УДК 681.3

Г. І. МІХАЛЕВСЬКА, В. Ц. МІХАЛЕВСЬКИЙ
Хмельницький національний університет

# ДЕЯКІ ЗАВДАННЯ СУЧАСНИХ КОРПОРАТИВНИХ СХОВИЩ ДАНИХ

*Розглянуті основні завдання та вимоги до сучасних корпоративних сховищ даних враховуючи зростаючі обсяги оброблюваної інформації і конкурентне середовище для IT-організації.*

Розвиток технології сховищ даних розпочався з необхідності розділити дані, які використовуються для операцій, і дані, які використовуються в аналітичних цілях. Сховище даних забезпечує можливості, найбільш придатні для звітності. Крім того, розділення користувачів, що виконують транзакції, і користувачів звітності, чиї нерегламентовані запити можуть негативно позначитися на ефективності оперативних систем, гарантують оптимальне використання ресурсів інфраструктури даних.

І хоча сховища дають організації чудову платформу звітності і аналізу, в реальному часі вони, як правило, не працюють. Через технологічні обмеження сховища зазвичай поповнюються вночі за допомогою пакетної передачі даних. Для цього використовується пакетна програма, яка виконує вертикальне читання всієї бази з пошуком змін. Дані, що поступають в сховище за допомогою такого ETL-підходу, - завжди застарілі (як правило, на добу).

В міру зростання об'єму оброблюваних даних, а також кількості і різноманітності систем обробки даних, збільшується час і складність процесу наповнення сховища. Разом з тим, глобалізація, збільшення тривалості експлуатації систем, обмеженість договорів про сервісне обслуговування приводять до необхідності скорочення пакетних операцій. Поєднання більшої кількості даних і конкурентного тиску створює серйозні проблеми для IT-організації.

Рішення, прийняті на основі вчорашніх даних, перестають задовольняти більшість організацій. Прийняття рішень в реальному часі вимагає і даних в реальному часі, що накладає особливі вимоги на інтеграцію даних для сховища.

Крім того, аналітичні операції, що виконуються в сховищі, необхідно знову передавати в OLTP-систему, звідки поступили дані. Таким чином відбувається централізація аналітичної обробки і гарантується передача рішень, прийнятих на агрегованих даних в сховищі, у відповідні OLTP-системи.

Ці тенденції реалізуються таким чином:

*1. Інтеграція даних в реальному часі для сховища даних.* Отримання і передача даних в реальному часі з операційних систем в сховище, що робить дані доступними для аналізу.

*2. Активне сховище даних.* СД в реальному часі, доповнюване інструментами Business Intelligence для обробки і виконання бізнес-рішень. Рішення автоматично передаються в OLTP-системи. В результаті формується замкнутий цикл обробки.

У прагненні добитися функціонування сховища в режимі реального часу, успіх часто залежить від грамотного вибору інструменту інтеграції і підходу до отримання даних, що забезпечує можливість підвищення якості і своєчасності інформації.

Для підтримки інтеграції в реальному часі пакетний підхід до витягання операційних даних повинен бути замінений на процеси, які постійно відстежують стан початкових систем, захоплюють і перетворять зміни в даних по мірі виникнення, потім завантажують їх в сховище в режимі, максимально наближеному до реального часу. Постійний збір даних дозволяє аналізувати прибуток і цінові елементи в будь-яких часових рамках. Тенденції можна аналізувати з будь-якою вибраною періодичністю і без затримки.

ETL є ідеальним рішенням задачі завантаження великих об'ємів даних в сховище, а також дає широкі можливості перетворення даних. Проте ETL-операції зазвичай виконуються в момент припинення оновлення початкової системи, щоб гарантувати, що у момент отримання даних джерело не змінюється. Це, у свою чергу, приводить до невідповідностей між OLTP-системами і сховищем. В результаті дані і додатки не завжди доступні бізнес-користувачам.

