УДК 681.3:658.115.6

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

ДЕЯКІ ЗАВДАННЯ ПРОСТОРІВ ДАНИХ ЯК АБСТРАКЦІЇ КЕРУВАННЯ ІНФОРМАЦІЄЮ

Розглянуто поняття просторів даних як нової абстракції керування інформацією та деякі їхні завдання. Визначено особливі властивості систем і компоненти побудови просторів даних, охарактеризовано зв'язки між їх учасниками.

There has been considered a concept of the data spaces as a new abstraction of controlling information and some their assignments. Determined the special properties of systems and components of building the data spaces, characterized the connections amongst their participants.

І бази даних, і сховища даних дозволяють опрацьовувати деталізовані та інтегровані дані, що побудовані на основі наперед допустимих моделей даних. В сучасних сценаріях управління даними все рідше трапляються випадки, коли всі дані можуть знаходитися під керуванням традиційної реляційної СКБД або будь-якої іншої моделі даних або системи. Замість цього розробники часто стикаються з набором слабо зв'язаних джерел даних і тому змушені кожного разу вирішувати повторювані низькорівневі завдання управління даними в різнорідних колекціях. В число цих завдань входять забезпечення можливостей пошуку і запиту даних; дотримання правил, обмежень цілісності, угод про іменування і т.д.; відстежування походження даних; забезпечення доступності, відновлення і контролю доступу; керований розвиток даних і метаданих.

Поняття простору даних вводиться [1] як нова абстракція керування даними в таких сценаріях. В якості ключової програми робіт в області керування даними пропонується проектування і розробка платформ підтримки просторів даних (DataSpace Support Platforms, DSSP). DSSP забезпечує набір взаємозв'язаних послуг і гарантує розробникам можливість концентруватися на специфічних проблемах їх додатків, а не на повторюваних завданнях, що виникають при потребі узгодженої та ефективної роботи з взаємозв'язаними, але роздільно керованими даними.

На рис. 1 показана класифікація існуючих рішень управління даними по двох вимірах. Вимір "Administrative Proximity" ("адміністративна близькість") показує, наскільки близькі різні джерела даних з точки зору адміністративного управління. "Near" ("близько") означає, що джерела знаходяться під єдиним або, принаймні, координованим управлінням, а "Far" (далеко) показує більш слабку координацію і навіть, можливо, повну відсутність координації. Чим ближче адміністративне управління групи джерел даних, тим сильніші гарантії (наприклад, узгодженість, стабільність), які можуть бути надані системою керування даними.

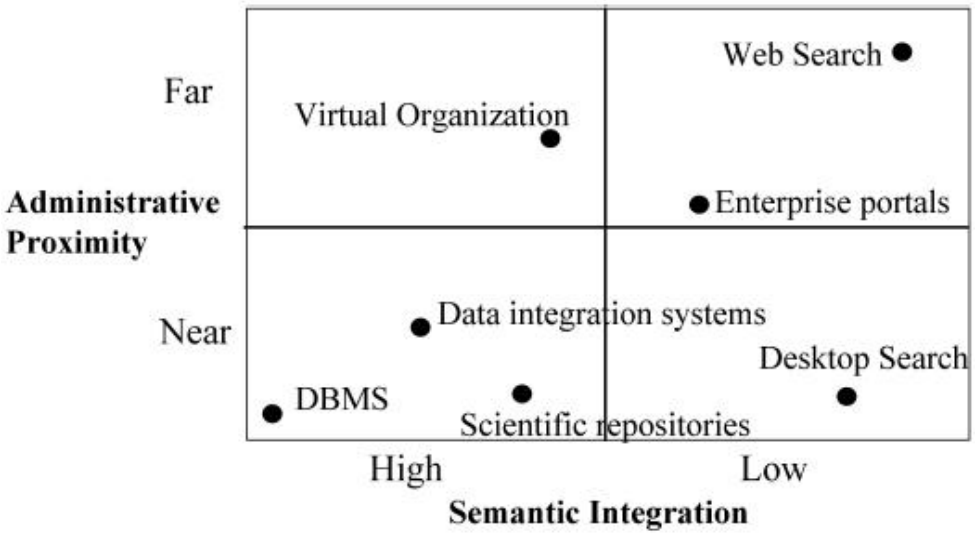

Рис. 1. Простір рішень управління даними

Вимір "Semantic Integration" ("семантична інтеграція") є мірою того, наскільки близько можуть бути зіставлені схеми різних джерел даних. Тобто наскільки добре відповідають типи, імена, одиниці виміру, вміст і т.д. даних в джерелах. На дальньому кінці ("low") інформація про схеми взагалі відсутня. У проміжку між "high" і "low" розміщуються різні рішення і підходи інтеграції даних, засновані на напівструктурованих даних і контрольованих словниках. Це вимір показує рівень, на якому можуть бути забезпечені семантично розвинені засоби запитів даних і маніпулювання даними над групою джерел даних, причому більш високий рівень інтеграції забезпечує більш розвинуті функціональні можливості.

