний характер (зведення про всі продажі за певний період часу, інформація про всіх споживачів певного регіону тощо), засоби вибірки, використовувані в традиційних системах управління базами даних (СУБД) й орієнтовані на вибірку індивідуальних записів, виявляються неадекватними. Крім цього, зрозуміло, що наша БД зберігає надлишкову інформацію. У випадку з продажами, наприклад, сума продажів деякого товару за квартал представляє суму сум продажів цього товару за окремі місяці. Отже, нормалізація стосовно нашої бази не має сенсу. А це приводить до висновку, що й організована наша база повинна бути інакше, ніж традиційні СУБД, або, точніше, вона не реляційна.

Щоб сформулювати останню вимогу до нашої БД, нам доведеться докладніше розглянути ситуацію з продажами, про яку вже йшлося. Що цікавить керівника або менеджера торговельної компанії? Ймовірно, не тільки обсяг продажів за квартал (місяць, рік тощо), а й розподіл продажів по окремих товарах і/або групах товарів. Це означає, що в базі потрібно зберігати агрегати даних, підсумовані не тільки за часом, а й за окремими товарами/групами товарів. Аналогічне судження можна привести і для регіонів, фірм-постачальників та багатьох інших аспектів або вимірів продажів. Прийнято представляти таку базу у вигляді багатомірного куба, або гіперкуба. Фактично, це узагальнення звичайної таблиці, у формі якої подаються звіти.

Підсумуємо сказане. Електронний архів (ЕА), або Сховище даних (СД), або Data Warehouse (DW) — це база даних, що зберігає дані, агреговані по багатьох вимірах. Поповнення СД відбувається на періодичній основі. При цьому на базі попередніх агрегатних даних автоматично формуються нові. Доступ до СД організований особливим чином на основі моделі багатомірного куба.

Трохи історії. Концепція DW була запропонована в останнє десятиріччя минулого століття Б. Інмоном і стала однією з домінуючих у розробленні інформаційних технологій опрацювання даних останнього десятиріччя минулого століття. На наш погляд, поява цієї концепції була наслідком неявного усвідомлення того факту, що існує два основних, функціонально різних, класи систем опрацювання інформації.

Перший клас систем базується на розробленні поточного потоку транзакцій і представляє знімок інформації, що охоплює поточний або невеликий часовий період; другий — на збиранні та підготовці великого за обсягом і часовим періодом (від п'яти років) масиву значимої інформації, потрібної для проведення аналізу даних. Розвиток концепції DW дав змогу провести межі між цими двома типами систем. Українською мовою термін Data Warehouse перекладається подвійно: як сховище даних і як інформаційне сховище. Однак термін "Information warehouse" був уведений корпорацією IBM на початку 80-х років XX ст. і, за твердженням її фахівців, означає щось більше, ніж DW за Б. Інмоном. Тому було б доцільно користуватися вже звичним терміном "сховище даних", хоча він трохи гірше передає суть концепції. Термінологія, використовувана нині у рамках концепції DW, наведена в глосарії [10].

Отже, сховище даних — це не автоматизована система прийняття рішень, не експертна система, не система логічного висновку, а "усього лише" оптимально організована база даних, що забезпечує максимально швидкий і комфортний доступ до інформації, необхідної для прийняття рішень.

Відповідно до класичного визначення Б. Інмона, DW є предметно орієнтований, інтегрований, незмінний, підтримуючий хронологію набір даних, призначений для підтримки прийняття рішень. Слід зазначити, що в цьому визначенні поєднані дві різні функції: а) збирання, організація і підготовка даних для аналізу у вигляді постійно нарощуваної бази даних; б) власне аналіз як елемент прийняття рішень. Прийняття рішень як сферу застосування DW істотно звужує визначення. Якщо у ньому залишити лише аналіз (як елемент наукових, технологічних і екологічних систем), коло використання цієї концепції може бути значно розширене.

Дуже важливий основний принцип дії DW: раз занесені в DW дані потім багаторазово витягаються з неї і використовуються для аналізу. Звідси випливає одна з основних переваг використання DW у роботі підприємства — контроль за критично важливою інформацією, отриманою з різних джерел, як за виробничим ресурсом.

Відзначимо, що найуразливішим місцем використання DW на підприємстві, з погляду бізнес-процесів, є коректність його даних, отриманих з різних джерел. Дані перед завантаженням у DW мають бути або "очищені від шуму", або оброблені методами нечіткої логіки, що допускає наявність суперечливих фактів. Наприклад, дані про підприємство-партнера можна отримати від різних експертів, чиї оцінки інколи бувають діаметрально протилежними.