EAI-рішення (Enterprise Application Integration - інтеграція корпоративних застосувань), раніше призначені для інтеграції додатків, сьогодні часто конкурують або співіснують з ETL-технологіями, будучи засобами інтеграції і отримання даних в реальному часі. EAI-рішення передають інформацію між початковою і цільовою системами, гарантують постачання даних, забезпечують розвинену підтримку потоку і спрощують основні елементи перетворення.

Проте EAI-технологія накладає обмеження на об'єми, оскільки початковим завданням цього методу була інтеграція саме додатків (а не даних), і суть його в запуску додатків та передачі інструкцій і повідомлень. Проте, можливість переміщати інформацію в реальному часі і підтримувати її цілісність в процесі інтеграції у ряді випадків робить технологію EAI придатною для взаємообміну між операційними системами і активним сховищем.

Іншим підходом до інтеграції даних в реальному часі є технологія управління транзакційними даними (transactional data management - TDM), призначена для отримання, передачі, перетворення, постачання і верифікації транзакційних даних в гетерогенному середовищі в реальному часі. TDM функціонує на виконаних

транзакціях: вибирає їх з OLTP-системи, застосовує основні методи перетворення і передає їх в сховище. По своїй архітектурі технологія асинхронна, проте забезпечує синхронну поведінку, працює із затримкою в частку секунди, підтримуючи цілісність даних в транзакції.

EAI і TDM призначені для передачі змін і оновлень даних, а не цілісних вибірок даних. Ні перше, ні друге не вимагає призупинки початкових систем, оскільки ці технології підтримують цілісність операцій мови маніпулювання даними (data manipulation language - DML). За рахунок цього істотно скорочується об'єм необхідних переміщень даних. І якщо ETL-засоби в основному призначені для початкового завантаження і перетворення даних, то EAI і TDM більше підходять для постійного збору даних.

Все більша кількість компаній використовують TDM-технологію з метою збору даних для сховища. TDM-засоби захоплюють, направляють, доставляють і перевіряють операції з даними в середовищі гетерогенних баз даних із затримкою в долі секунди.

Передача змінених даних на рівні транзакції дозволяє системі працювати в активному режимі і обробляти операції одночасно з наповненням сховища. В цьому випадку повністю усувається залежність інтервалу пакетної обробки і зберігається цілісність кожної транзакції.

Інтеграція сховища і OLTP-системи передбачає отримання і передачу транзакційних даних в сховище *одночасно з передачею даних про прийняті рішення* на основі даних СД в одну або декілька оперативних систем. Такий замкнутий цикл роботи також забезпечується засобами TDM.

В задачі інтеграції DW і OLTP можливе комбінування TDM і ETL-процесів. Зокрема для обробки даних в реальному часі, постійному захопленні і витяганні даних на транзакційному рівні. Засоби TDM можуть передавати дані в реальному часі в проміжний рівень зберігання цільової БД, де ETL-сервер перехоплюватиме дані і, застосувавши до них перетворення, завантажуватиме в сховище. У такого підходу є недоліки (зокрема, додаткова затримка і необхідність підтримувати ETL-сервер), проте вони обгрунтовані у випадку, якщо вимоги до перетворення даних дуже високі.

Переваги в тому, що нові транзакційні дані негайно захоплюються з дуже малим ефектом по продуктивності на OLTP-систему (в порівнянні із звичайним ETL-процесом).

При побудові сховища даних необхідно забезпечити можливість простого і ефективного доступу користувачів до аналізованої інформації. Для вирішення цього завдання використовують інструменти Business Intelligence. Продукти даного класу надають можливість проведення OLAP-аналізу (обертання даними, проведення деталізації, сортування і т.д.), а також дозволяють проглядати інформацію в зручному для сприйняття вигляді (графіки, зведені таблиці, звіти), що дозволяє приймати обгрунтовані рішення.

