УДК 004.658:004.72

Архітектура корпоративного сховища даних

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

Розглянуті компоненти корпоративного сховища даних: підсистеми введення, зберігання, індексування, відображення інформації, аналізу, управління потоками, адміністрування науково-технічного супроводу і специфічна особливість сховища даних — надання повнотекстового пошуку.

Ключові слова: *інформаційне сховище даних, повнотекстовий пошук, сервери уведення й опрацювання інформації.*

В опублікованій раніше роботі [1] були розглянуті основні компоненти інформаційного сховища даних в електронному документообігу. Тепер перейдемо до опису його архітектури. Загальну ідею можна змалювати в такий спосіб. Організовується розгортання високопродуктивної мережі, що включає графічні робочі станції й потужні сервери введення та опрацювання інформації. Для введення документів з паперових носіїв низької якості використовуються промислові сканери потокового введення й відповідні українізовані (або русифіковані) програмні засоби. Система забезпечує ефективне індексування й повнотекстовий пошук неструктурованої інформації великого обсягу. Дані, необхідні для пошуку документів, зберігаються у високопродуктивній відмовостійкій системі пам'яті, а графічні образи документів — у вигляді зображень на носіях, що характеризуються тривалим часом зберігання й дешевизною. Розглянемо основні функції сховища даних (СД): сканування; розпізнавання й коректування помилок; створення та міграція електронних документів й образів; індексування документів; оперативний пошук і відображення документів; їхній аналіз; управління функціонуванням системи.

Для реалізації даних функцій у СД необхідні компоненти введення, зберігання, індексування, пошуку і відображення інформації, аналізу, управління потоками, адміністрування й науково-технічного супроводу. Специфіка впровадження системи електронного архівування полягає в тому, що, насамперед, необхідно ввести в базу даних системи повний обсяг документів. Це надзвичайно тривалий і трудомісткий процес, що потребує максимальної автоматизації — відсторонення оператора від будь-якої участі в процесі введення, розпізнавання, коректування та індексування документів. Із цим пов'язана інша специфічна риса СД — забезпечення повнотекстового пошуку. Побудова і підтримка системи атрибутивного пошуку, характерного для систем управління документообігом, виявляються неприйнятними через тимчасові та цінові обмеження.

Розглянемо такий приклад. Припустимо, паперовий архів нараховує 50 млн документів. На перевірку одного розпізнаного документа, класифікацію-рубрикацію, введення атрибутів оператор, у середньому, витрачає дві хвилини. Отже, для введення всіх документів у режимі стандартного робочого тижня потрібно 1112 років. З іншого боку, при автоматичному введенні документів основною вузькою ланкою системи буде продуктивність сканерів і потужність сервера, що виконує розпізнавання й індексування. З урахуванням оптимізації потоків підсистеми введення можна чекати, що аналогічний обсяг буде повністю виконаний за 5—15 років, тобто ще за життя оператора.

Іншою відмінністю й важливою особливістю СД є те, що воно включає як засоби оперативного пошуку інформації (On-line Time Processing — OLTP), так і засоби оперативного аналізу інформації (On-line Analysis Processing — OLAP).

Функціонування OLAP-засобів базується на використанні ідеології багатомірних кубів, тобто кожне значення відповідає декільком вимірам (наприклад, факт відвантаження продукції є приналежністю не тільки типу продукції, а й підрозділу, що її випускає, менеджера з продажів, відповідального за цей тип продукції, тощо). Таким чином, стає можливим одержання інформації "у декількох вимірах", що вкрай важливо у випадку запиту складного виду.

Компоненти корпоративного сховища даних. Тепер, розглянувши основні функції СД, коротко визначимо його ключові програмно-апаратні компоненти. Спочатку подамо перелік основних високопродуктивних апаратних засобів: потокові сканери, що забезпечують надійне введення паперових матеріалів низької якості (40 ст./хв. і більше); масштабовані сервери, що забезпечують паралельне опрацювання запитів; обчислювальна мережа, орієнтована на введення та обробляння графічних документів; RAID-масиви, що забезпечують наднадійний доступ до пошукових даних системи; автоматичні бібліотеки компакт-дисків, flash-пам'ять тощо, що уможливлюють довгострокове зберігання величезних масивів інформації; системи перенесення даних на компакт-диски та резервного копіювання; робочі місця-клієнти, орієнтовані на опрацювання графічної інформації; робочі місця розроблювачів конкретних додатків; системи забезпечення безаварійного живлення; принтери й модеми.

