УДК 004.652.4+004.827

Я. І. Виклюк, д-р техн. наук,
Н. Б. Шаховська, канд. техн. наук

СХОВИЩА ТА ПРОСТОРИ ДАНИХ – ІНФОРМАЦІЙНИЙ ФУНДАМЕНТ СИСТЕМ ПРИЙНЯТТЯ РІШЕНЬ

Анотація *У статті наведено порівняння інформаційних технологій аналізу та інтеграції даних, таких як сховища даних і простори даних.*

Ключові слова: *простори даних, бази даних, системи підтримки прийнятя рішень.*

Y. Vyklyuk, DSc,
N. Shahovska, PhD

DATA WAREHOUSE AND DATA SPACE – INFORMATION FOUNDATION OF DECISION SUPPORT SYSTEMS

Abstract. *In the article there are given a comparison of information technology for integration and data analysis such as datawarehouse and dataspace.*

Keywords: *spatial data, databases, systems support decisions.*

Я. И. Виклюк д-р техн. наук,
Н. Б. Шаховская, канд. техн. наук

ХРАНИЛИЩА И ПРОСТРАНСТВА ДАННЫХ – ИНФОРМАЦИОННЫЙ ФУНДАМЕНТ СИСТЕМ ПРИНЯТИЯ РЕШЕНИЙ

Аннотация. *В статье приведены сравнения информационных технологий анализа и интеграции данных, таких как хранилища данных и пространства данных.*

Ключевые слова: *пространства данных, базы данных, системы поддержки принятия решений.*

Вступ. Традиційно при розв'язанні задач фахівці використовують звичні для них джерела інформації. При такому підході очевидною є неповнота інформації, яку вдається опрацювати. Безліч джерел даних і сервісів, що існують в Інтернеті, їхня розмаїтість, викликають потребу в радикальній зміні методів отримання даних. Ця зміна полягає в тому, що задачі мають формулюватися незалежно від існуючих джерел даних. Після їх формулювання має здійснюватися ідентифікація релевантних джерел, приведення даних до необхідного виду, інтеграція, ідентифікація сервісів, які дають змогу розв'язати окремі частини задачі [1].

Для прийняття адекватних рішень потрібно, щоб дані, які надходять із різних джерел, задовольняли такі вимоги:

були повними, несуперечливими та надходили вчасно;

були інформативними, оскільки повинні застосовуватися для прийняття рішень;

були однакової структури, щоб мати можливість завантажити їх у єдине сховище даних та проаналізувати;

зберігалися в однакових моделях даних та були незалежними від платформи розроблення, щоб мати можливість використання цих даних в інших засобах.

Однак, на сьогодні немає жодної методики опрацювання даних, яка б задовольняла всі наведені вимоги до опрацювання даних.

Проблема інтеграції розрізненої інформації з метою її подальшого опрацювання та прийняття рішень на її основі постала разом із появою сховищ даних, ще у 80-х роках минулого століття. Передумовою її виникнення було зростання інтересу до розподілених баз даних та їх масове впровадження у бізнесові структури. Значний внесок у вирішення цієї проблеми зробили вчені: Colin White, A. Sheth, J. Larson, К. В. Антипін, А. В. Фомичев, М. Н. Гриньов, С. Д. Кузнецов та ін.

Розроблені на сьогодні підходи інтеграції даних за своєю функціональністю поділяються на два типи: інтеграція вебзастосувань та інтеграція на основі сховищ даних (з утворенням локального сховища даних). Проте специфіка опрацювання даних усієї предметної областізагальна, характеризується:

наявністю великої кількості джерел даних, інформація в яких має різну структуру, не виключаються протиріччя та суперечливість інформації,

наявністю великої кількості моделей зберігання джерел даних (реляційні бази даних (РБД), сховища даних (СД), напівструктуровані текстові файли, електронні таблиці, статичні та динамічні веб-сайти тощо),

відсутністю стандартів наіменування об'єктів і суб'єктів певних галузей діяльності (туристична, наукова тощо).

