УДК 004.652.4+004.827

Н. Шаховська, Я. Виклюк
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ЗАСТОСУВАННЯ СХОВИЩ ТА ПРОСТОРИ ДАНИХ У СИСТЕМАХ ПРИЙНЯТТЯ РІШЕНЬ

© Шаховська Н., Виклюк Я., 2012

Описано методи побудови інтелектуальних систем прийняття рішень. Для опрацювання різнотипних даних обрано простір даних.

Ключові слова: сховище даних, простір даних, системи прийняття рішень, якість даних.

This article is described construction methods of intellectual systems of decision-making. To process data different is selected dataspace.

Key words: datawarehouse, dataspase, decision support systems, data quality.

Вступ

Введемо ряд означень.

Інформаційний ресурс (ІР) –документи і масиви документів в інформаційних системах: бібліотеках, архівах, фондах, банках даних, інших класах інформаційних систем, організовані для багаторазового використання та задоволення потреб користувача.

Структура даних ІР (СДІР) – опис складних інформаційних об’єктів засобами простіших типів даних. Характеризується множиною допустимих значень; множиною допустимих операцій; характером організованості.

Інформаційний продукт (ІП) – документований інформаційний ресурс, підготовлений відповідно до потреб користувачів і поданий у формі товару. Інформаційними продуктами є програмні продукти, текстові файли, веб-сторінки, електронні таблиці, xml-файли, бази даних, сховища даних та ін.

Каталог ІП – метадані про інформаційні продукти. Описує місцезнаходження інформаційного продукту, його структуру даних, методи доступу до інформаційних ресурсів тощо.

1. Постановка задачі

Важливою науковою проблемою, яка виникає у слабкоструктурованому середовищі, є розроблення і удосконалення методів опрацювання різнотипних даних з метою підтримки прийняття рішень. Опрацювання інформаційних ресурсів, що використовують різні моделі даних, схеми керування тощо вимагає розроблення уніфікованого методу доступу до них для того, щоб надати можливість користувачу вибирати адекватний інструментарій для вивчення та використання різних засобів опрацювання даних. Необхідність у цьому виникає в організацій, робота яких полягає в опрацюванні великої кількості різнотипних, взаємонезалежних джерел даних, для яких не всі семантичні взаємозв'язки відомі та вказані. У деяких випадках семантичні зв'язки невідомі через невизначену кількість початкових джерел або через брак людей, кваліфікованих у визначенні таких зв'язків. У інших випадках не всі семантичні зв'язки необхідні для класифікації послуг користувачам. Тому в користувачів немає єдиної технології, за якою вони можуть створювати запити відносно цільових задач.

Для прийняття адекватних рішень необхідно, 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, щоби можна було завантажити їх у єдине сховище даних та проаналізувати;

– зберігалися в однакових моделях даних та були незалежними від платформи розроблення, щоби можна було використовувати їх в інших засобах.

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

У статті розглянуто відмінності між засобами зберігання та опрацювання даних та визначено їх місце в системах прийняття рішень.

2. Аніліз літературних джерел

2.1. Методи опрацювання даних з різних джерел

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

1. *Теорія* інтеграції [1] є підмножиною теорії баз даних і формалізує основні поняття за допомогою логіки першого порядку.

Система інтеграції даних формально визначається як трійка <*G*, *S*, *M*>, де *G* є глобальною схемою (або схемою посередника), *S* – множина схем різнорідних джерел даних, *M* – відображення між запитом джерела і запитом глобальної схеми. *G* і *S* подаються виразами мови, алфавіт якої складається зі символів, що встановлюють відображення між символами, притаманних обом схемам (тут теорія інтеграції переплетена з роботами щодо інтеграції даних через створення метамови). Відображення *M* складається з тверджень між запитами на G і запитами на S. У випадку, коли користувачі створюють запити за системою інтеграції даних, вони встановлюють зв'язки між елементами в глобальній схемі та схемі джерела.

База даних через схеми визначається як множина наборів – по одному для кожного відношення (у реляційної бази даних). База даних відповідного джерела схеми *S* складатиметься з множин кортежів для кожного з різнорідних джерел даних. Це єдине джерело бази даних насправді може бути множиною непов'язаних баз даних. База даних, схема якої відповідає віртуальній схемі посередника *G*, називається глобальною базою даних. Глобальна база даних має задовольняти відображення *M* стосовно вихідної бази даних. Чинність цього відображення залежить від характеру зв’язку між *G* і *S*. Є два способи моделювання цієї відповідності: глобальний (Global As View, GAV), локальний (Local As View, LAV).

У GAV підмножина кортежів, що відображається посередником, є набагато меншою, ніж множина кортежів джерел даних. У LAV кількість кортежів глобальної схеми є набагато більшою, ніж кількість кортежів у джерелах даних. Тому у LAV системах часто доводиться зустрічатися з неповними відповідями.

Опрацювання запитів у системах інтеграції даних зазвичай відображається за допомогою об’єднання. У GAV-системах розробник посередника пише код, щоб перевизначити запис. Кожен елемент в запиті користувача відповідає правилу заміни так само, як кожен елемент глобальної схеми відповідає запиту до джерела. Опрацювання запитів просто розширює підцілі за запитом користувача відповідно до правила, зазначеного у посередника, тобто в результаті запити стають еквівалентні. Найефективнішим з алгоритмів перезапису запитів для GAV є Tsimmis [2].

У LAV-системах процес переписування запитів є радикальніший, оскільки немає посередника, який може встановити відповідність з глобальною схемою. Системі інтеграції необхідно виконати пошук по всьому простору можливих питань для того, щоб знайти ті, які відповідають запиту користувача. У результаті перезапису є ймовірність отримати нееквівалентний запит, але такий, що найбільше відповідає запиту користувача, внаслідок чого і виникає невизначеність у відповіді на запит. Станом на 2009 алгоритм MiniCon є найкращим серед алгоритмів перезапису запитів для LAV [3].

Загалом складність перезапису запитів є NP-повною. Тому для простору даних, що складається зі сотень джерел, алгоритми перезапису можуть використовуватися лише для пошуку даних у визначеній наперед невеликій підмножині джерел.

Іншою назвою «теорії інтеграції» є семантична інтеграція (визначення в ISO 15926). Зазвичай під семантичною інтеграцією розуміють підхід GAV. Для реалізації пропонується використання онтологій та посередника.

