УДК 004.453.4::664.6

РОЗРОБЛЕННЯ АЛГОРИТМІВ ТА МЕТОДІВ ЗАВАНТАЖЕННЯ ІНФОРМАЦІЇ ДО СХОВИЩА ДАНИХ ІНФОРМАЦІЙНОЇ СИСТЕМИ ВАТ «МАКАРОННА ФАБРИКА»

С.В. Грибков
Старший викладач*
Контактний тел.: (044) 287-96-77
E-mail: sergio_nuft@ukr.net

Л.Г. Загоровська
Кандидат технічних наук, доцент*
Контактний тел.: (044) 287-96-77
E-mail: zagorov@i.ua
*Кафедра інформаційних систем
Національний університет харчових технологій
вул. Володимирська, 68, м. Київ, Україна, 01601

Розглянуто проблему інтеграції різних інформаційних джерел в інформаційній системі та запропоновано підхід використання єдиного сховища даних як інформаційної інфраструктури підприємства. Розроблено алгоритми та методи завантаження інформації до сховища даних

Ключові слова: інформаційна система, база даних, сховище даних, алгоритми, методи завантаження даних

Рассмотрена проблема интеграции разных информационных источников в информационной системе и предложен подход использования единственного хранилища данных как информационной инфраструктуры предприятия. Разработаны алгоритмы и методы загрузки информации в хранилище данных

Ключевые слова: информационная система, база данных, хранилище данных, алгоритмы, методы загрузки данных

The problem of integration of different informative sources is considered in the informative system and offered approach the use of only depository of data as an informative infrastructure of enterprise. Algorithms and methods of loading of information are worked out to the depository of data

Keywords: informative system, database, depository of data, algorithms, methods of loading of data

Вступ

В наш час ефективність функціонування будь-якого підприємства значною мірою залежить від стану комп'ютеризації виробничих процесів та управління. На сьогоднішній день на ВАТ «Макаронна фабрика» автоматизовані системи технологічними процесами повністю забезпечують комп'ютерну підтримку технології виготовлення макаронних виробів. Крім автоматизації технологічних ліній на фабриці використовуються комп'ютерні засоби для забезпечення успішної роботи структурних підрозділів та керівного персоналу. Усі автоматизовані системи підприємства в даний час працюють з різними базами даних (БД), представлених різними типами файлів та не поєднаних між собою. Така ситуація обумовлює розмежований доступ до різнобічної інформації різних виробничих підрозділів та керівництва підприємства і не задовольняє сучасних потреб щодо використання ними спільних наборів даних.

Постановка задачі

Для збереження існуючих автоматизованих систем та окремих засобів автоматизації, інтеграції наявних різноплатформенних інформаційних джерел з метою створення репозиторію інформаційної інфраструктури підприємства обрано підхід використання єдиного сховища даних (СД) як інформаційного джерела інформаційні системи (ІС) підприємства [1,5,6]. Ефективність використання СД в якості інформаційної ін-

фраструктури підприємства значною мірою залежить від ефективності та об'єктивності його формування, а тому розроблення алгоритмів та пакетів завантаження інформації до СД є досить актуальною задачею.

Методика розроблення

Структуру СД реалізовано на основі стандартного архіву даних та тематичних областей [6]. Стандартний архів даних забезпечує процес семантичної інтеграції й проміжного зберігання інформації, а також координацію даних з різних джерел для заповнення ними тематичних областей, що являють собою вітрини даних, призначені для вирішення спеціалізованого кола задач. Інформація у стандартному архіві зберігається у деталізованому вигляді, а тематичні області містять агреговані та відфільтровані дані. Стандартний архів фактично являється репозиторієм інформаційної інфраструктури підприємства та ґрунтується на реляційному підході зберігання даних. Тематичні області призначені для підтримки бізнес-процесів певних напрямків. Виходячи з того, що ступінь агрегації та представлення даних для кожної з них різні, тому доцільно використовувати змішані підходи при побудові структур даних. У випадку виникнення потреби у більш детальній чи розширеній інформації, що не відображена у тематичній області, при вирішенні певної задачі користувачеві буде забезпечено доступ до інформації у стандартному архіві даних з урахуванням прав його доступу.

Інформація повинна завантажуватись до СД в автоматичному режимі, бути прихованою від користувачів та утворювати хронологічну послідовність. При завантаженні інформації до СД в цілому, або ж до окремої тематичної області, слід виконати наступну послідовність дій:

отримати набір даних з інформаційного джерела;

перетворити значення полів до визначеного типу даних у СД;

виконати задані розрахунки, операції агрегації, перетворення з одних вимірів до інших, тощо;

перевірити існування заданого набору даних;

занести інформацію до СД.

Зазначена послідовність дій повинна виконуватись з певною періодичністю, або ж при кожному зверненні користувача. Інтервал виконання пакетів для кожної тематичної області буде різним.