Насамперед, треба звернути увагу на те, що йдеться не про традиційну автоматизацію каталогів паперових бібліотек, а про побудову інтегрованої системи масштабу галузі або корпорації, чим забезпечується ефективний доступ і збереження величезних обсягів документів в електронному вигляді. Потреба в такій системі з'явилася досить давно і час від часу "підхльостувалася" зрослим інтересом до відомчих і державних архівів, що містять унікальні запаси історичної та довідкової інформації. Справа в тому, що архіви, де працюють "дідівськими" методами з паперовими бібліотечними каталогами, уже перестали забезпечувати потрібну оперативність, повноту і вірогідність виконання запитів до фондів документів, які мають тенденцію до непомірного розростання. Більше того, паперові цінності, як відомо, з часом стають непридатними і безповоротно зникають. Величезний потік документів та інформаційних матеріалів, наявних в обігу всередині великих корпоративних структур, додає новий імпульс побудові архівів електронних документів. І тут справа стосується вже не тільки компактного, безпечного збереження і швидкого пошуку документів, а й питань оперативного аналізу, мета якого — прогнозування ринкових колізій і виявлення закономірностей.

Усе це зумовило актуальність створення нової інформаційної технології побудови корпоративного електронного архіву, здатного ефективно опрацьовувати масиви даних обсягом у десятки терабайтів. Причому, технологія має включати як засоби створення/наповнення супербанку даних, так і засоби забезпечення його належного функціонування і розвитку. Однак, якщо потреба в такій технології назріла вже багато років тому, то технічна можливість її реалізації з'явилася відносно недавно, як наслідок комбінації таких факторів:

- з'явилися дешеві носії — бібліотеки компакт- і магнітооптичних дисків;
- різко знизився показник вартість/продуктивність для високошвидкісних обчислювальних систем, мереж і пристроїв;
- одержали розвиток апаратно-програмні системи, що реалізують рівнобіжне опрацювання запитів;
- підвищився рівень інтерфейсу роботи із СУБД;
- з'явилися нові інформаційні технології індексування надвеликих масивів даних;
- розроблені й розвиваються вітчизняні технології і програмні продукти розпізнавання й аналізу російськомовних (україномовних) текстів;
- намітився напрям упровадження засобів штучного інтелекту, що уможливлюють моделювання й аналізування великих масивів інформації.

Концепція сховища даних. Кожен, хто стикався з проектуванням бази даних, знає, наскільки це складний і дорогий процес, до того ж він може призвести до багатьох помилок. Дійсно, побудувати нормалізовану БД реального підприємства, що містить сотні або тисячі таблиць, — надзвичайно складне завдання навіть при наявності гарних CASE-засобів. Здавалося б, у випадку СД, коли йдеться про п'ять-десять вимірів, особливих труднощів виникати не повинно. На жаль, ускладнення, хоча й іншого характеру, зберігаються. Повернемося знову до нашого прикладу торговельної компанії.

Згадаємо, що основна мета СД — підтримка процесу прийняття рішень. Насамперед, про які рішення мовиться? Допустимо, потрібно з'ясувати, чи має сенс знизити ціну на якесь друковане видання при торгівлі вроздріб у цьому регіоні. Щоб відповісти на це запитання, необхідно провести повноцінне маркетингове дослідження. Можливо, потрібно вивчити динаміку реалізації за часом, з'ясувати її динаміку в компаній-конкурентів, врахувати витрати на рекламу й оцінити її дієвість щодо саме цього видання у цьому регіоні. Необхідно оцінити його запаси на складах і вартість збереження, час доставляння товару від постачальника і так далі. Досвідчений маркетолог легко розширить цей перелік. Але за кожним з перерахованих пунктів стоїть цілком певний запит до СД. Невже, перш ніж приступити до проектування СД, потрібно передбачити всі можливі запити до бази даних? Як це не сумно звучить, правильною відповіддю буде "так". Більше того, нездатність СД ефективно підтримати якесь відповідальне рішення може перекреслити всі його переваги.

Ясно, що неможливо скласти заздалегідь перелік усіх мислимих рішень, які доведеться приймати в кожній конкретній компанії. Саме тому закінченої структури електронного архіву не існує, та й не може існувати. Кожне СД — результат об'єднаних зусиль консультантів, майбутніх користувачів і розроблювачів. Обов'язки всіх учасників у проекті заздалегідь відомі. Відповідно до методології розроблення електронних архівів фірми Price Waterhouse, проект загалом організується в такий спосіб [3—8].

Консультанти, тісно взаємодіючи з користувачами, з'ясовують коло бізнес-понять, прийнятих на фірмі, вивчають бізнес-процеси і потоки даних. Саме їм належить вирішальне слово в тому, що стосується набору рішень, прийнятих за допомогою СД. Робота консультантів, як стверджують американські експерти з СД, на 10 % визначається технологією і на 90 % — досвідом, причому важливий досвід саме в конкретній галузі діяльності, оскільки консультанти повинні говорити тією ж мовою, що і користувачі.