2. Побудова канонічних систем. У роботах [4, 5] передбачено, що проектування інформаційної системи (ІС) для вирішення завдань над множинними неоднорідними джерелами інформації є композиційним, основна ідея якого полягає в тому, щоб побудувати композицію специфікацій існуючих, релевантних задачі компонентів (інформаційних, програмних, процесних) так, щоб вона уточнювала абстрактну специфікацію.

З метою проектування специфікації компонентів та ІС приводяться до однорідного подання у канонічній інформаційній моделі. Принциповою складовою процесу композиційного проектування є формальне доведення факту уточнення специфікації ІС композицією специфікацій компонентів. Уточнення системи у системі А означає, що користувач може використовувати систему В замість системи А, не помічаючи факту заміни А на В.

У контексті простору даних ідеї Л.С. Калініченка доцільно застосувати для побудови єдиної схеми джерел даних, які використовуються для відповіді на запит користувача. Це означатиме, що користувач, обираючи в побудованій схемі необхідні йому атрибути, звертатиметься до джерела даних як до власної системи.

Для доведення процесу уточнення в Лабораторії композиційних методів проектування інформаційних систем ІПУ РАН (Росія) розроблено:

– формальну семантику канонічної інформаційної моделі (мови СИНТЕЗ) в нотації абстрактна машина (Abstract Machine Notation, AMN). Як ядро канонічної інформаційної моделі, призначеної для уніфікованого подання специфікацій під час композиційного проектування систем, використовується гібридна об'єктна мова. AMN є формальною мовою специфікацій, що ґрунтується на логіці предикатів першого порядку і теорії множин і призначена для побудови математичних моделей ІС;

– інструментальні засоби, які автоматично відображають специфікації канонічної моделі в AMN.

Ці методи успішно використовуються для опрацювання неоднорідних джерел, що створюються за метаописами мови СИНТЕЗ. Проте у реальних предметних областях вже існують джерела, які не відповідають цим метаописам. Тому теорія Л.С. Калініченка потребує уточнення для розроблення канонічних моделей для джерел з невідомими структурами даних. Проте ідея побудови відображення джерел даних у канонічній формі використано в роботі для побудови сховища консолідованих даних як результату інтеграції даних з джерел.

3. Онтології. Цей метод є розширенням методу пошуку на основі метаданих [6, 7]. Насамперед 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 самого виду (клас, тип, атрибут, функція, параметр), якщо між відповідними для них онтологічними поняттями встановлена позитивна асоціація або асоціація узагальнення/спеціалізації.

Кожний клас («концепт») може бути співвіднесений з іншим подібним концептом завдяки доповненню тегів метаданих, що вказують на властивості, загальні риси, розбіжності тощо. Розширення моделей тегами дає змогу створювати такі структури, яких раніше не могло бути [7]. У семантичній моделі будь-яка інформаційна одиниця подається графом, що спрощує її модернізацію; наприклад, злиття двох моделей зводиться до об'єднання їхніх графів. Інформаційну одиницю може бути подано ідентифікатором Uniform Resource Identifier (URI), за допомогою якого можна встановити зв’язки між двома або більше інформаційними одиницями.

Є декілька проектів систем, що використовують погляд на світ з боку семантики, закладеної в метаданих, і застосовують онтології. До таких проектів відносяться SIMS, HERMES, InfoSleuth, TSIMMIS, Information Manifold [8]. Ці проекти надають доступ до гетерогенних і розподілених інформаційних ресурсів.

У Росії розроблено систему інтеграції даних на основі онтологічного підходу для нафтової галузі [9]. Інтегруються реляційні бази даних за допомогою RDF-RIF-описів. Для інтеграції джерел інших типів (XML) вимагається їх ручне описання в RDF. Перевагою підходу є використання дескриптивної логіки (правила, сформовані в RIF), що дає можливість виводити нові знання. Для порівняння двох баз даних (атрибутів) використовуються міри семантичної близькості. Визначаються на основі нестрогого порядку розміщення в онтології.

Проте метод А.Ф.Тузловського [9] вимагає роботи винятково зі структурованими джерелами і ґрунтується насамперед, на модифікованому підході Калініченка. Окрім того, єдиною мовою формування запитів є SPARQL, що вимагає від користувача деякої підготовки.

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

3. Означення сховища та простору даних

Означення 1. Сховище даних – це агрегований інформаційний ресурс, що містить консолідовану інформацію з усієї проблемної області та використовується для підтримки прийняття рішень.

Означення 2. Консолідована інформація – це одержані з декількох джерел та системно інтегровані різнотипні інформаційні ресурси, які в сукупності наділені ознаками повноти, цілісності, несуперечності та становлять адекватну інформаційну модель проблемної області з метою аналізу її опрацювання та ефективного використання в процесах підтримки прийняття рішень.

Означення 3. Простір даних DS – це множина усіх інформаційних продуктів предметної області

$$DS = \mathbf{<DB, DW, Wb, Nd, Gr>}, \quad (1)$$

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 ІР залежно від його СДІР. Хоча інформаційні продукти, що входять в 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 є визначені для усього простору даних.

Основним завданням простору даних є надання можливості користувачеві працювати з джерелами даних, не знаючи їхніх СДІР, розміщення, методів доступу тощо.

4. Завдання, що вирішуються за допомогою сховищ 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.

Прикладом предметної області, для якої доцільно будувати сховище даних, є університет. Цей об’єкт характеризується наявністю ієрархічної структури, великими обсягами інформації, необхідністю розв’язання задач комплексного аналізу. Для інформатизації ВНЗ зазвичай необхідно інтегрувати дані, оскільки на момент розроблення єдиного сховища даних вже є ряд інформаційних продуктів, які повинні обмінюватися між собою інформацією, а також надавати частину інформації у корпоративне сховище даних з метою її подальшого аналітичного опрацювання. До аналітичних

задач, які необхідно розв'язати аналітикам університету, належать: пошук залежностей між отриманими оцінками студентів за предметами та результатами вступу; пошук дисциплін, у яких показники «Успішність», «Якість» або дуже високі, або дуже низькі; пошук залежностей між результатами наукової діяльності студентів та їх практичними здобутками у вигляді проходження практик, участі в олімпіадах, конкурсах робіт тощо.