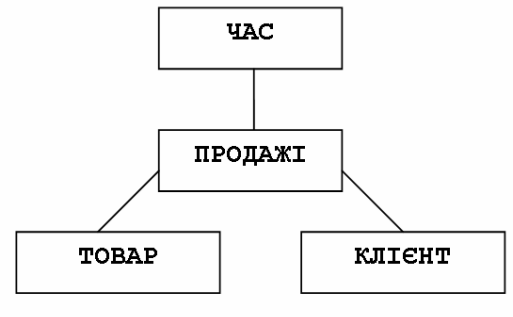
Для забезпечення можливості використання засобів Business Intelligence розробникам сховищ даних доводиться вирішувати ряд завдань: *проектування сховища з використанням багатовимірної моделі даних; розробка процедур ETL; забезпечення прийнятної якості даних.*

При побудові сховища зазвичай використовують багатовимірну модель даних. При такому підході інформація розбивається на два класи: факти і виміри. Факти - це числові характеристики, що позначають деяку подію. Наприклад, на рис. 1 в центрі схеми зображений факт (“продажі”), який визначає суму, витрачену клієнтом.

Факти завжди оточені текстовим контекстом - вимірами. На малюнку зображено три виміри, в яких задається інформація про товари, час здійснення операції і клієнтів (“товар”, “час”, “клієнт”).

Для наповнення сховища інформацією використовується програмне забезпечення класу ETL (Extract transfer load). Програмне забезпечення цього класу призначене для витягання, приведення до загального формату, перетворення і завантаження даних в сховище. Існують два підходи до написання ETL-процедур: їх можна написати вручну або скористатися спеціалізованими засобами ETL. Кожний з підходів має ряд переваг і недоліків, і вибір одного або іншого методу написання процедур ETL визначається вимогами до підсистеми завантаження даних в кожному конкретному випадку.

Не дивлячись на досвід і методики, накопичені за більш ніж 30-річну історію, проекти по створенню сховищ даних залишаються дуже ризикованими. За статистикою: 37% проектів припиняються, не отримавши яких-небудь результатів; 50% проектів доводяться до логічного завершення, але при цьому перевищуються терміни або бюджет на 20% і більше; 13% складають успішні системи.




*Багатомірна модель даних*

**Рис. 1**

При цьому основним чинником ризику, що визначає успішність проекту по створенню сховищ даних, є проблема якості даних, що включає: *коректність* (всі значення, що містяться в сховищі даних, є достовірними і безпомилковими); *однозначність* (будь-яка запитана інформація повинна мати єдине значення, щоб вона не могла тлумачитися різними користувачами по-різному); *узгодженість* (інформація, що поступає в сховище даних, повинна відповідати єдиним вимогам); *повнота* (забезпечення того, щоб всі необхідні величини містили непорожні значення, і забезпечення контролю попадання в сховище даних всіх необхідних записів).

Для вирішення проблеми якості даних розробник може скористатися програмними засобами, що існують на ринку, такими як: системи профілізації інформації, системи моніторингу даних, засоби очищення інформації. Корисним може стати використання спеціалізованих засобів управління довідниками, які дозволяють управляти інформацією в предметній області, пов'язаній з певним вимірами. Як приклад наведемо засіб Oracle Customer Hub, який дозволяє управляти інформацією про клієнтів. Проте, використання даних подібного засобу в більшості проектів виявляється недостатнім, і розробникам доводиться реалізовувати додаткову логіку контролю якості даних на етапі ETL.

**Висновки**

Сховища даних раніше призначалися тільки для звітності, аналізу і прогнозування. Сьогодні все більше компаній прагне до активних операційних сховищ, а тому важливою вимогою стає обмін даними між СД і OLTP-засобом - в реальному часі і з мінімальною затримкою. Інструменти інтеграції даних, зокрема TDM, дають можливість рішення цієї задачі.

**Список використаних джерел**

1. Эрик Спирли. Корпоративные хранилища данных. Планирование, разработка и реализация. – Т.1. – Вильямс, 2007. – 400 с.

2. Kimball R., Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Confirming and Delivering Data. Wiley, 2004. – 525 p.