Майбутні користувачі надають усю необхідну консультантам інформацію. Це зовсім не таке просте завдання, як може здатися. Справа в тому, що поки СД немає, і рішення приймаються "вручну", ніхто не може точно визначити потребу в тих або інших рішеннях. Навіть керівництво фірми досить часто не в змозі точно сказати, які рішення доведеться приймати в майбутньому, оскільки, через обставини, змушене безупинно вирішувати поточні питання.

Розроблювачі одержують від консультантів технічний опис СД і приступають до його реалізації (відзначимо, що консультанти мають уміти сформулювати вимоги до СД мовою, зрозумілою для розроблювачів). У ході реалізації ті неминуче зіштовхуються з проблемою вибору програмних і апаратних засобів (кількість яких безупинно зростає) і знову звертаються по допомогу до консультантів. Останні або самі вирішують питання, або знову звертаються до користувача, і процес повторюється.

Таким чином, процес створення СД по своїй природі циклічний. Ясно також, що варто викинути з ланцюга консультант→користувач→розроблювач хоча б одну ланку, і процес неминуче зайде в безвихідь, тому СД може бути реалізоване тільки у взаємодії різних невзаємозамінних фахівців, об'єднаних загальною метою.

До проектування сховищ даних звичайно ставляться такі вимоги [11, 12]:

- структура даних сховища має бути зрозуміла користувачам;
- повнота і ймовірність збережених даних;
- потрібно виділити статичні дані, що регулярно модифікуються: щодня, щотижня, щокварталу;
- підтримка внутрішньої несуперечності даних;
- слід спростити вимоги до запитів з метою вилучення тих, що могли б потребувати множинних тверджень SQL у традиційних реляційних СУБД;
- підтримка високої швидкості одержання даних зі сховища;
- потрібно забезпечити підтримку складних запитів SQL, що вимагають послідовного опрацювання тисяч або мільйонів записів;
- можливість одержання і порівняння так званих зрізів даних (slice and dice);
- наявність зручних утиліт перегляду даних у сховищі;
- підтримка якісного процесу поповнення даних.

Ці вимоги істотно відрізняють структуру реляційних СУБД і сховищ даних. Нормалізація даних у реляційних СУБД спричиняє створення безлічі пов'язаних між собою таблиць. У результаті, виконання складних запитів неминуче приводить до об'єднання багатьох таблиць, що істотно збільшує час чекання відгуку. Проектування сховища даних має на увазі створення денормалізованої структури даних (допускається надмірність даних і можливість виникнення аномалій при маніпулюванні даними), орієнтованої насамперед на високу продуктивність при виконанні аналітичних запитів. Нормалізація робить модель сховища занадто громіздкою, ускладняє її розуміння і знижує ефективність виконання запиту.

У наступній публікації обговоримо стандарти моделей для розмірного моделювання електронних сховищ даних.

Список використаної літератури

1. *Фатеев Ф.* Архив технической документации предприятия / Ф. Фатеев, Н. Ширяев // Электрон. офис. — 1998. — № 2.
2. *Марков А.* Концепция построения электронного архива / А. Марков // Открытые системы. — 1997. — № 1. — С. 54—58.
3. *Hagen P.* Smart Personalization, The Forrester Report, Forrester Research / P. Hagen // Cambridge. Mass. — 1999. — July.
4. *Comm. ACM,* Special Issue on Recommender Systems. — 1997. — Vol. 40, 3.
5. *Pazzani M.* A Framework for Collaborative, Content-Based and Demographic Filtering [text] / M. Pazzani // Artificial Intelligence Review. — 1999. — Dec. — P. 393—408.
6. *Adomavicius G.* Expert-Driven Validation of Rule-Based User Models in Personalization Applications / G. Adomavicius, A. Tuzhilin // J. Data Mining and Knowledge Discovery. — 2001. — Jan. — P. 33—58.
7. *Fast* Discovery of Association Rules / R. Agrawal etc. // Advances in Knowledge Discovery and Data Mining. — Menlo Park, Calif : AAAI Press, 1996. — Chap. 12.
8. *Srikant R.* Mining Association Rules with Item Constraints / R. Srikant, Q. Vu, R. Agrawal // Proc. Third Int'l Conf. Knowledge Discovery and Data Mining. — Menlo Park, Calif.: AAAI Press, 1997.
9. *Каменнова М.* Управление электронными документами: технологии и решения / М. Каменнова // Открытые системы. — 1995. — № 4.
10. *Асеев Г. Г.* Электронный документооборот : учебник / Г. Г. Асеев. — К. : Кондор, 2007.
11. *Kimball R.* The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data / Ralph Kimball. — Warehouses : John Wiley & Sons, 1996.
12. *Kimball R.* The Data Webhouse Toolkit: Building the Web-Enabled Data / Ralph Kimbal. — Warehouse : John Wiley & Sons, 2000.