Системні програмні засоби включають: операційне мережне середовище, орієнтоване на мультипотокове обробляння в мережі й сертифіковане з безпеки; системи управління даними (СУД), орієнтовані на опрацювання надвеликих масивів даних, відображення та захист інформації.

І, нарешті, найважливіше — спеціальні програмні засоби: розпізнавання україномовних і російськомовних текстів; розроблення та оптимізації запитів; повнотекстового індексування і пошуку інформації; її аналізу.

Технічна реалізація — проблема вибору. При існуючій різноманітності програмного забезпечення сьогодні відсутні будь-які продукти СУД, що уможливили б забезпечення всіх основних функцій електронного документообігу й архівування для роботи з надвеликими обсягами документів. З іншого боку, більшість компонентів СД, як технічних, так і програмних, є унікальними й налаштованими зразками. Тому, *при проектуванні СД виникають кілька взаємозалежних проблем: оцінка й вибір компо-*

нентів; інтегрування технологій, програмних продуктів і технічних засобів.

Вибір деяких компонентів, наприклад, високопродуктивної мережі, конкретних моделей серверів або RAID-масиву збігається із сучасними технологічними рішеннями побудови традиційних автоматизованих систем збирання, зберігання й опрацювання інформації. Однак, ряд компонентів має унікальну орієнтацію саме на електронне архівування. Тут ми будемо розглядати тільки проблему вибору специфічних для СД компонентів.

У найближчі роки сховища даних і на веб складуть основу планів більшості організацій. Звертання до технологій Internet/intranet, а точніше — до веб-технологій, обумовлене, насамперед, низькою ціною управління та впровадження додатків, легкістю в застосуванні, а також можливістю створення простого у застосуванні користувальницького інтерфейсу на основі веб-браузера. Концепція "стратегічного оточення" дає розроблювачам змогу будувати ODBC (Open Database Connectivity)-подібні інтерфейси для значної частини платформ, ніби "занурюючи" існуючі додатки в середовище Інтернет і створюючи тим самим зручну графічну оболонку для користувачів. Вони одержують доступ до інформації й додатків за допомогою будь-якого веб-браузера практично однаково, незалежно від того, на якій платформі вони працюють [2].

Крім того, завдяки високому ступеню інтеграції сховищ даних з існуючими інформаційними системами, організації та фірми можуть додавати сховища даних на веб в існуючі мережі, де на внутрішніх серверах розміщені бази даних, на серверах проміжного шару — прикладні системи, а персональні або мережні комп'ютери користувачів містять засоби доступу.

Для того, щоб вирішити проблеми, пов'язані з доступом до сховища документів на веб, потрібне сполучення "товстих" та "тонких" типів клієнтських додатків.

Програмне забезпечення клієнта включає ПЗ сервера додатків і надтонкого клієнта. Сервер додатків, створений на базі Active Server Page-технології, забезпечує функціонування надтонкого клієнта — опрацьовує запити, спрямовує їх до сервера БД і формує статичну html-сторінку, що відображає отриману вибірку даних.

Надтонкий клієнт є інтерфейсом у вигляді інформаційно-пошукової системи (ІПС), яка працює під управлінням стандартних веб-браузерів, що функціонують у різних операційних системах (ОС), а також на різних апаратних платформах (IBM, MAC, HP, DEC, Sun, Seguent тощо) у єдиній інформаційній системі.

Таким чином, пошук документів побудований на основі принципу QBE (Query By Example), тобто для того, щоб знайти документ, користувач заповнює структуровані поля екранної форми ІПС, уводячи його ключові слова або назву. У відповідь на запит система видає список документів у форматах HTML, RTF або PDF та інших, відповідно до введених даних і прав доступу користувача. Адаптація надтонких клієнтських місць до вимог різних груп користувачів відбувається за рахунок контролю за доступом до веб-сервера й сервера БД.

Адміністрування містить: застосування масштабованих серверів (секціонування опрацювання й розподіл завантаження сховища; конфігурування компонентів, розподіл розширень plug-in серед користувачів при супроводі тонких клієнтів, резервне копіювання, багаторівневий авторизований доступ) і підтримку чистоти даних (верифікація даних, управління документами із закінченим строком зберігання).

Електронне сховище документів забезпечує: сканування; створення електронних документів; індексування документів; їхні оперативний пошук і відображення; управління функціонуванням системи; функціонування різнотипних комп'ютерів, що працюють під управлінням різних ОС; швидкісне опрацювання інформації в мережному середовищі; підготовку гіпертекстових баз даних; оперативне внесення змін у бази даних; доступ до інформації різних категорій користувачів за рахунок WWW-інтерфейсу; контроль наповнення, цілісність даних й багаторівневого доступу до інформації; створення вітрин даних; віртуальне складування даних; аналіз документів з можливістю подальшого прогнозування (OLAP); швидку адаптацію до зміни програмних і технічних засобів; сумісність використання різних апаратних платформ; високу надійність і безперебійну експлуатацію за рахунок резервного копіювання й кластерної архітектури.