Існуючі програмні продукти виконують поставлені перед ним задачі і не завжди дають змогу розв'язати нові. Наприклад, для автоматизації та супроводу навчального процесу «Львівської політехніки» розроблено такі інформаційні системи:

– «Абітурієнт» – облік вступників на перші курси бакалаврату та магістратури, формування наказів на зарахування тощо;

– «Навчальні плани» – розроблення та облік навчальних планів за спеціальностями;

– «Деканат» – облік студентів, облік індивідуальних навчальних планів, облік та аналіз успішності студентів;

– «Розклад» – облік аудиторного фонду, формування розкладу занять та екзаменів;

– «Випускник-працевлаштування» – аналіз якості підготовки випускників «Львівської політехніки», облік практик та дипломних робіт студентів.

Перелічені системи зберігаються на одному сервері під керуванням СКБД SQL Server. Як метод інтеграції використовується інтеграція на рівні сховища даних, технологія тиражування – копіювання визначеної частини даних з одної системи в іншу за певним розкладом. Схему взаємодії основних БД університету зображено на рис. 1.

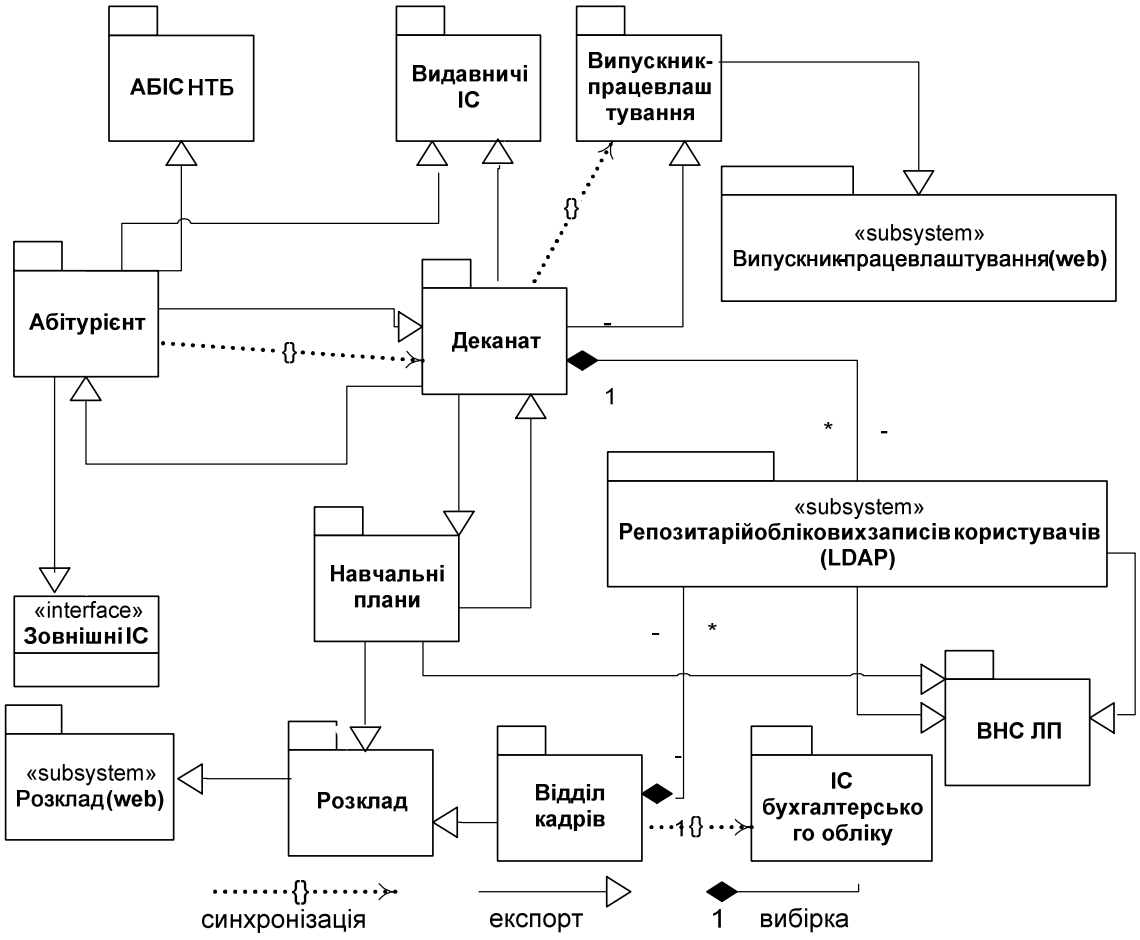


Рис. 1. Схема взаємодії основних БД ВНЗ

Зазвичай для тиражування використовуються три методи обміну даними між інформаційними продуктами (рис. 2):

- синхронізація – порівняння даних ;
- експорт – копіювання даних таблиці за певний період;
- вибірка – копіювання частин таблиці за певний період за параметрами користувача.