Ієрархічне впорядкування об'єктів складних галузей та агрегування інформації під час передавання її до верхніх рівнів ієрархії вказує на те, що для врахування інформації від усіх об'єктів галузі необхідно поєднати обидва типи інтеграції та вдосконалити наявні моделі зберігання даних.

Стаття присвячена визначенню відмінностей між засобами зберігання та опрацювання даних та встановленню їх місця в системах прийняття рішень.

2. Огляд літературних джерел

2.1. Методи опрацювання даних з різних джерел

Є такі методи опрацювання даних з джерел з різними структурами даних.

1. *Пошук інформаційних джерел на основі метаданих*. Це найпоширеніший метод пошуку інформації, необхідної для користувача [10, 15], робота якого базується на основі посередника. Програма посередника отримує модель вимог користувача до даних, що він шукає. Пошук здійснюється серед джерел даних, інформація про яких розміщена у репозиторії посередника. Також додаються нефункціональні вимоги до інформаційних джерел (наприклад, необхідні характеристики якості даних, такі як точність, частота оновлення). У свою чергу, специфікація посередника шукає джерела, що задовольняють вказані вимоги, визначає структури даних та допустимі операції над даними джерела.

Для специфікації метаданих моделі вимог використовують типи метаданих. Тип метаданих, визначений для посередника, є частиною його моделі вимог і може не збігатись з типами метаданих потенційних джерел. Самі вимоги вводяться як інваріанти типа метаданих. Моделі вимог на класи посередників поширюється за допомогою метакласів, екземплярами яких стають класи посередника. Кожен клас посередника може бути екземпляром одного або декількох метакласів.

До переваг методу пошуку на основі метаданих слід віднести швидкий доступ до даних і точність результатів пошуку.

Недоліком пошуку інформації на основі метаданих є неможливість фізичного доступу до них через відсутність знань про структуру джерела, яке ще не перебуває у репозиторії метаданих.

2. *Відбір джерел з використанням онтологічних визначень*. Є розширенням методу пошуку на основі метаданих. Перш за все будується онтологічна модель предметної області, задаються онтологічні специфікації понять предметної області і зв'язків між ними. Онтологічні специфікації використовуються для пошуку класів і типів інформаційних джерел, релевантних класам і типам посередника. Елемент специфікації джерела вважається онтологічно релевантним елементу специфікації посередника того ж виду (клас, тип, атрибут, функція, параметр), якщо між відповідними їм онтологічними поняттями встановлена позитивна асоціація або асоціація узагальнення/спеціалізації.

Здебільшого онтологічні (їх ще називають семантичні) моделі будуються на основі одного з напрямків логіки першого порядку (числення предикатів) – дескрипційних логіках, які є сімейством мов, що дають смогу формально й однозначно описувати поняття в певній предметній області. Кожний клас («концепт») може бути співвіднесений з іншим подібним йому концептом шляхом додавання тегів метаданих, що вказують на властивості, загальні риси, розбіжності тощо. Розширення моделей тегами дає змогу створювати такі структури, яких раніше не могло бути. У семантичній моделі будь-яка інформаційна одиниця подається графом, що спрощує її модернізацію; наприклад, злиття двох моделей зводиться до об'єднання їхніх графів. Інформаційна одиниця може бути представлена ідентифікатором Uniform Resource Identifier (URI), за допомогою якого можуть бути встановлені зв'язки між двома або більше інформаційними одиницями. Семантичні моделі можуть створюватись з використан-

ням Resource Description Framework (RDF) моделі для подання даних, розробленої W3C мовою Web Ontology Language (OWL).

Цей підхід на теоретичному рівні описано в роботі [7], проте ще немає практичних розробок, оскільки онтології розроблені для невеликої кількості предметних областей. Для української мови кількість онтологій взагалі обмежена.

3. *Інтеграція на рівні сховищ даних.* Метою інтеграції даних на рівні їх сховища є отримання єдиної і цілісної картини корпоративних даних предметної області [2, 4]. Типовим підходом до інтеграції інформації в масштабах підприємства є побудова сховищ та вітрин даних на основі отримання оперативних даних, їхньої трансформації до єдиної схеми і завантаження даних у сховище (процедура ETL – extraction, transfotamation, loading).