Технології індексування та пошуку. Ядром корпоративного електронного архіву по праву можна вважати технології індексування й пошуку. Сьогодні намітилося кілька напрямів побудови електронних архівів, залежно від використовуваних у них методів пошуку.

Перший напрям, іменований також корпоративним електронним архівом, належить до класу традиційних інформаційно-пошукових систем, заснованих на атрибутивному пошуку структурованих даних. Як приклади можна навести системи побудови електронних архівів на базі програмних продуктів типу DOCS Open (PC DOCS), XDOC (Rank Xerox), SoftSolution (Novell), PaperWise (PaperWise) та ін. Точніше кажучи, цей напрям не є технологією корпоративного електронного архівування як такого. Проведені розрахунки по введенню повного масиву документів показують, що навіть невелика затримка на кілька секунд при введенні документів виливається в додаткові кілька років, необхідні для введення повного обсягу документів. Візуальний контроль і напівавтоматизоване заповнення атрибутів практично не реалізовані в основній масі документів великого архіву.

Альтернативний напрям електронного архівування базується на принципі повнотекстового індексування неструктурованих даних і включає два види індексування: контекстно-незалежне індексування, яке не залежить від природної мови через бінарну або словникову індексацію та контекстно-залежне індексування, що уможливлює оптимізацію індексації й пошуку з урахуванням специфіки морфології й семантики природної мови.

Відомо кілька методів контекстно-незалежного індексування [3, 4]. Найпоширеніший — індексація на базі інвертованої матриці, де словам або нормалізованим словоформам ставляться у відповідність адреси документів. Тут, зазвичай, використовується стоп-словник слів, що не індексуються, і словник синонімів. Інший метод — бінарне індексування, наприклад, на базі теорії нейронних мереж. При використанні теорії розпізнавання образів цей метод уможливлює нечіткий пошук подібних, з погляду бінарних одиниць, слів або, інакше, "пошук з помилками". Нечіткий пошук надає величезні можливості для виявлення слів, що містять перекручування або помилки. Наприклад: текст після розпізнавання, перекладені російською мовою назви фірм або іноземні прізвища. Однак при нечіткому пошуку користувач стикається із проблемою відсіювання шуму — документів, де зустрілися подібні по синтаксису, але не за змістом слова. У цілому, технологія повнотекстового електронного архіву представлена двома магістральними напрямами: технологія електронного архівування, де використані можливості сучасних промислових СУБД, і технологія, заснована на спеціалізованих системах індексування й пошуку.

Перший підхід базується на використанні засобів відомих SQL-СУБД, типу: Oracle, Informix, Sybase й інших, здатних підтримувати надвеликі бази даних. Звичайно, ці

СУБД не мають засобу повнотекстової індексації типу інвертованої матриці, через це обсяг індексу може становити 30—35% від загального обсягу бази. Процентний розкид залежить від ступеня нормалізації індексованих слів тексту — приведення до початкової форми іменників, прикметників і дієслів. До переваг цього методу можна віднести ось що: крім функцій індексування, у СУБД присутня безліч корисних і необхідних функціональних, сервісних і технологічних функцій підтримки якісної архівної діяльності й документообігу. Ці засоби істотно спрощують завдання інтегрування засобів і функцій, захисту інформації тощо; СУБД досить поширені, що виключає необхідність освоєння нових продуктів; ці засоби пройшли багаторічну апробацію в рамках додатків СУБД, перевірені на практиці й, безсумнівно, будуть підтримуватися й розвиватися ще досить довго.

До основних недоліків варто віднести те, що СУБД, особливо реляційного типу, з самого початку не орієнтовані на інтенсивне обробляння надвеликого обсягу інформації. Тому ряд функцій з повнотекстового пошуку й побудови запитів, швидкості пошуку реалізується менш ефективно та якісно, ніж у спеціалізованих пакетах. Наприклад, більшість СУБД поки що не мають засобів підтримки нечіткого пошуку. У результаті, необхідний додатковий етап верифікації введеного тексту для виправлення можливих помилок сканування й розпізнавання. Однак нині виникла нова тенденція — випускаються нові модулі або версії програмних продуктів, орієнтованих на опрацювання надвеликих обсягів традиційних даних і даних мультимедіа. Прикладами реалізації зазначеного напряму є програмні засоби індексування й пошуку російськомовних текстів, розроблені фірмами LVS і Cognitive Technologies. Сьогодні відповідні засоби працюють у рамках СУБД Oracle й OB2.