Традиційні СКБД представляють тільки одну точку (хоч і дуже важливу) в просторі рішень управління даними. СКБД вимагають, щоб всі дані знаходилися під єдиним адміністративним керуванням і відповідали єдиній схемі. У відповідь на задоволення цих обмежень СКБД можуть забезпечити розвинені засоби маніпулювання даними та опрацювання запитів зі зрозумілою і строгою семантикою, а також строгі транзакційні гарантії оновлень, паралельного доступу і довготривалого зберігання (так звані властивості "ACID"). Важливою точкою на рис. 1 є "системи інтеграції даних". Насправді, системи інтеграції даних і обміну даними традиційно призначаються для підтримки багатьох інших осмислених служб в системах просторів даних. Особливість полягає в тому, що в системах інтеграції даних потрібна семантична інтеграція до того, як можуть бути забезпечені які-небудь інші послуги. Тому, хоч і відсутня єдина схема, якій відповідають всі дані, система повинна знати точні взаємозв'язки між елементами, що використовуються в кожній схемі. В результаті для створення системи інтеграції даних потрібна значна попередня робота.

Простори даних не є підходом до інтеграції даних; швидше, це підхід співіснування даних. Мета підтримки простору даних полягає в забезпеченні базового набору функцій над усіма джерелами даних, а не в їх інтеграції. Наприклад, DSSP може забезпечити над усіма своїми джерелами даних пошук за ключовими словами, аналогічно тому, як забезпечують існуючі пошукові системи в десктопах.

Аналогічна гнучкість є і у вимірі адміністративної близькості рис. 1. Якщо бажано наявність адміністративної автономії, то DSSP не зможе гарантувати узгодженість, стійкість результатів операцій оновлення і т.д. Для задоволення потреби в більш строгих гарантіях потрібні додаткові зусилля для досягнення угод між власниками джерел даних і відкриття деяких інтерфейсів (наприклад, для протоколів фіксації транзакцій).

Таким чином, особливими властивостями систем просторів даних є наступні:

1. DSSP повинні працювати з даними і додатками в різноманітних форматах, доступних для багатьох систем через різні інтерфейси. Від DSSP потрібна підтримка всіх даних простору даних, без будь-яких винятків (як це буває при використанні СКБД).

2. Хоча DSSP забезпечує засоби інтегрованого пошуку, запиту, оновлення та адміністрування просторів даних, ті ж самі дані часто можуть бути доступні для читання і оновлення через власний інтерфейс системи, який безпосередньо управляє даними. Тому, на відміну від СКБД, DSSP не має повного контролю над своїми даними.

3. Можуть забезпечуватися різні рівні послуг з обробки запитів до DSSP, і в деяких випадках вони можуть повертати найкращі з можливих наближених відповідей. Наприклад, якщо деякі джерела даних стають недоступними, DSSP може забезпечити найкращий з можливих результатів на основі даних, доступних під час виконання запиту.

4. DSSP повинні підтримувати засоби для забезпечення більш тісної інтеграції даних простору, якщо це стає необхідно.

На даний час дослідження з управління даними відбуваються все активніше та енергійніше. Розглянемо два сценарії просторів даних.

Управління персональною інформацією: Мета управління персональною інформацією (Personal Information Management, PIM) полягає в забезпеченні простого доступу та маніпулювання всією інформацією на персональному комп'ютері з можливими розширеннями до мобільних пристроїв, персональної інформації в Web і навіть всієї інформації, накопиченої протягом життя людини. Пошукові засоби, доступні на десктопах в даний час, представляють важливий перший крок для PIM, але вони обмежуються запитами на основі ключових слів. Наші десктопи зазвичай містять деякі структуровані дані (наприклад, електронні таблиці), і між різними елементами десктопу є важливі асоціації. Тому на наступному кроці розвитку PIM користувачу повинно бути дозволено робити пошук в десктопі більш осмисленим чином.