Перевагами інтеграції на рівні сховищ даних є добре пророблене теоретичне та практичне підгрунтя, використання алгоритмів ETL в сучасних СУБД, розролення серверів інтеграції.

Цей підхід придатний для використання на підприємстві з декількома десятками операційних баз даних, що перебувають під єдиним контролем.

Проте для джерел даних, що не входять в корпорацію (зберігаються в Internet-середовищі), парадигма ETL або не може застосовуватись, або її застосування обмежене. Як правило, об'єкти глобальної мережі не дозволять у масовому порядку витягувати дані зі своїх операційних баз даних, до них можна буде адресувати лише одиничні запити. У результаті цього потрібно здійснювати інтеграцію, можливо, мільйонів інформаційних джерел "на льоту". Тому існують такі невирішені проблеми [4]:семантична неоднорідність; неповнота і неточність даних; обмеженість доступу до конфіденційних даних і т.ін.

4. *Web-інтеграція.* Web-інтеграція – це методи опрацювання і подання інформаційних ресурсів за допомогою Web-технологій. Інтеграція на основі Web-сервісів має декілька рівнів. Рівень наповнення (даних) передбачає інтеграцію даних з відомими СДІР і є найпростішим. Наступний рівень – об'єктна взаємодія. Тут мова йде про те, що програмне застосування, розташоване на одному сервері, може запускати програмні процеси на іншому. Третій рівень інтеграції – інтеграція на рівні стандартної семантики. На цьому рівні сервіси можуть «спілкуватися спільною мовою», обходячи технологічні розбіжності.

Якщо семантичних розбіжностей між джерелеми даних немає, інтеграція не має особливих труднощів. Тобто, якщо використовується специфікація WSDL, програмне застосування може «говорити» системно-незалежною мовою. Системна незалежність застосувань можлива при використанні мови XML при створенні WSDL-описів та протоколу SOAP [7].

Отже, на рівні сховища даних доцільно використовувати традиційні методи інтеграції, а на рівні простору даних – семантичну інтеграцію або розширення традиційної шляхом попереднього визначення структури даних джерела та методів доступу до даних.

3. Означення сховища та простору даних

Сховище даних – агрегований інформаційний ресурс, що містить консолідовану інформацію з усієї проблемної області та використовується для підтримки прийняття рішень.

Консолідована інформація – це одержані з декількох джерел та системно інтегровані різнотипні інформаційні ресурси, які в сукупності наділені ознаками повноти, цілісності, несуперечності та складають адекватну інформаційну модель проблемної області з метою її аналізу опрацювання та ефективного використання в процесах підтримки прийняття рішень.

Простір даних DS – множина усіх інформаційних продуктів предметної області

$$DS = <DB, DW, Wb, Nd, Gr>,$$

де DB, DW, Wb, Nd, Gr – інформаційні продукти, що подають множини баз даних, сховищ даних, веб-сторінок, текстових файлів, електронних таблиць, графічних даних відповідно.

Говорячи про інформаційний продукт, матимемо на увазі його вміст (інформаційний ресурс), а також множину відомостей про нього (розміщення, схема доступу, швидкість

оновлення інформації тощо). Також нас цікавитимуть операції, які виконуються над IP залежно від його СДІР.

Хоча інформаційні продукти, що входять у ПД, мають різні структури даних, методи доступу, проте вони усі *виконують однакову роль*: надають дані для простору даних через фіксацію свого стану та забезпечують виконання притаманних для них операцій, причому ці операції та їх результати визначені для усього простору даних.

Спектр застосування технології сховищ даних достатньо широкий. Задачі, що розв'язуються за допомогою сховищ даних, як правило, належать до класу задач керівного аналізу і стратегічного планування. Нижче наведено приклади типових питань, на які можна відповідати за допомогою сховищ даних: фінансовий аналіз, аналіз продажу, аналіз прибутковості, аналіз каналів продажу, аналіз клієнтської бази, маркетинг, аналіз якості обслуговування клієнтів, аналіз складських запасів, аналіз постачальників, аналіз персоналу.