Другий підхід, що включає повнотекстове індексування й пошук, базується на використанні додаткових спеціалізованих пакетів повнотекстової індексації, зокрема на базі нейронних мереж. Багато хто з аналітиків вважає, що традиційні системи не придатні для вирішення завдань СД, де потрібні винятково потужні процесори даних, оптимізовані за критерієм швидкості доступу. У таких системах застосована бінарна індексація й реалізується нечіткий пошук. Переваги систем: вони мають якісніші можливості з індексування, пошуку й аналізу, зокрема, реалізують нечіткий пошук, що дає змогу відмовитися від проблеми виправлення помилок після розпізнавання; характеризуються винятково високою швидкістю доступу; обсяг індексу не перевищує 30% обсягу текстових даних; крім неструктурованих даних, звичайно, підтримують різні мультимедіадані.

Які тут виникають супутні проблеми? *По-перше*, результати нечіткого пошуку прямо залежать від якості заданого запиту, і перед користувачами постає проблема шуму — одержання нерелевантних документів. *По-друге*, названі системи розраховані на потужні паралельні обчислювальні системи й поки не дуже поширені на платформі Intel. Але головний недолік полягає в тому, що це системи винятково індексування й пошуку — в них істотно обмежені функції управління документами. На розроблювачів покладають непрості завдання — створення власних технологічних і сервісних функцій, інтегрування технологій і програмно-апаратних засобів тощо.

Піонером представленого напряму, який на європейському ринку досить новий, є американська компанія Excalibur Technologies, що має сьогодні представництва по всьому світу. Фірма пропонує два програмних продукти: Excalibur EFS й RetrievalWare. Перший — це "коробковий" продукт, він дає можливість виконувати повнотекстову індексацію й пошук інформації, що збережена у файлових системах або СУБД. Другий продукт — потужний інструментальний засіб створення систем повнотекстового пошуку на базі теорії нейронних мереж. RetrievalWare включає два компоненти або програмні бібліотеки: текстовий сервер і сервер зображень. Бібліотеки можуть бути підключені до базового ядра інформаційної системи, наприклад, до СУБД Sybase.

Сервер зображень може ефективно використовуватися для пошуку тривимірних зображень, сигналів, фотографій, відбитків пальців, усного мовлення тощо. Текстовий сервер тут істотно перспективніший, ніж Excalibur EFS — він включає не просто механізм пошуку неструктурованої інформації, а й семантичний аналізатор. Цей продукт уможливлює створення семантичної мережі між поняттями мови, а отже, — істотне розширення можливості пошуку. Наприклад, англійський варіант семантичної мережі включає понад 0,5 млн слів й 1,5 млн зв'язків між ними. Сьогодні є також відповідні розроблення у сфері "русифікації" семантичного сервера. RetrievalWare також включає компонент веб, що дає змогу працювати в середовищі Internet або intranet. Незважаючи на розходження в розглянутих підходах до індексування й пошуку, можлива їхня комбінація при реалізації конкретного електронного архіву.

Пристрої зберігання даних. Поняття "зберігання даних" має цілу низку аспектів: безпеку зберігання даних; швидкий і легкий доступ до них; резервне копіювання даних з таким розрахунком, щоб у випадку збою системи або виникнення якої-небудь надзвичайної ситуації дані залишилися цілими й придатними для відновлення, а функціонування СЕД тривало у звичайному режимі.

Як ми вже відзначали, всі дані в системі можуть перебувати у двох видах: пошуковий образ й образ власне документа. Через високі вимоги до швидкості доступу до пошукового образу документа та його цілісності, його потрібно зберігати у високошвидкісних відмовостійких системах зберігання, наприклад RAID-масивах.

Для зберігання образу документів найпридатнішими носіями можуть бути фазоінверсійні (PD/CD), компакт- (DVD-RW й CD-RW) і WORM-диски. Для автоматизації пошуку інформації, розміщеної на цих дисках, її видобування й роботи власне з дисками використовуються автоматичні бібліотеки або, як їх ще називають, оптичні дискові автомати (JukeBox). Сьогодні відомі бібліотеки, які мають до 60 дисководів і до 3 тис. гнізд для дисків, які вибираються механізованим способом. Автоматичні бібліотеки можуть бути багатофункціональними, наприклад, одночасно підтримувати фазоінверсійні й компакт-диски.

