УДК 004.4:004.6

Безверхий А. І., Бельков О. С., Кулик І. В.

ТЕХНОЛОГІЯ СХОВИЩА ДАНИХ ТА ЇЇ АНАЛІТИЧНІ ДОДАТКИ

На сучасному етапі розвитку обчислювальні системи та комп'ютерні мережі дозволяють накопичувати великі масиви даних для вирішення задач обробки та аналізу. Великий об'єм інформації, з одного боку, дозволяє отримати більш точні результати аналізу, з іншого – перетворює пошук рішень у складне завдання [1–2]. Таким чином, аналіз накопиченої упродовж декількох років інформації допоміг би співробітникам вищого навчального закладу, насамперед викладачам та працівникам деканатів, у своєчасному прийнятті вірних рішень, щодо подальшої долі студентів [3]. Наприклад, в разі необхідності додатково мотивувати його, щоби заздалегідь попередити проблеми з навчанням.

Мета роботи:

– спроектувати та реалізувати систему сховища даних та засоби для ефективної взаємодії з нею;

– підготувати дані, на основі яких проводитиметься аналіз;

– побудувати аналітичні моделі та знайти закономірності у навчанні студентів.

В основі досліджень даної роботи лежать дані, накопичені під час роботи приймальної комісії ЗДІА з 2003 по 2008 роки, та дані з деканатів, що накопичувалися під час навчання студентів з моменту впровадження модульно-рейтингової системи. Під час досліджень були задіяні наступні середовища:

– Microsoft SQL Server 2005 Standard Edition;

– Microsoft Visual Studio 2005;

– Microsoft Business Intelligence.

Вибір вказаних середовищ обумовлений тим, що вони мають інструменти інтеграційної взаємодії, що забезпечує зручність та ефективність розробки власних програмних проектів [4].

Система сховища даних.

Виходячи з вимог до системи, що розглядається, та можливостей, які вона в змозі забезпечити, було обрано архітектуру системи сховища даних з єдиним вимірним інформаційним складом даних (ВІСД) (Single Dimensional Data Store) [2] (див. рис. 1).

ETL-пакет видобуває дані з різних систем-джерел та розташовує їх у буфері. Буфер є необхідним у тому випадку, коли перетворення даних є складним (не може виконуватися оперативно у пам'яті), коли об'єм даних є надто великим (через нестачу місця в оперативній пам'яті), або коли дані видобуваються за допомогою декількох ETL-пакетів. Фізично у ролі буфера може виступати база даних або файли.

В даній системі одна ETL-система видобуває дані із систем-джерел та розташовує їх у базі даних або записує їх до файлів. Друга ETL-система, оперуючи над даними з декількох джерельних систем у буфері, об'єднує їх, застосовує до них бізнес-правила та завантажує вже очищені дані до ВІСД.

Застосунки, що взаємодіють зі сховищем даних, зчитують інформацію з ВІСД та підготовлюють її до взаємодії з користувачами. Дані можуть також бути завантажені до багатовимірних баз даних та можуть використовуватися користувачами через OLAP та Data Mining застосунки. Перевага даної архітектури полягає в тому, що вона є найбільш простою, бо дані з буферу завантажуються безпосередньо до ВІСД без будь-якого проміжного нормалізованого зберігання. Головним же її недоліком є складність створення додаткових інформаційних складів.

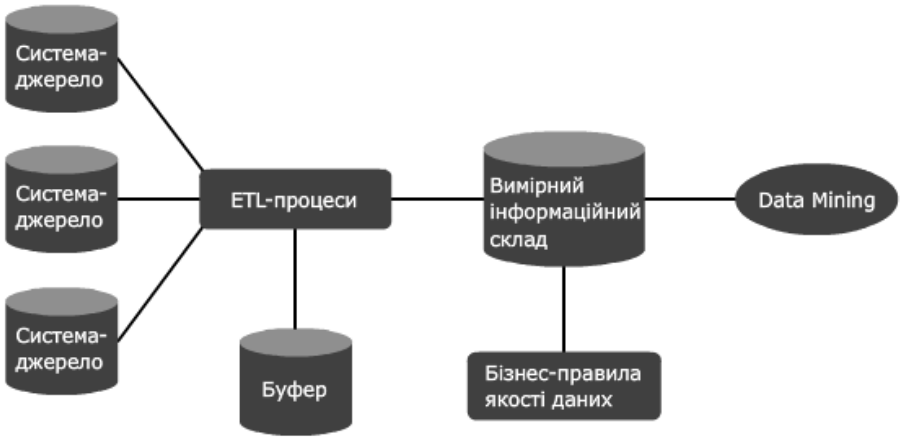

Рис. 1. Архітектура системи сховища даних

Обрання цієї архітектури є найбільш оптимальним через те, що джерельні системи мають невелику кількість корисних для аналізу атрибутів сутностей і одного ВІСД цілком вистачатиме для оперування над даними.