Управління науковими даними: Для наукового-дослідницької групи, що працює в області спостережень за навколишнім середовищем і передбачень його поведінки, може здійснюватися моніторинг прибережної екосистеми з використанням метеостанцій, датчиків, встановлених на берегових стійках і буях, і віддалених пристроїв одержання зображень. Крім того, можуть використовуватись атмосферні і гідродинамічні моделі, що імітують минулі, поточні і майбутні умови. Для обчислень можуть бути потрібні дані і модельні результати від інших груп, що забезпечують прогнози річкових стоків і океанічної циркуляції. Спостереження та моделювання забезпечують вхідні дані для програм, що генерують широкий діапазон продуктів даних для користування цією групою та іншими групами: діаграми порівняння спостерігалися і в модельних даних, зображеннях розподілів поверхневої температури, анімації надходження води в устя річок та інше. Така група легко накопичить мільйони зразків даних протягом короткого часу.

Простір даних повинен містити всю інформацію, доречну для конкретної організації, незважаючи на формат і місце розташування цієї інформації, а також моделювати розвинений набір зв'язків між репозиторіями даних. Отже, ми моделюємо простір даних як набір учасників та зв'язків.

Учасниками простору даних є індивідуальні джерела даних: вони можуть бути реляційними базами даних, репозиторіями XML, текстовими базами даних, Web-сервісами та пакетами програмного забезпечення. Вони можуть зберігатися або бути потоками даних (локально керованими системами потоків даних) або навіть сенсорними установками. Деякі учасники можуть підтримувати виразні мови запитів, а інші – бути неінтелектуальними і підтримувати лише обмежені інтерфейси для формулювання запитів (наприклад, структуровані файли, Web-сервіси або інші пакети). Учасники можуть бути дуже структурованими (наприклад,

реляційними базами даних), напівструктурованими (XML, колекції коду) або повністю неструктурованими. Деякі джерела будуть підтримувати традиційні операції оновлення, інші – допускати тільки доповнення (з метою архівації), а треті можуть бути повністю незмінними.

Простір даних повинен вміти моделювати будь-який вид зв'язку між двома (або декількома) учасниками. У більш традиційному варіанті ми повинні вміти моделювати ситуації, коли один учасник є уявленням або реплікою іншого учасника, або відображати одну на іншу схему двох учасників. Простори даних можуть вкладатися один в один (наприклад, простір даних факультету вкладається в простір даних університету), і вони можуть перекриватися (наприклад, простір даних одного факультету може розділяти деяких учасників з іншим факультетом). Тому в просторі даних повинні міститися правила розмежування доступу.

Однією з основних служб простору даних є каталогізація елементів даних від учасників. Каталог – це реєстр ресурсів даних, що містить найбільш базову інформацію про кожного з них: джерело, ім'я, місцеположення в джерелі, розмір, дата створення, власник і т.д. Каталог є інфраструктурою для більшості інших сервісів простору даних, але він також може підтримувати базовий призначений для користувача інтерфейс перегляду простору даних.

Іншими основними службами, які будуть підтримуватися в DSSP, є пошук і запит даних. В той час як СКБД відрізняються поліпшеною підтримкою запитів, пошук є основним механізмом роботи кінцевих користувачів з великими колекціями невідомих даних. Пошук менш вимогливий, ніж запит даних, оскільки він заснований на схожості, наданні кінцевим користувачам ранжированих результатів та підтримці інтерактивного вдосконалення, так що користувачі можуть досліджувати набір даних і інкрементно покращувати свої результати. DSSP повинні дозволяти користувачам задавати запит і ітераційно його вдосконалювати, якщо це доречно, до виду запиту в стилі бази даних. Ключовий принцип просторів даних полягає в тому, що пошук повинен бути застосовний до всього вмісту простору даних, незалежно від форматів даних.

Універсальні можливості пошуку і запитів повинні поширюватися не тільки на дані, але й на метадані. У користувачів повинні бути можливості знаходження необхідних джерел даних і отримання інформації про їх складності, коректності та актуальності. DSSP будуть також підтримувати і оновлення даних. Очевидно, що ефекти оновлень будуть визначатися рівнем мінливості відповідних джерел даних. Однією з основних дослідницьких проблем просторів даних є розробка та забезпечення гарантованої семантики оновлень в різнорідному середовищі з високим рівнем автономності компонентів. Інші ключові сервіси DSSP включають моніторинг, виявлення подій і підтримку складних потоків робіт. Наприклад, ми можемо захотіти зробити обчислення при надходженні нової частини даних та поширити результати цього обчислення в набір прийомних джерел даних. Аналогічно, в DSSP повинні підтримуватися різні форми аналізу даних.