Компакт-диски уможливлюють перезаписування інформації. Більшість технологічних рішень електронного архівування підтримує технологію міграції даних саме на компакт-диски, які стійкі до помилок запису, мають високу швидкість читання та тривалий гарантійний термін зберігання інформації, який може становити понад 100 років. З урахуванням того, що більшість архівних документів, практично, не підлягають модифікації й видаленню, бібліотекам на компакт-дисках може бути віддана перевага. Крім того, вони зручніші в роботі: їхнє автономне читання можна здійснювати на будь-якому ПК, комплектованому DVD або CD-ROM-приводом.

Не викликає сумніву, що вся інформація в системі повинна мати резервні копії. Для графічних образів збереження інформації може бути забезпечене створенням дубльованих компакт-дисків. Для зберігання мінливої пошукової інформації у якості зберігаючих накопичувачів зручніше використовувати системи резервного копіювання на магнітних стрічках. Застосовувані в персональних сис-

темах технології (DC2000/Travan, DC6000, DAT) непридатні через обмеження в обсязі. Можливим варіантом можуть стати DLT-стримери, восьмиміліметрові бібліотеки Exabyte (Mammoth) або спеціалізовані котушкові системи.

Архітектура обчислювальної системи. Не зупиняючись на виборі конкретного сервера, відзначимо лише особливості архітектури обчислювальної системи. Дослідження показують, що для підтримки системи класу СД сьогодні придатні тільки потужні масштабовані RISC-платформи, орієнтовані на паралельні обчислення. Важливим критерієм при виборі перспективного сервера є підтримка 64-розрядності, необхідна при введенні й опрацюванні великих обсягів мультимедійних даних. Сьогодні цю можливість надають компанії DEC, SGI, Oracle й Sybase. У найближчому майбутньому очікується, що до них приєднаються HP й SUN Microsystems.

Деякі рекомендації. Побудова електронного архіву — справа суто індивідуальна. І якщо архіву фото- і кінематографії, скажемо, додатково потрібні функції обробляння відеозображень й аудіосигналів, то архіву МВС — пошук відбитків пальців і фотографій. Кожна організація унікальна й вимагає врахування специфіки роботи, ступеня її автоматизації, наявного парку технічних засобів, кваліфікації фахівців і, нарешті, платоспроможності.

Які ж інженерно-технічні труднощі впровадження технології електронного архівування? Це вирішення двох великих завдань: наповнення електронного архіву й забезпечення ефективного пошуку. Перше охоплює ряд інженерно-технічних проблем, вирішення яких потребує істотних тимчасових витрат. Цим обумовлена важливість ефективної організації процесу розроблення, що включає оптимальне планування процесів, аналіз і синтез напрацьованих технологій, створення системи управління якістю тощо. Негнучкість економічної діяльності в більшості держструктур призводить до обмеження на поетапність і нарощування державних електронних архівів. Це підвищує вимоги до системного й детального проектування, створення дослідного зразка, організації випробувань і тестування. При цьому варто пам'ятати, що систему не можна вважати закінченою, поки не буде введений основний нагромаджений обсяг документів.

Незважаючи на те, що масове введення є найважливішим і найбільш трудомістким завданням СД, воно не самоціль. Забезпечення ефективного доступу до наявних даних із застосуванням інтелектуальних засобів — цільове завдання СД. На цьому етапі найактуальнішими є питання оптимізації запитів за критерієм швидкості виконання.

Важливим фактором є обсяг дискового простору**.** Сховища даних вимагають великого обсягу дискового простору. При його оцінці не варто брати до уваги тільки сучасні виробничі системи. Треба пам'ятати, що сформована система буде зберігати нагромаджувані дані. Будь-яка організація прагне зберігати важливі дані як мінімум за рік, а якщо на майбутнє запланований аналіз тенденцій, знадобляться дані за 10—15 років. Крім того, звіти й аналізи звичайно не обходяться одним індексом. Отже, необхідна правильна оцінка дискового простору. Немає нічого незвичайного в сховищі даних, вимірюваному терабайтами, а в деяких великих організаціях рахунок йде навіть на петабайти[1].