Дослідження ефективності використання інструментів Integration Services

Проведемо порівняльний аналіз послідовного використання SQL-скриптів та засобів (компонентів) SSIS для маніпуляції даними.

Дослідження (див. табл. 1, рис. 2) проводилися на інформаційно-обчислювальній машині з наступним апаратним та програмним забезпеченням:

- процесор: Intel Core 2 Duo T5750 2.00 GHz;
- оперативна пам’ять: 2 GB;
- операційна система: Windows XP Professional SP2.

Таблиця 1

Усереднені дані по виконанню операцій різними методами

№ п/п	Назва операції	SQL-скрипти, (хв. : сек.)	Засоби SSIS, (хв. : сек.)
1	Повне виконання пакету операцій	01 : 24,6	01 : 14,5
2	Заповнення буферу	00 : 11,5	00 : 02,8
3	Заповнення ВІСД та очищення буферу	00 : 14,7	00 : 05,7

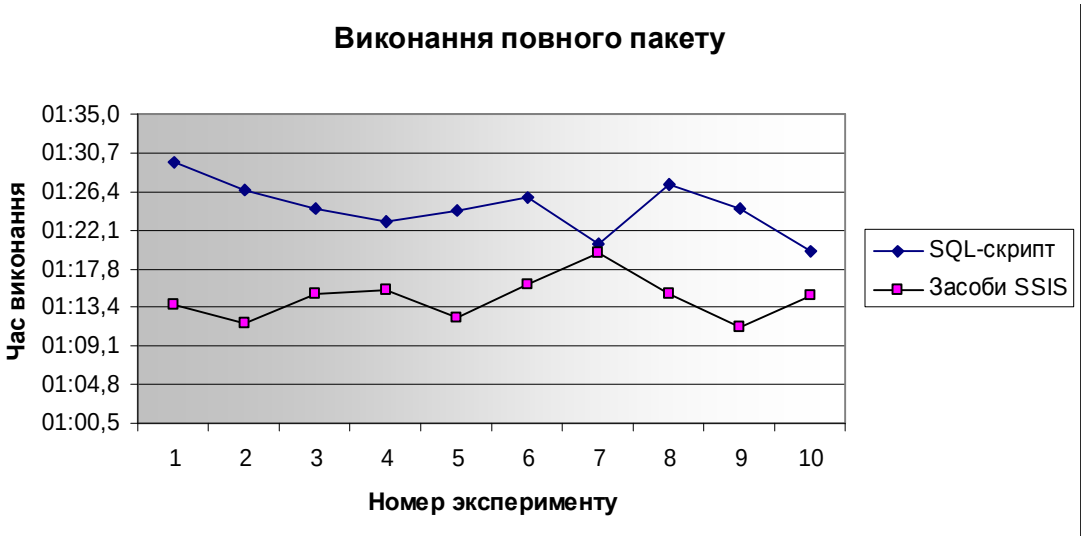

Рис. 2. Виконання повного пакету

Висновки по дослідженню Integration Services представлені у табл. 2.

Таблиця 2

Переваги та недоліки підходів до транспортування даних

	Послідовного використання SQL-скриптів	Засоби SSIS
Переваги	1. Контроль виконання ETL-завдань забезпечується засобами SQL, що є зручнішими для відладки. 2. Зручність у додаванні ETL-завдання для завантаження даних з нової бази.	1. Швидке виконання завдяки паралельному виконанню процесів. 2. Візуальні засоби розробки та модифікації компонентів. 3. Вбудовані засоби перевірки, що попереджують, наприклад, помилки пов'язані з конвертуванням типів даних.
Недоліки	1. Програш у швидкості через послідовність виконання ETL-завдань. 2. При проектуванні моделі транспортування необхідно ретельно перевірити поля таблиці-джерела та цільової таблиці аби запобігти втраті інформації.	1. Додавання ETL-завдання для завантаження нової інформації веде за собою модифікацію певної кількості компонентів. 2. Велика кількість компонентів сприяє ускладненню знаходження помилок при проектуванні пакетів.

Дослідження закономірностей впливу на навчання студентів

Модель «Стан навчання. Загальні закономірності».

Метою побудови даної моделі є сформувати уявлення про вплив таких вагомих загальних показників, як середня шкільна оцінка, середня оцінка за іспити, середня оцінка за напівсеместр, форма фінансування, на стан навчання та дату зміни стану навчання. Дана модель використовує алгоритм кластеризації [1].