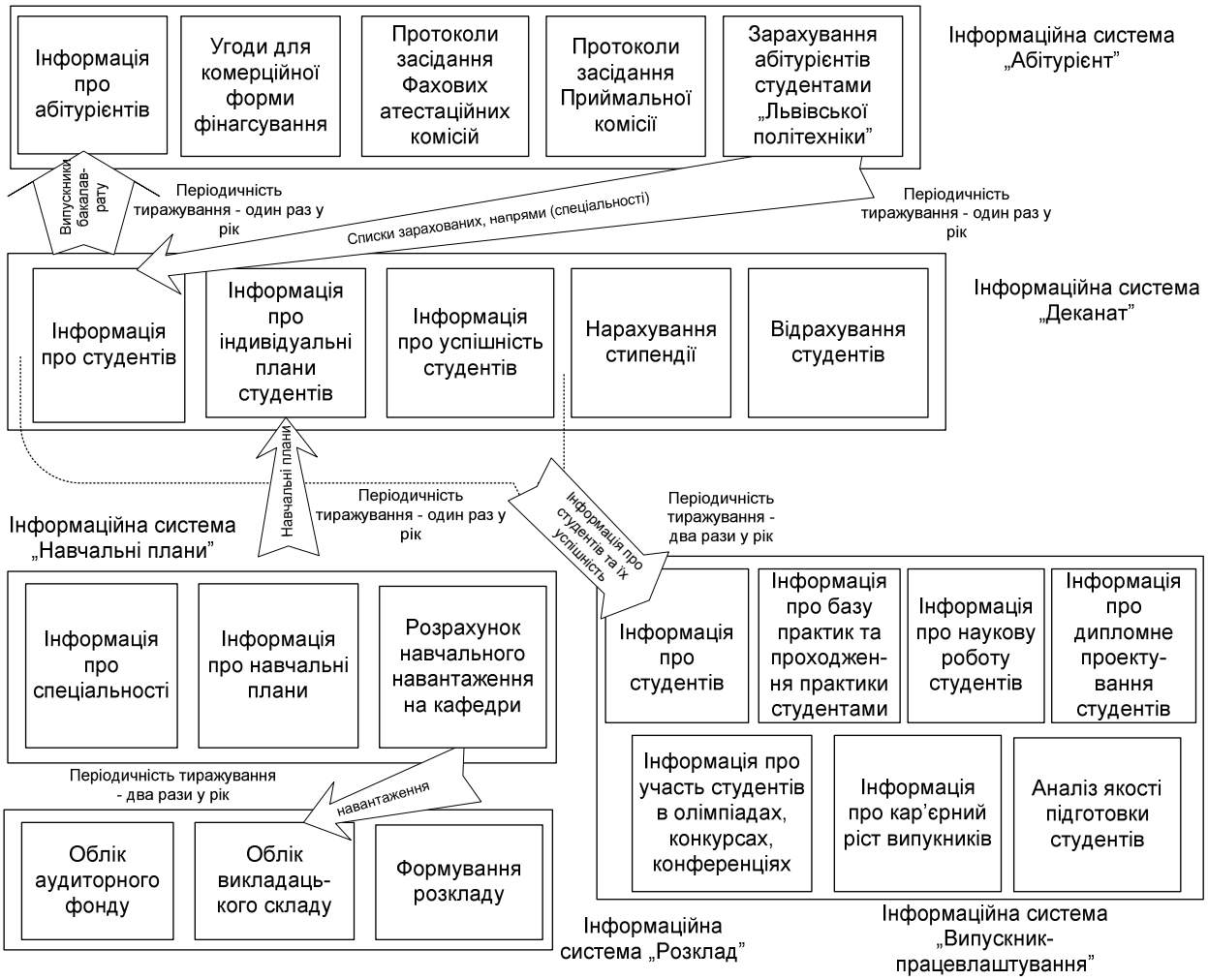


Рис. 2. Тиражування даних у інформаційних системах ВНЗ за допомогою переглядів

Методи обміну даними, показані на рис. 3, реалізуються такими засобами СУД MS SQL Server: DB Snapshot, перегляд, збережена процедура, розподілена транзакція, репліка.

Оскільки, як вже зазначалося, інформаційні продукти університету розроблялися ітераційно та на момент проектування сховища даних вже певний час функціонували, то для розроблення централізованого сховища даних використано підхід Інмона. Згідно з вимогами Інмона, сховище даних університету має виглядати так (рис. 3):

a. вітрини даних інститутів, які містять локальні копії систем «Деканат», «Випускник-працевлаштування» з первинними схемами даних (висока нормалізація) – інститути вносять детальну інформацію про свої потоки даних та використовують вітрини для оперативного аналізу даних;

b. центральне сховище університету, що містить інформацію з систем «Навчальні плани», «Розклад», «Абітурієнт», а також денормалізовану агреговану інформацію вітрин інститутів – інформація про навчальні плани та розклад тиражуються у вітрини даних, а агреговна інформація про успішність студентів консолідується з вітрин у центральне сховище.

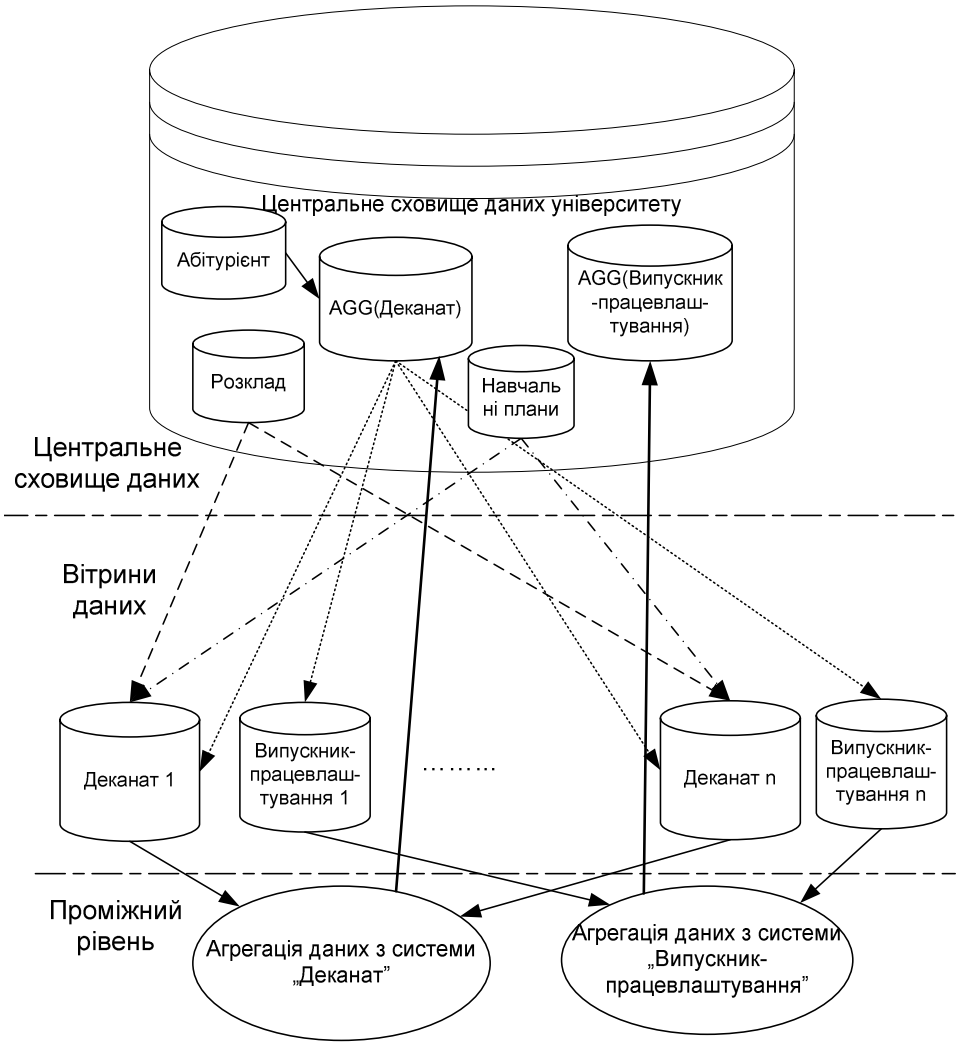


Рис. 3. Схема корпоративної фабрики сховища даних університету

5. Предметні області, для яких доцільно розробляти простори даних

Простори даних доцільно використовувати у різних галузях. Особливостями цих галузей є:

- ієрархія об’єктів;
- використання різних засобів опрацювання та аналізу даних;
- опрацювання даних з джерел, які наперед не були пристосовані для інтеграції;
- необхідність опрацювання потокових даних та даних, що надходять зі запізненням.

Спроектуємо простір даних для керування науковими даними.

Університети і дослідницькі інститути у всьому світі активно планують і реалізують архіви своєї наукової продукції. Крім того, Інтернет подає нові можливості для своєчасного поширення наукової інформації. Правильний вибір програмного забезпечення, яке б максимально повно задовольнило потреби навчальних і наукових організацій, за мінімуму прямих і непрямих витрат – один з аспектів розв’язання цієї задачі.

Простір даних обліку та супроводу наукових досліджень об'єднує інформацію про (рис. 4):

- інженерні та технічні розрахунки експериментів чи досліджень, подані у вигляді таблиць Excel, текстових файлів з розділювачами тощо;
- наукові статті та тези, описані у внутрішніх базах даних кафедр чи наукових підрозділів;
- програми, що використовуються для ведення розрахунків та експериментів;
- наукові звіти кафедр та наукових підрозділів про роботу працівників;
- літературні джерела (бібліотечні фонди), подані у вигляді внутрішніх баз даних чи загальноуніверситетського сховища даних, url-посилання, мультимедійна та графічна інформація, збережена на файл-сервері університету чи подана у глобальній мережі.

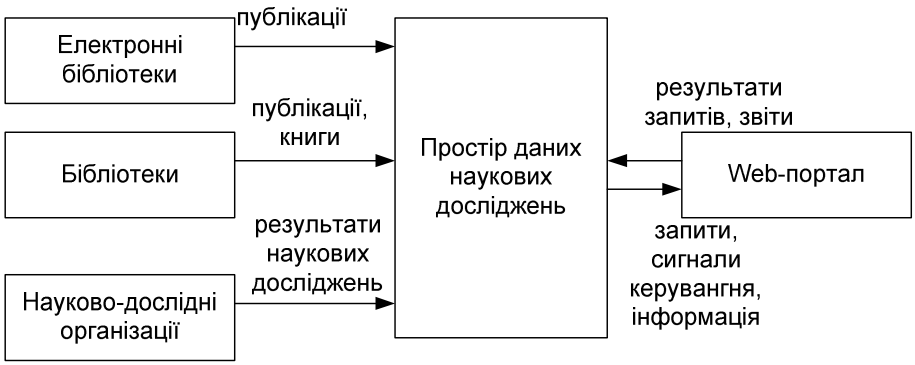


Рис. 4. Схема функціонування простору даних наукових досліджень

Простір даних стосовно аналізу наукових досліджень повинен забезпечувати:

1. Облік документів за результатами наукових досліджень – збереження структурованих даних про документи, авторів і зміст документів. Задача обліку полягає у розборі текстових даних і перетворенні їх на структуровану інформацію – опис документів, а також класифікація документів за галузями наукових досліджень. Також необхідним є накопичення такої інформації в базі знань.
2. Створення онтологій предметних областей – визначення сутностей предметної області і зв'язків між ними. До виконання цієї роботи залучаються експерти в предметних галузях. Результати оформляються у вигляді бази знань на основі онтологій.
3. Визначення змісту наукових досліджень. Розбір змістовного наповнення документа і його прив'язка до певної предметної області. Визначення термінів і понять, використаних в документі, і прив'язка до онтологій предметної області дослідження.
4. Пошук інформації в базі знань. Можливі два варіанти розвитку подій:
   a. пошук документів за запитом користувача;
   b. пошук документів, які пов'язані з документом або схожі на заданий документ.
5. Кількісний аналіз – визначення кількості дослідників (досліджень) за галузями і тематиками. Такий аналіз необхідний для визначення перспективності досліджень в певній галузі чи темі.
6. Інтеграція з іншими системами – використання електронних бібліотек і надання власного програмного інтерфейсу (API) для доступу до системи.

Простір даних наукових досліджень призначений для застосування, насамперед, у галузі науки та освіти, а також в промисловості для пошуку нових технологій і слідкування за їх розвитком з метою майбутнього використання. Впровадження пропонується в університетах та науково-дослідних інститутах, профільних міністерствах, Вищій атестаційній комісії. Також можливе незалежне застосування для інтеграції даних з різних організацій та установ.

Розроблювана програма має містити такі модулі (рис. 5):

1) база даних і знань;
2) підсистема інтеграції і збирання інформації;
3) підсистема аналізу;
4) Web-портал для керування системою і відображення результатів роботи системи.

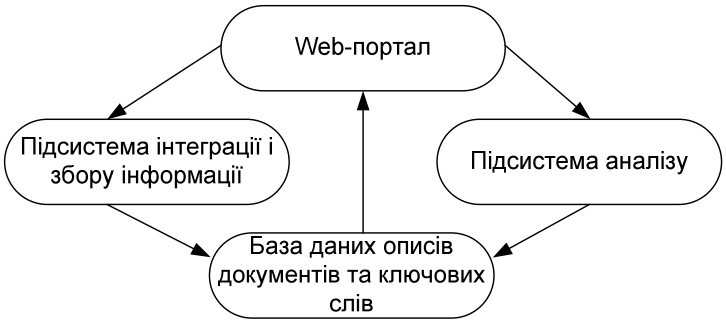


Рис. 5. Діаграма залежностей компонентів

База даних і знань призначена для накопичення структурованих даних і метаданих про наукові публікації. Вона є центральною частиною програми, оскільки надає інформацію іншим модулям Також в базі даних зберігаються метадані про документи – тобто перетворені в зрозумілий для програми формат з метою подальшого аналізу.