Очевидно, що розглянута технологія досить дорого коштує й "по плечу" тільки великим організаціям. Тому, з огляду на певні витрати на створення системи, наведемо перелік основних переваг електронного архіву. *По-перше,* підвищення повноти й оперативності відпрацьовування запитів до архіву. Особливо це ефективно при виконанні нестандартного нерегламентованого запиту. Відповідь, якої раніше чекали місяцями, причому без будь-якої впевненості, що вона виявиться позитивною, тепер можна одержати за секунди й у зовсім іншій якості. *По-друге*, компактність і надійність зберігання. Можна відмовитися від дорогих сховищ документів, скорочувати витрати й займані площі. Звуження кола допущених осіб, контроль й облік доступу до системи дадуть змогу підвищити не тільки збереження, а й безпеку конфіденційної інформації. Зберігання документів в електронному вигляді приводить до того, що ряд архівних функцій: ксерокопіювання, мікрофільмування, ведення автоматизованих картотек будуть зведені до мінімуму. *По-третє*, створюється можливість проведення оперативного аналізу наявної інформації, що підвищить обґрунтованість рішень, прийнятих вищою й середньою ланками керівників, які поки що покладаються тільки на свій досвід та інтуїцію.

Середовище Microsoft Data Warehousing Framework. Процеси створення, підтримання і використання сховищ даних традиційно вимагали значних витрат, що насамперед викликано високою вартістю спеціалізованих інструментів, які пропонує ринок. Ці інструменти практично не інтегрувалися між собою, оскільки базувались не на відкритих і стандартних, а на приватних і закритих протоколах, інтерфейсах тощо. Ускладнення при створенні та висока вартість робили практично неможливим використання сховищ даних у невеликих і середніх фірмах, але ж необхідність аналізу даних існує у будь-якій, незалежно від її масштабу.

Корпорація Microsoft вжила певних заходів, пов'язаних зі сховищами даних і необхідністю створення інструментального й технологічного середовища, що дало б змогу мінімізувати витрати на створення сховищ даних і зробило б цей процес доступним для масового користувача. Це призвело до створення Microsoft Data Warehousing Framework (http://www.mf.grsu.by/other/lib/olap/bd_wh/doc18.htm) — специфікації середовища створення та використання сховищ даних. Ця специфікація визначає розвиток не тільки нової лінії продуктів Microsoft (наприклад, Microsoft SQL Server), а й технологій, що забезпечують інтеграцію продуктів різних виробників. Відкритість середовища Microsoft Data Warehousing Framework забезпечила його підтримку багатьма виробниками ПЗ, що, у свою чергу, дає можливість кінцевим користувачам обирати найкращі інструменти для реалізації своїх рішень. Мета Microsoft Data Warehousing Framework — спростити розроблення, впровадження й адміністрування рішень із використанням можливостей сховищ даних. Ця специфікація покликана забезпечити: відкриту архітектуру, що легко інтегрується й розширюється третіми фірмами; експорт й імпорт гетерогенних даних одночасно з їхньою перевіркою, очищенням і можливим веденням історії нагромадження; доступ до поділюваних метаданих у ході процесів розроблювання сховища, видобування й трансформації даних, управління сервером й аналізом даних кінцевими користувачами; вбудовані служби планування завдань, управління дисковою пам'яттю, моніторинг продуктивності, оповіщення й реакції на події.

Основними компонентами Microsoft Data Warehousing Framework є: OLE DB — стандарт обміну даними; метадані; засоби зберігання даних; засоби OLAP-аналізу; засоби перенесення та трансформації даних; засоби подання й аналізу даних; засоби адміністрування.

OLE DB — стандарт обміну даними. Побудова сховищ даних вимагає, з одного боку, взаємодії з різними оперативними БД для видобування даних, а з іншого — обміну даними та метаданими між різними компонентами. І те, й інше завдання вирішується вкрай складно — за відсутності єдиного інтерфейсу для доступу до різнорідних даних. Але такий інтерфейс існує — це OLE DB, який пов-

[1] 1 Пбайт = 1024 Тбайт.

ністю заснований на відкритій моделі COM (Component Object Model) і є набором інтерфейсів, які можуть бути використані, наприклад, у додатках на Visual C++. Для спрощення використання OLE DB створений набір ActiveX-компонентів — Active Data Objects (ADO). Ці компоненти можуть викликатися з додатків на Visual Basic, Access, Excel, вбудовуватися в активні веб-сторінки тощо. Практично всі компоненти використають OLE DB для доступу до даних. OLE DB забезпечує доступ не тільки до реляційних даних, а й до таких ресурсів, як поштові повідомлення, файлові каталоги, повнотекстові індекси тощо.