Простори даних використовуються для стратегічного планування на корпоративному, галузевому та державному рівнях, оскільки поєднують дані з баз даних, сховищ даних, текстових файлів тощо і використовують їх для підтримки прийняття рішень.

4. Засоби реалізації сховищ та просторів даних

Для реалізації сховищ та просторів даних використовують системи управління базами даних, засоби обміну даними та інтеграції.

Джерела даних, такі як електронні таблиці, мультимедійна інформація тощо, можуть мати свої власні засоби зберігання та опрацювання, і тоді завданням засобів інтеграції є розпізнавання цих інформаційних ресурсів та організація доступу до них. Коли йде мова про сховища даних, то структура джерел є відомою наперед і основним завданням є очищення та завантаження самих даних.

Для просторів даних необхідно забезпечити можливість працювати з програмним продуктом, якого теоретично може і не бути на робочій станції користувача. Якщо не передбачати такої можливості, то потрібно передбачити розроблення сховища даних такої структури, щоб у нього можна було завантажити дані з джерел даних для забезпечення відповіді на запит користувача.

Порівняльна характеристика деяких засобів реалізації сховищ даних подана в таблиці.

Засоби інтеграції можна розділити на два умовних класи: інтеграції застосувань та інтеграції веб-застосувань. Засоби інтеграції застосувань реалізуються за допомогою проміжного шару, спеціалізованих засобів та серверів інтеграції. Засоби інтеграції застосувань розроблено в Oracle, SQL Server. Для інтеграції геоінформації та баз даних використовують спеціалізовані засоби, які перетворюють векторні дані у спеціальний формат.

Простір даних вимагає набагато більше технологічних та методичних рішень, оскільки у ньому опрацьовується інформація з різними структурами даних, наперед невизначеними, а також використовуються різні засоби опрацювання та зберігання даних.

Якщо розглядати технології, які дають змогу реалізувати можливості просторів даних, то в першу чергу необхідно зупинитися на *грід* та *хмарних обчисленнях* (англ. Cloud Computing).

Сервіс-орієнтірована грід-технологія надає нові можливості, яких не було в мережах, організованих за схемою peer-to-peer або клієнт-сервер. Функціонування і взаємодія сервісів близькі до технології мультиагентних систем, а роль інтелектуальних агентів виконують грід-сервіси. При цьому в такому поданні вони мають ряд переваг перед web-сервісами. Серед них:

можливість реалізації функціональності пошуку даних, не обмеженої набором процедур, реалізованих на сервері сховища даних;

можливість проводити аналіз як в глобальних, так і в корпоративних мережах;

можливість продовження роботи сервісів-агентів пошуку та збору даних і після виконання конкретного запиту;

вбудована можливість передачі прав доступу до даних від користувача до усієї послідовності грід-сервісів за допомогою цифрового сертифікату.

1. Порівняльна характеристика засобів реалізації сховищ даних

Засіб	Переваги	Недоліки
Oracle Warehouse Builder Oracle Data Integration Oracle Optimized Warehouse Hyperion	СУБД рівня корпорації; можна використовувати як компоненту орієнтованої на дані архітектури в середовищі SOA або BI. Включає: переміщення даних, синхронізацію, перевірку якості даних, керування даними, сервіси перевірки актуальності даних	Для інтеграції необхідно попередньо описати джерела даних та в ручному режимі налагодити процедури перевірки якості даних
Database Application Server	Платформа для створення і розгортання розрахованих на багато користувачів мережевих програм для Web, клієнтами яких можуть бути як стандартні браузери, так й Java-застосування і аплети)	ПЗ проміжного рівня
SQL Server 2008	Містить засоби інтеграції Integration Services, аналізу Analysis Services, формування звітів Reporting Services, набір інструментів керування сховищами даних Management Studio, набір інструментів розроблення застосувань інтелектуального опрацювання даних Business Intelligence Development Studio	Як і для Oracle, працює з наперед відомими джерелами
Biz Talk	Сервер інтеграції, дозволяє аналізувати текстові дані та записувати у сховище даних; функціонує як на внутрішньокорпоратвиному, так частково і на міжкорпоративному рівнях	Функціонує за принципом сповіщень, що робить якість інтеграції сильно залежною від користувача

Сервіс-орієнтірований підхід до слабкозв'язаних масивів даних як до простору даних дозволяє вже сьогодні створювати сервіси нового рівня, що оперують не тільки БД або метаданими, але і працюють безпосередньо з Web-даними та іншими слабкоструктурованими ресурсами. При цьому відчувається потреба застосування доопрацьованих технологій СУБД для просторів даних, що в сумі уможливлює вирішення проблеми інтеграції федеральних гетерогенних інформаційних ресурсів з нових позицій.