Результати застосування моделі:

1. Переважна більшість студентів залишаються навчатися до кінця терміну навчання.

2. Відмінники навчання у ЗДІА поділяються на два класи. Студенти, що мають відносно більшу середню шкільну оцінку мають більшу імовірність вчитися за контрактом, ніж ті, що мають порівняно нижчу середню шкільну оцінку.

3. Середня оцінка за іспити у студентів, що найімовірніше зазнають переведення або відрахування, ледве перевищує прохідний бал. Переведення або відрахування відбуваються або у першому півріччі навчання, або наприкінці першого курсу (найбільша кількість), або на початку другого курсу.

4. Існує категорія студентів, що мали добрі або навіть дуже добрі оцінки у школі, проте вже іспити вони здавали на більш посередні оцінки, а під час навчання у ЗДІА результативність падала до низького рівня.

Модель «Вплив шкільних оцінок на успішність навчання у ЗДІА».

Дана аналітична модель побудована з метою генерації правил, які б проводили паралелі між успішністю навчання у стінах ЗДІА, шкільними оцінками та оцінками за іспити. У моделі використовується алгоритм асоціацій [1]. Асоціації між значеннями атрибутів стають правилом, якщо вони справедливі для більш ніж 12-ти студентів.

Результати застосування моделі.

1. Відмінники навчання отримують високий середній бал на вступних іспитах.

2. Відмінні оцінки з фізики та іноземної мови позитивно впливають на подальше навчання у ЗДІА.

3. Студенти, що мають добрі оцінки під час навчання у ЗДІА, отримують середні оцінки за вступні іспити та мають досить високу шкільну оцінку з фізики. Крім цього, середня оцінка з точних наук збігається з загальною шкільною середньою оцінкою.

4. Студенти з незадовільним рівнем навчання мають середню шкільну оцінку на межі посереднього та середнього рівнів. Шкільна оцінка з гуманітарних наук знаходиться на посередньому або навіть на низькому рівні. Шкільна оцінка з математики знаходиться на рівні «добре». Такі студенти показують низькі, та все ж прохідні, результати на вступних іспитах.

Модель «Вплив шкільних оцінок та оцінок за вступні іспити на дату зміни стану навчання».

Дана аналітична модель використовує логістичну регресію у якості алгоритму аналізу. Для детального вивчення обрані такі фактори впливу, як оцінки за іспити та шкільні оцінки.

Отже, порівнюючи оцінки студентів, що відраховувалися у дані два періоди (І-й період – з 2-го по 5-й напівсеместр, ІІ-й – з 5-го по 8-й), виявляються очевидними наступні висновки:

1. Студенти, що відраховуються у І-й період, мають гіршу середню шкільну оцінку, у порівнянні зі студентами, що відраховувалися у ІІ-й період.

2. Студенти, що відраховуються у І-й період, мають кращу шкільну оцінку з фізики, у порівнянні зі студентами, що відраховувалися у ІІ-й період.

3. Студенти, що відраховуються у ІІ-й період, показують добрі результати на іспиті з математики.

ВИСНОВКИ

1. Розроблено механізм зберігання даних для аналізу, використовуючи систему сховища даних, у якій задіяні дані з приймальної комісії ЗДІА про абітурієнтів та дані з деканатів про навчання студентів.

2. Проведено порівняльний аналіз методів автоматичного транспортування даних засобами Integration Services, що входить до складу пакету Business Intelligence.

3. Побудовані аналітичні моделі, в яких задіяні алгоритми технології Data Mining.

4. Проаналізовано результати застосування аналітичних моделей та виявлено вплив низки ознак на подальше навчання студентів.

ЛІТЕРАТУРА

1. Методы и модели анализа данных : OLAP и Data Mining / [Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И.]. – СПб. : БХВ-Петербург, 2004. – 336 с.

2. Rainardi V. Building a Data Warehouse : With Examples in SQL Server / V. Rainardi. – Apress, 2008. – 523 с.

3. Пожуєв В. І. Дослідження зв'язку між даними про абітурієнтів та їх подальшою успішністю за допомогою технології Data Mining / В. І. Пожуєв, А. І. Безверхий, О. С. Бельков, І. В. Кулик // Програмне забезпечення в освіті та науці. – 2009. – №. 1 – С. 78–79.

4. Microsoft SQL 2005 Analysis Services. OLAP и многомерный анализ данных / [Бергер А. Б., Горбач И. В., Меломед Э. Л., Щербин В. А., Степаненко В. П.]. – СПб. : БХВ-Петербург, 2007. – 928 с.