Метадані. Одне з найважливіших завдань при побудові сховища даних — інтеграція різних компонентів й інструментів, використовуваних для проектування, зберігання даних, перенесення й трансформації, а також аналізу даних. Ключовим моментом при такій інтеграції є можливість використання поділюваних метаданих (тобто дані про дані). Центральним компонентом Data Warehousing Framework є сховище метаданих (репозитарій) — Microsoft Repository, що постачається як один із компонентів Microsoft SQL Server. Microsoft Repository — це база даних, що зберігає описову інформацію про компоненти програмного забезпечення й про їхній зв'язок. Microsoft Repository складається з набору відкритих інформаційних моделей (Open Information Model, OIM), а також набору опублікованих COM-інтерфейсів. Відкриті інформаційні моделі — це об'єктні моделі певного типу інформації, при цьому вони досить гнучкі, забезпечують підтримку нових типів інформації. Корпорація Microsoft, спираючись на співпрацю із представниками галузі, вже розробила моделі OIM для схеми баз даних (Database Schema), перетворення даних (Data Transformations) і OLAP. Наступні моделі будуть підтримувати реплікацію, планування завдань, семантичні моделі й інформаційний довідник, що призначений для забезпечення метаданими кінцевого користувача. Коаліція метаданих (The Metadata Coalition), галузевий консорціум 53 виробників, заявила про підтримку Microsoft Repository. Відкриті інформаційні моделі одержали широку підтримку в незалежних розроблювачів ПЗ.

Засоби зберігання даних. Серцевиною сховища даних є, безумовно, СУБД, що забезпечує надійне й продуктивне зберігання й опрацювання даних. Як правило, дані з оперативних БД переміщуються в реляційне сховище, де вони стають доступними для аналізу; надалі, при використанні OLAP-засобів, можуть бути переміщені в багатомірну СУБД або будуть вибиратися процесором багатомірних запитів прямо з реляційних таблиць. Microsoft SQL Server забезпечує як реляційний, так і багатомірний вид зберігання. Докладну інформацію про Microsoft SQL Server можна знайти в розділі "Microsoft SQL Server". Нижче коротко перераховані його основні характеристики: спочатку можливості реляційної СУБД, а потім — багатомірної.

Microsoft SQL Server притаманна низка властивостей, що роблять його чудовою платформою для побудови сховищ даних: підтримка баз даних, розмір яких обчислюється терабайтами; поліпшене обробляння запитів, що забезпечує оптимізацію й ефективне виконання складних запитів, типових для сховищ даних, зокрема, запитів за схемою типу "зірка"; засоби паралельного виконання складних запитів; ефективні засоби настроювання продуктивності, завантаження даних і побудови індексів; розподілені запити, що уможливлюють вибирання зв'язаних даних із різних OLE DB-джерел; надійні й ефективні засоби тиражування даних, незамінні при підтримці декількох зв'язаних сховищ або кіосків даних; масштабованість як "нагору" — вбік сучасних наймогутніших апаратних платформ для підтримання дуже великих баз даних, так і "вниз" — убік серверів невеликих робочих груп і навіть настільних і мобільних комп'ютерів (при цьому забезпечується повна сумісність).

Крім того, засоби тиражування, як і раніше, залишаються одним із механізмів переміщення даних з оперативної БД у сховище. Нижче розглядається ряд механізмів, що входять до складу SQL Server.

Засоби OLAP-аналізу. OLAP усе більш популярна з-поміж інших аналогічних технологій, яка може докорінно вдосконалити аналіз даних. Microsoft SQL Server OLAP Services — це новий, повнофункціональний OLAP-сервер, що постачається в складі SQL Server. OLAP Services містить власне сервер, доступний за протоколом OLE DB for OLAP, а також клієнтський компонент, що є постачальником протоколу OLE DB for OLAP і забезпечує ефективне кешування й можливість локального збереження багатомірних вибірок для їхнього подальшого аналізу без підключення до OLAP-сервера.

Традиційно OLAP характеризувався дорогим інструментарієм і складним процесом реалізації. Включення OLAP-функціональності в Microsoft SQL Server зробить багатомірний аналіз дешевшим, що особливо важливо для невеликих і середніх організацій. Крім того, усі підрозділи організацій також зможуть повною мірою скористатися новими можливостями аналізу — від складання звітності до просунутих систем прийняття рішень.

Засоби перенесення й трансформації даних. Організація видобування даних з оперативних БД, їхнє очищення, інтеграція й розміщення в сховищі може вимагати значних зусиль і витрат, якщо не користуватися вбудованою в Microsoft SQL Server службою — Data Transformation Services (DTS). DTS має такі властивості: вона сповна використовує OLE DB для доступу як до джерела, так і до приймача даних. Завдяки цьому, DTS може витягати й перетворювати дані практично з будь-яких джерел (і, відповідно, поміщати їх у будь-які приймачі даних); для перенесення й трансформації даних використовується розширюваний набір ActiveX-об'єктів, якими легко керувати за допомогою мови сценаріїв, наприклад VBScript або JavaScript. Таким чином, є практично необмежені можливості управління перенесенням і перетворенням даних; DTS здатна інтегруватися з Microsoft Repository для використання метаданих про джерело, приймач і схему перетворення даних; завдання щодо перенесення та перетворення даних (а це значна кількість послідовних операцій), оформляються у вигляді пакетів (DTS Package), які можна зберігати в сховищі метаданих (Repository), у базі SQL Server або у файлі. Пакети можуть потім автоматично виконуватися за розкладом за допомогою сервісу SQL Server Agent.