Хмарні обчислення – технологія опрацювання даних, у якій програмне забезпечення надається користувачеві як Інтернет-сервіс. Користувач має доступ до власних даних, але не може керувати і не повинен піклуватися про інфраструктуру, операційну систему і власне програмне забезпечення, з яким він працює.

Надання у користування обчислювальних потужностей і баз даних датацентра може здійснюватися декількома варіантами: Saa, Paa, Haa, Iaa, Caa – as a Service – як Інтернет-Сервісу. *S, P, H, I, C* – Software,

Platform, Hardware, Infrastructure, Communication – відповідно, програмне забезпечення, платформа, апаратне забезпечення, інфраструктура, комунікації. Платформу парадигми Cloud Computing формують складові частини: віртуалізація, SAAS, SOA (Service-Oriented Architecture) і S+S (Software + Services).

На сьогодні є такі *реалізації хмарних обчислень*, пов'язані з просторами даних.

Windows Azure – хмарна операційна система компанії Microsoft, призначена для розроблення й запуску веб-застосувань, які виконуються на сервері постачальника, а не на комп'ютері користувача. Входить до складу платформи Microsoft Azure. Використовує реляційні структури для зберігання даних. Також передбачається використання SharePoint в хмарі.

Google App Engine – це платформа, яка дає змогу використовувати одну інфраструктуру для створення й хостинга своїх застосувань. Використовує нереляційне розподілене сховище даних.

Висновки

Стаття присвячена аналізові таких засобів побудови систем прийняття рішень, як сховища та простори даних, а також загальному аналізу галузі комп'ютингу.

Наукова новизна статті – встановлення формальних відмінностей між такими об'єктами, як бази даних, сховища даних та простори даних.

Практичне значення статті полягає у визначенні основних задач компонент систем прийняття рішень та зв'язку між ними.

Список використаної літератури

1. Turban, E. Decision support and expert systems: management support systems / E. Turban. – Englewood Cliffs, N. J.: Prentice Hall, 1995.

2. iMeMex:Personal Dataspace Management System. – [Електронний ресурс]. – [Режим доступу] http://www.dbis.ethz.ch/research/current_projects/iMeMex

3. Dan Linstedt. Data Vaulttm overview the next evolution in data modeling / Dan Linstedt – 2005, http://www.tdan.com/i021hy01.htm.

4. Огляд технологій інтеграції інформаційних систем, 2006, http://www.microsoft.com/Ukraine/Government/Analytics/IntegrationTechnologies/Overview.msp

5. Кузнецов, С. Пространства данных: исследовательский полигон или путь к новому поколению систем управления данными? http://synthesis.ipi.ac.ru/sigmod/seminar/s20060420.

6 Donald Kossmann, Jens-Peter Dittrich. Personal Data Spaces. http://www.inf.ethz.ch/news/focus/res_focus/feb_2006/index_DE.

7. Рогушина, Ю. В. Формирование тезауруса предметной области как средства моделирования информационных потребностей пользователя при поиске в Интернете / Ю. В. Рогушина, А. Я. Гладун // Вестник компьютерных и информационных технологий. – М. : 2007. – № 1. – С. 26–33.

8. ETH – Databases and Information Systems – iMeMex, www.dbis.ethz.ch/research/current_projects/iMeMex

9. Processing of natural language queries to a relational database. Samsonova, M., Pisarev, A., Blagov, M. http://www.cs.dartmouth.edu/~brd/Teaching/AI/Lectures/Summaries/natlang.html /

10. Основные концепции и подходы при создании контекстно-поисковых систем на основе реляционных баз данных. http://www.citforum.ru/database/articles/search_sys.shtml.