Оскільки система передбачає інтеграцію з іншими системами і джерелами документів про наукові дослідження, вона має мати функціональність, яка б давала змогу зчитувати файли в різних форматах і надавати свої дані для інших систем. За це відповідає Підсистема інтеграції і збору інформації. Через цю підсистему наповнюється база даних. Результатом роботи модуля є структурована інформація про документ. Підсистема аналізу аналізує документи на основі метаданих та реалізує кілька різних стратегій порівняння документів.

Першим етапом реалізації простору даних наукових досліджень є проектування бази даних (БД). Схему БД подано на рис. 6.

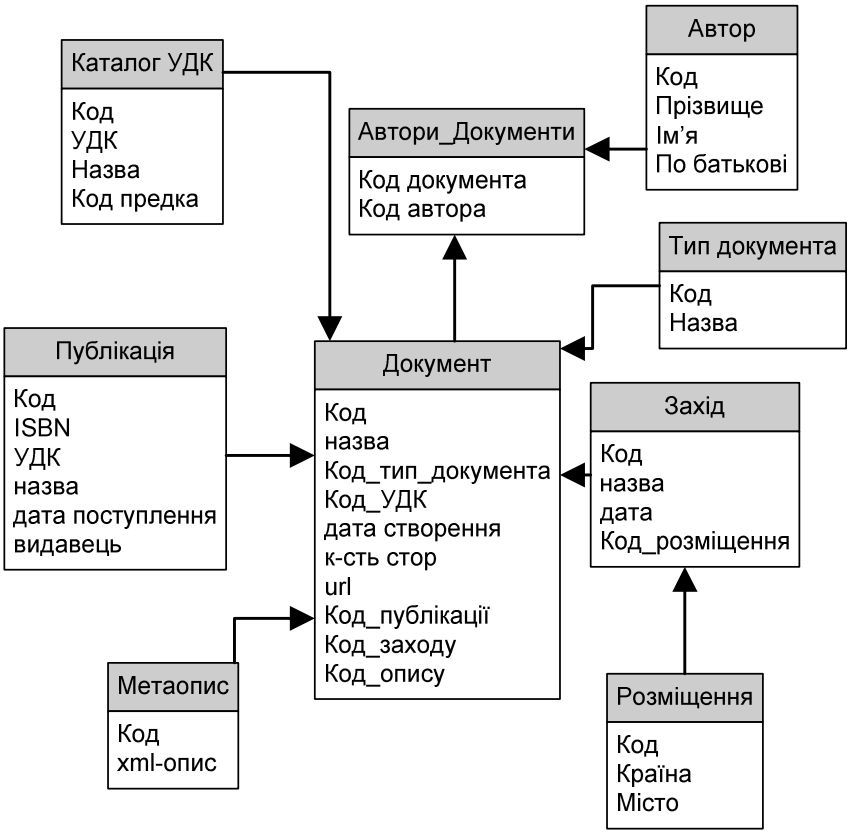


Рис. 6. Схема БД обліку наукових досліджень

База даних реалізована в СКБД SQL Server 2005, що дає змогу використовувати великий арсенал готових рішень для аналізу даних і текстів.

Додавання документа до системи відбувається через його розбір. Оскільки документи відрізняються як за форматом, так і за змістом, то необхідно передбачити можливість додавання розширення можливостей для розбору.

6. 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 йдеться про сховища даних, то структура джерел є відомою наперед, і основним завданням є очищення та завантаження самих даних.