Засоби подання й аналізу даних. Саме у цій сфері варто очікувати (і вже можна бачити) найбільшу кількість продуктів третіх фірм, хоча й Microsoft пропонує тут не тільки базові технології, а й засоби для кінцевого користувача. До них належать компоненти нового покоління Microsoft Office — Office 2008, насамперед Microsoft Excel. Його популярний засіб аналізу даних PivotTable тепер зможе задіяти всю потужність OLAP-сервера, підключаючись до нього через уже згадуваний клієнтський компонент PivotTable Services.

Засоби адміністрування. Значною складовою витрат на впровадження сховища даних є витрати на поточний супровід й адміністрування сховища. Засоби адміністрування, у тому числі засоби автоматизації виконання адміністративних завдань, передбачені Data Warehousing Framework і включені до складу Microsoft SQL Server, дають змогу

значно скоротити ці витрати. Єдиним середовищем адміністрування різних компонентів є Microsoft Management Console, а засобами управління кожним конкретним компонентом (наприклад, SQL Server або OLAP Services) — так звані snap-in, тобто модуль адміністрування, що використовує єдині засоби користувальницького інтерфейсу. Засіб управління SQL Server — SQL Enterprise Manager містить понад 25 програм-майстрів (Wizards), що допомагають не надто кваліфікованому адміністраторові вирішувати найважливіші завдання, у тому числі створювати й копіювати бази даних, робити настроювання тиражування, імпорт/експорт даних, управляти правами користувачів тощо. Крім того, в SQL Enterprise Manager входять засоби створення й редагування графічних діаграм баз даних, що значно полегшує створення й модифікацію структури сховища. Засоби автоматизації адміністрування дають можливість створювати багатоступінчасті завдання, що складаються як з команд мови Transact-SQL, так і зі сценаріїв на мовах VBScript або JavaScript. При цьому виконання наступних операцій може залежати від результатів виконання попередніх. Ці завдання можуть охоплювати величезну кількість серверів і виконуються за заданим розкладом.

Основні постачальники ПЗ сховищ даних: корпорація "Електронний Архів", Arbor, Business Objests, Carleton, Cognos, Hewlett-Packard, IBM, Information Builders, Informix, Intellidex, Microsoft, MSP, NCR, Oracle, Platinum Technology, Praxis, Prism, Pyramid, Red Brick, SAS Institute, Sequent, Software AG, Sybase, Tandem та ін. Всі ці фірми мають сторінки в Інтернеті, де наводяться докладні відомості про їхні продукти й послуги. Варто окремо відзначити альянс Arbor і Seagate при вбудовуванні OLAP у Crystal Info для СУБД Essbase.

У цій роботі часто згадувалися засоби *OLAP-аналізу*. Саме їм буде присвячена наступна публікація.

Список використаної літератури

1. *Асєєв Г. Г.* Концепція електронного сховища даних / Георгій Асєєв // Вісн. Кн. палати. — 2009. — № 2. — С. 28—30.
2. *Вовчок В. А.* Документальное хранилище на Web / В. А. Вовчок. — Режим доступа : http://www.inftech.webservis.ru/it/conference/scm/2000/session5/. — Загл. с экрана.
3. *Hill W.* Recommending and Evaluating Choices in a Virtual Community of Use / W. Hill [et al.] // Proc. Conf. Human Factors in Computing Systems (CHI95), ACM Press. — New York, 2005. — P. 194—201.
4. *Pazzani M.* A Framework for Collaborative, Content-Based and Demographic Filtering / M. Pazzani // Artificial Intelligence Review. — Dec., 2007. — P. 393—408.

Рассмотрены компоненты корпоративного хранилища данных: подсистемы ввода, хранения, индексирования, отображения информации, анализа, управления потоками, администрирования научно-технического сопровождения и специфическая особенность хранилища данных — предоставление полнотекстового поиска.

Components of enterprise data warehouse: subsystems input, storage, indexing, information display, analysis, flow control, administration, scientific and technical support and specific features of the data warehouse — providing full-text search are considered.