Не кожен учасник простору даних буде обов'язково забезпечувати інтерфейси, необхідні для підтримки всіх функцій DSSP. Тому з'явиться потреба в різних розширеннях джерел даних. Джерело не обов'язково буде зберігати свої власні метадані, тому для таких джерел буде потрібний незалежний репозиторій метаданих. На відміну від СКБД, в DSSP не передбачається наявність повного контролю над даними у просторі даних. Замість цього, DSSP дозволяє керувати даними системам-учасникам, але забезпечує новий набір служб поверх всіх цих систем, дотримуючись їх потреби в автономності. Крім того, у нас може бути декілька DSSP, які обслуговують один і той ж простір даних в деякому розумінні, у DSSP може бути своє власне уявлення про конкретний простір даних.

Моделювання даних і базові можливості запитів: На відміну від СКБД, в ядрі DSSP потрібна підтримка декількох моделей даних, щоб природним чином підтримувалося якомога більше типів учасників.

Моделі даних, які підтримує DSSP, будуть утворювати ієрархію відповідно до їх потужності. Кожен учасник простору даних підтримує певну модель даних і деяку мову запитів, що відповідає цій моделі. Наприклад, на самому верхньому рівні ієрархії (найбільш загальному) знаходяться колекції іменованих ресурсів, можливо, з базовими властивостями – розмір, дата створення та тип (наприклад, зображення JPEG, база даних MySQL). "Запит" до такої моделі даних відповідає тому, що зазвичай підтримується в файлових системах по відношенню до їх директорій: зіставлення імен, пошук в діапазоні дат, сортування за розміром файлу і т. д. На наступному рівні DSSP повинні підтримувати модель даних мультимножини слів, з чого випливає, що має бути можливість формулювання запитів за ключовими словами для будь-якого учасника простору даних і, отже, можливість перегляду вмісту учасників простору даних.

Нижче рівня моделі мультимножини слів в ієрархії може розташовуватися модель напівструктурованих даних, заснована на позначених графах. Якщо учасник підтримує деяку структуру, ми повинні мати можливість формулювання і простих запитів, і більш складних запитів, заснованих на моделі напівструктурованих даних. В ієрархії будуть присутні й інші моделі даних: реляційна модель, XML зі схемою, RDF, OWL (Web Ontology Language). При наявності деякого середовища ключова проблема полягає в знаходженні методів інтерпретації запитів на різних мовах на учасниках, що підтримують деякі моделі. Більш точно, проблема полягає в переформулюванні запиту, представленого на складній мові, для джерела, що підтримує більш слабку модель даних, і навпаки, переформулювання запиту, представленого простою мовою, для джерела, що підтримує більш виразні модель даних і мову запитів (наприклад, запит за ключовими словами до реляційної бази даних).

Відповідальним компонентом побудови простору даних є розкриття його учасників і зв'язків між ними. Дуже поширена проблема сьогоднішніх великих підприємств полягає в тому, що вони навіть не знають, які джерела даних є в організації. Остаточною метою розкриття простору даних є виявлення учасників простору

даних, створення зв'язків між ними та підвищення точності існуючих зв'язків між учасниками. Основними компонентами системи розкриття простору даних є (1) виявлення учасників в організації; (2) напівавтоматичний засіб для кластеризації і знаходження зв'язків між учасниками і (3) засіб для створення більш точних зв'язків між учасниками (в межах відображень схем).

Висновки. Найбільш гострі проблеми управління інформацією в організаціях сьогодні залежать від наявності в організаціях багатьох різнотипних, але часто взаємозалежних джерел даних. Ідея просторів даних і розробки платформ підтримки просторів даних (DataSpace Support Platforms, DSSP) є засобом вирішення цих проблем. Призначенням DSSP є звільнення розробників від потреби постійної повторної реалізації основних функцій управління даними при роботі зі складними, різнорідними, взаємопов'язаними джерелами даних, подібно до того, як традиційні СКБД забезпечили аналогічні можливості для роботи зі структурованими реляційними базами даних. Однак, на відміну від СКБД, в DSSP не передбачається наявність повного контролю над даними у просторі даних. Замість цього, DSSP дозволяє керувати даними системам-учасникам, але забезпечує новий набір служб над усіма системами, задовольняючи їх вимоги автономності.

Список використаних джерел

1. Michael Franklin. From Databases to Dataspaces: A New Abstraction for Information Managementcord / Michael Franklin, Alon Halevy, David Maier // SIGMOD Record. – 2005. – Vol. 34. – № 4, Dec.

2. Кузнецов С. От баз данных к пространствам данных: новая абстракция управления информацией [Електронний ресурс] / Кузнецов С. – 2006. – Режим доступу : http://www.citforum.ru/database/articles/from_db_to_ds.

3. Черняк Л. Машины для обработки событий / Л. Черняк // Открытые системы. – 2006. – № 9.