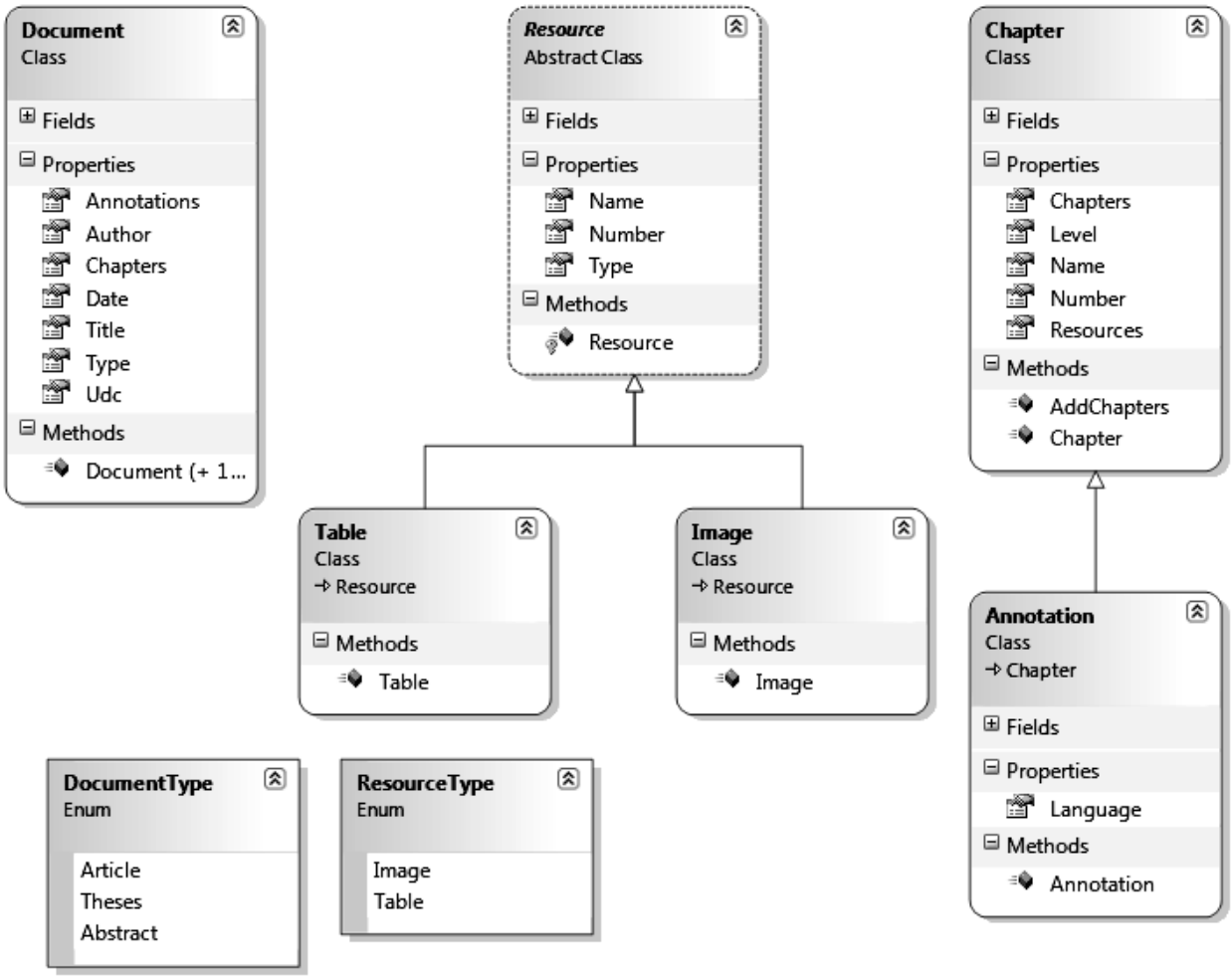


Рис. 7. Діаграма класів агрегату Document

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 необхідно передбачити розроблення сховища даних такої структури, щоб у нього можна було завантажити дані з джерел даних для забезпечення відповіді на запит користувача.

6.1. Технології та засоби реалізації сховищ даних

Порівняльну характеристику засобів реалізації сховищ даних подано в таблиці.

Порівняльна характеристика засобів реалізації сховищ даних

Засіб	Переваги	Недоліки
1	2	3
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. Містить: 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 і Java-застосування і аплети)	ПЗ проміжного рівня

Продовження таблиці

1	2	3
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 дає змогу аналізувати текстові дані та записувати у сховище даних; функціонує як на внутрішньокорпоративиному, так частково і на міжкорпоративному рівнях	Функціонує за принципом сповіщень, що робить якість інтеграції сильно залежною від користувача
Informix MetaCube	Один з продуктів (INFORMIX-MetaCube for Excel) дає змогу здійснювати багатомірний аналіз дуже великих даних безпосередньо з Excel-таблиць. INFORMIX-MetaCube Agents дає змогу опрацьовувати запити у фоновому режимі.	Розрахований на розроблення сховищ даних, але засоби інтеграції поступаються аналогічним в SQL Server чи Oracle
Visual Warehouse (IBM)	До сімейства програмних продуктів IBM входить проміжний сервер DataJoiner, який дає змогу інтегрувати дані з реляційних БД та завантажувати дані з текстових файлів	Обмежений перелік підтримуваних операційних систем – для версії 5.2 від квітня 2009 р.
Netezza	Архітектура з масовим паралелізмом без поділу ресурсів (на рівні зберігання) і симетрична багатопроцесорна архітектура (на рівні хоста), прискорення опрацювання даних під час передавання до сховища	Закрита технологія, масштабованість до сотень терабайтів, недостатньо розроблені засоби інтеграції
Teradata	За рахунок технології масового паралелізму сховище Teradata можна масштабувати до декількох петабайтів	Закрита архітектура, не підтримує SaaS

Засоби інтеграції можна розділити на два умовні класи: інтеграція застосувань та інтеграція веб-застосувань. Засоби інтеграції застосувань реалізуються за допомогою проміжного шару, спеціалізованих засобів та серверів інтеграції. Засоби інтеграції застосувань розроблено в Oracle, SQL Server. Узагальнену схему інтеграції даних подано на рис. 8.

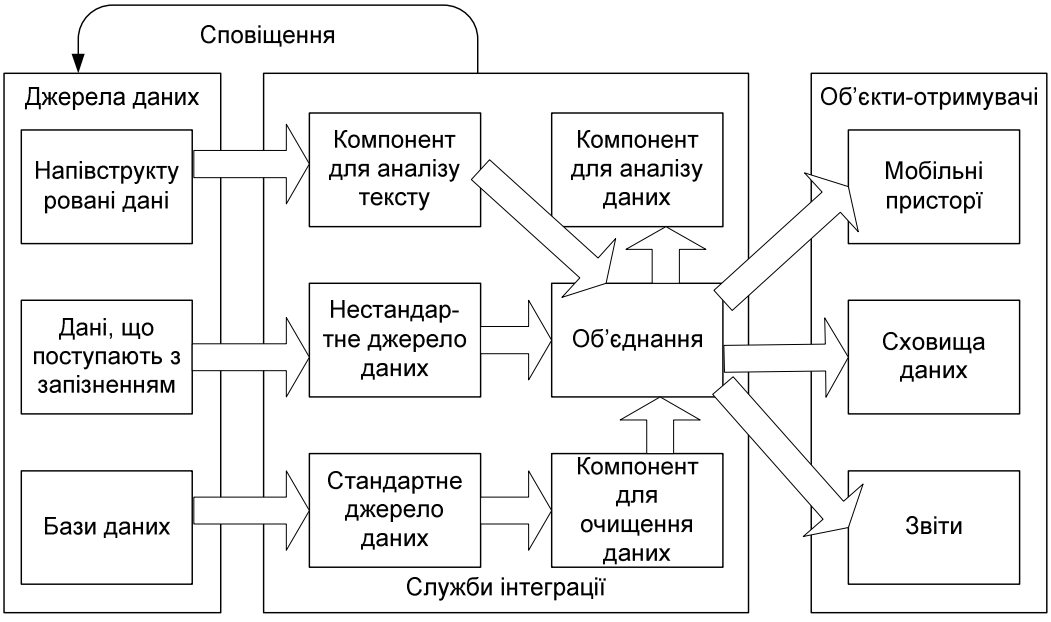


Рис. 8. Узагальнена схема інтеграції даних

Для інтеграції геоінформації та баз даних використовують спеціалізовані засоби, які перетворюють векторні дані на 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.