11. Особенности построения хранилищ данных. http://citforum.uar.net/seminars/cis99/sch.shtml/

12. Kacprzyk J. Database Queries with Fuzzy Linguistic Quantifiers / J. Kacprzyk, A. Ziolkowski // IEEE Transactions on Systems, Man, and Cybernetics. SMC-16, 1996. – P. 512–529.

13. Fuzzy Grouping в Microsoft SQL Server 2005 http://msdn.microsoft.com/msdnmag/issues/05/09/SQLServer2005/default.aspx.

14. Пелещишин, А. М. Методи та алгоритми моделювання Web-систем / А. М. Пе-

лещишин // Вісник ДУ "Львівська Політехніка". –Львів : – 2000. – № 406. – С.199–211.

15. Тенденции в области Хранилищ данных на 2007 год / [Електронний ресурс] / TWAN. – 2007. — Режим доступу: http://citcity.ru/15272/.

Отримано 30.09.2012

References

1. Turban, E. Decision support and expert systems: management support systems / Turban, E. – Englewood Cliffs, N.J.: Prentice Hall, 1995 [in English].

2. iMeMex: Personal Dataspace Management System. – [Електронний ресурс]. – [Режим доступу] http://www.dbis.ethz.ch/research/current_projects/iMeMex [in English].

3. Dan Linstedt. Data Vaulttm overview the next evolution in data modeling. – 2005, http://www.tdan.com/i021hy01.htm [in English].

4. Information systems integration technology owerview, 2006, http://www.microsoft.com/Ukraine/Government/Analytics/IntegrationTechnologies/Overview.msp [in English].

5. Kuznetcov, Sergey. Data spaces: research polygon or an approach to data control new generation? http://synthesis.ipi.ac.ru/sigmod/seminar/s20060420 [in Russian].

6 Donald Kossmann, Jens-Peter Dittrich. Personal Data Spaces. http://www.inf.ethz.ch/news/focus/res_focus/feb_2006/index_DE [in English].

7. Rogushina, Y. V. Formation of thesaurus as a means of modeling user information needs when searching on the Internet / Y. V. Rogushina, A. J. Gladun // Computer and Information Technology. – New York : – 2007. – № 1. – P. 26–33 [in Russian].

8. ETH – Databases and Information Systems – iMeMex, www.dbis.ethz.ch/research/current_projects/iMeMex [in English].

9. Processing of natural language queries to a relational database. Samsonova, M., Pisarev, A., Blagov, M. http://www.cs.dartmouth.edu/~brd/Teaching/AI/Lectures/Summaries/natlang.html [in English].

10. Basic concepts and approaches to create context-search systems based on relational databases. http://www.citforum.ru/database/articles/search_sys.shtml [in Russian].

11. Characteristics of building a data warehouse. http://citforum.uar.net/seminars/cis99/sch.shtml/ [in Russian].

12. Kacprzyk, J. Database Queries with Fuzzy Linguistic Quantifiers / J. Kacprzyk, A. Ziolkowski // IEEE Transactions on Systems, Man, and Cybernetics. SMC–16. – 1996. – P. 512–529 [in English].

13. Fuzzy Grouping в Microsoft SQL Server 2005 http://msdn.microsoft.com/msdnmag/issues/05/09/SQLServer2005/default.aspx [in English].

14. Peleshchyshyn, A. M. Methods and algorithms for modeling Web-systems / A. M. Peleshchyshyn // Bulletin of control "Lviv Polytechnic". – Lviv : – 2000. – № 406. – P. 199–211 [in Ukrainian].

15. Trends in the field repository of data for 2007 / [electronic resource] / TWAN. – 2007. - Mode of access: http://citcity.ru/15272/ [in Russian].

Шаховська
Наталя Богданіна,
к.т.н., доц., каф. Інформаційних систем та мереж Нац. ун-ту «Львівська політехніка»,
тел. (063)275025

Виклюк
Ярослав Ігорович,
д.т.н., доц. проректор з наукової роботи та міжнародних зв'язків, декан факультету Комп'ютерних систем і технологій Буковинського ун-ту,
м. Чернівці