6.2. Технології та засоби реалізації просторів даних

Якщо розглядати технології, які дають змогу реалізувати можливості просторів даних, то насамперед 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 низку переваг перед web-сервісами. Серед них:

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

– можливість здійснювати аналіз як в глобальних, так і в корпоративних мережах;

– можливість продовжувати роботу сервісів-агентів пошуку та збирання даних і після виконання конкретного запиту;

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

Сервіс-орієнтірований підхід до слабкозв'язаних масивів даних як до простору даних дає змогу 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 СКБД для просторів даних для того, щоби з нових позицій вирішувати проблему інтеграції федеральних гетерогенних інформаційних ресурсів.

Хмарні обчислення— технологія опрацювання даних, за якою 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.

Сьогодні є такі *реалізації хмарних обчислень*, пов’язані з просторами даних.

a. 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.

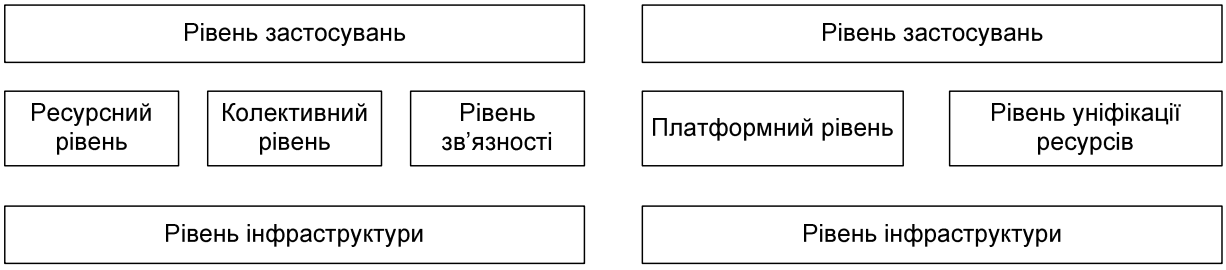
b. Google App Engine – це платформа, яка дає змогу використовувати одну інфраструктуру для створення й хостинга своїх застосувань. Використовує нереляційне розподілене сховище даних.

Порівняємо обидві технології.

У хмарі, як і в грід, акумулюються потужності, щоб одержати економічніше, масштабованіше рішення і позбавити користувача необхідності працювати винятково на власному апаратному та програмному забезпеченні. Проте за десять років з моменту появи ідеї грід принципово змінилися обсяги опрацьованих даних, з'явилися віртуалізаційні рішення.

Архітектури грід і хмар відрізняються, оскільки вони створювалися за різних передумов. На першу вплинуло прагнення ефективніше використовувати дорогі розподілені обчислювальні ресурси, зробити їх динамічними й однорідними. Тому архітектура побудована на інтеграції вже існуючих ресурсів, враховуючи апаратне та програмне забезпечення, операційні системи, локальні засоби, що забезпечують керування й безпеку. У результаті створюється «віртуальна організація», ресурси якої, переведені в логічну форму, можуть використовуватись членами тільки цієї організації. Існування цієї організації підтримується п'ятьма рівнями протоколів, інструментами й сервісами, побудованими поверх них (рис. 9, а). Нижнім є інфраструктурний рівень, що поєднує

комп'ютери, системи зберігання, мережі, репозиторії кодів. Вище нього розташований рівень зв'язності, на якому визначені комунікаційні протоколи й протоколи аутентифікації. Ресурсний рівень забезпечує надання ресурсів, можливості керування ними, поділ між окремими користувачами й оплату. Колективний рівень доповнює ресурсний, даючи змогу оперувати наборами ресурсів. Рівень додатків слугує для підтримки роботи додатків.



а *б*

Рис. 9. Порівняння архітектури грід та хмарного обчислення

Архітектура хмар відкрита для доступу через Мережу, а не тільки в рамках грід. Звертаються до пулів обчислювальних ресурсів і систем зберігання даних за стандартними протоколами, наприклад таким, як WSDL і SOAP, або за допомогою технологій Web 2.0 (REST, RSS, AJAX), а також через існуючі технології grid. Протоколи хмар можна розділити на чотири рівні (рис. 9 б). Інфраструктурний рівень містить «сирі» комп'ютерні ресурси (сервери, системи зберігання, мережі). Рівень уніфікації ресурсів містить ті самі ресурси, але в абстрагованому виді – їх можна подати користувачам верхнього рівня як віртуалізовані сервери, кластери серверів, файлові системи й СУБД. Рівень платформ додає набір спеціалізованих інструментів, що зв'язує ПЗ й сервіси поверх універсальних ресурсів, утворюючи середовище для розроблення й впровадження додатків.

Висновки

Розглянуто аналіз таких засобів побудови систем прийняття рішень, як сховища та простори даних, а також загальний аналіз галузі комп'ютингу.

Наукова новизна статті полягає у встановленні 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.

1. Lenzerini M. Data Integration: A Theoretical Perspective // PODS 2002. pp. 233–246. [Електронний ресурс].[Режимдоступу] - http://www.dis.uniroma1.it/~lenzerin/homepagine/talks/TutorialPODS02.pdf. 2. Papakonstantinou Y. Object Exchange Across Heterogeneous Information Sources/ Papakonstantinou Y., Garcia-Molina H., Widom J. // IEEE International Conference on Data Engineering, Taipei, Taiwan, 2005. – P. 251–260. 3. Halevy A. Answering queries using views: A survey / Halevy A. // The VLDB Journal. – 2001. – № 10. – P. 270–294. 4. Kalinichenko L.A. Extensible ontological modeling framework for subject mediation / Kalinichenko L.A., Skvortsov N.A. // Proc. of the Fourth Russian Conference on Digital Libraries (RCDL'2002). – Dubna: JINR, 2002. – V. 1. – P. 99–119. 5. Калиниченко Л.А. Методы и средства интеграции неоднородных баз данных / Л.А. Калиниченко. – М.: Наука, 1983. – 420 с. 6. Основные концепции и подходы при создании контекстно-поисковых систем на основе реляционных баз данных. – [Електронний ресурс]. – [Режим доступу] http://www.citforum.ru/database/articles/search_sys.shtml. 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. Селиверстова А.В. Совершенствование механизма отбора информации для принятия управленческих решений: Автореф. – Челябинск, 2002. – 20 с. 9. Тузовский А.Ф. Системы управления знаниями (методы и технологии) / Под общ. ред. В.З. Ямпольского. – Томск: Изд-во НТЛ, 2005. – 260 с.