Визнаними лідерами на ринку інформаційних технологій, які пропонують СУБД з підтримкою технологій СД на рівні підприємств, є IBM, Oracle, Teradata та Microsoft [4]. Кожен з них пропонує свої підходи для забезпечення виконання операцій перетворення даних та завантаження їх до СД.

Доцільність використання з поміж інших саме СУБД MS SQL Server 2005 та супровідних сервісів обумовлена функціональною достатністю та повнотою, забезпеченням спадковості та гнучкого переходу до нових версій продукту, наявністю вбудованих сервісів взаємодії між різними СУБД інших виробників, а також прийнятною ціною для підприємств середнього та малого бізнесу. Дана СУБД має вбудовані засоби Integration Services (SSIS), що забезпечують підтримку гнучкої та потужної архітектури для ефективної інтеграції даних, розроблення та використання пакетів перетворення даних за заданим алгоритмом розробника, глибоку інтеграцію пакетів з інструментарієм аналітичної обробки даних Data Mining в Analysis Services [2,3,4]. Крім цього забезпечується підтримка систематизації та аналізу текстової інформації без попередньої обробки, а також висування припущень щодо її корисності на основі заданих бізнес-правил. Можливість використання спеціалізованого інструментарію для створення пакетів завантаження даних та забезпечення встановлення параметрів роботи за розкладом є переконливим аргументом на користь використання даного засобу.

На початковому етапі побудови основи пакета використовувалися майстри, а потім його функціональні можливості доопрацьовувалися в інтегрованому середовищі Business Intelligence Development Studio (BIDS). Пакети завантаження з файлів, що не мають чіткої структури та потребують складної обробки, розроблялися в середовищі BIDS завдяки його можливостям створювати та редагувати в графічному режимі пакети перетворення будь-якої складності з візуалізацією структур та алгоритмів їх роботи. Крім цього у даному середовищі розроблялись пакети завантаження інформації для моделей аналітичної обробки Analysis Services та підготовки даних до використання в Reporting Services.

Побудова кожного пакету включає в себе визначення операцій потоку управління, алгоритм оброблення потоку даних та налаштування його виконання. Потік управління визначає алгоритм виконання дій до оброблення потоку даних та після нього. Він забезпечує виконання системних команд, сценаріїв та операцій взаємодії з об'єктами й службами операційної системи, а також елементами та сервісами обраної СУБД. Потік даних забезпечує алгоритм зчитування, перетворення та запису інформації. При створенні пакетів пов'язані операції потоку даних об'єднують у контейнери, а для забезпечення повторення певних операцій використовують цикли. Кожен алгоритм оброблення потоку даних визначає джерело та приймача інформації. В якості джерела даних виступають таблиці та представлення БД, текстові файли, електронні таблиці, xml- та html-файли. Приймачами даних, окрім БД та файлів можуть бути OLAP-куби, багатовимірні структури та пакети аналітичної обробки даних.

За функціональними можливостями пакети завантаження розділено на дві групи: пакети завантаження даних до стандартного архіву та пакети завантаження даних до тематичних областей.

Перша група призначена для зчитування інформації з джерел, перетворення її до форматів, визначених у стандартному архіві та завантаження до нього. Джерела інформації різноплатформенні, територіально розташовані на відстанях, що потребує значних системних ресурсів на їх оновлення. Тому при розробленні пакетів використано засоби, що сприяють зменшенню навантаження на сервер БД та локальну мережу.

Для зменшення часу виконання пакетів завантаження зроблено акцент на зчитування, оброблення та передачу лише тих даних, що змінилися з часу останнього виконання пакету. Для цього до стандартного СД введено таблицю, що містить статистику виконання пакетів та складається з наступних полів: *ідентифікатор пакета* (поле для збереження інформації, що забезпечує ідентифікацію всього запису до відповідного

пакета), *ідентифікатор джерела, час останнього виконання, вказівник на останній оброблений елемент, відмітка про виконання* (або помилка у разі виникнення). Зчитування та занесення даних до цієї таблиці відбувається в пакеті. Вказівник для кожного пакету має свою структуру, це викликано різними методами збереження інформації в кожному джерелі даних. Для спрощення збереження вказівника використано поле зі строковим типом даних. Робота кожного пакету розпочинається зі зчитування значення вказівника та перетворення його зі строкового типу до набору значень, що записуються у відповідні змінні пакету для пошуку в джерелі останнього зчитаного запису. Після знаходження початку нових записів пакет їх обробляє, після чого останнє зчитане значення перетворюється у строкову змінну та записується в таблицю статистики виконання пакетів. У випадку необхідності зчитування даних з кількох джерел для кожного з них в пакеті передбачено здійснення відповідної відмітки в таблиці статистики, при цьому в поле *ідентифікатор* джерела вноситься відповідний запис. Таким чином відбувається оброблення не всієї інформації з джерела, а лише нової.

Для зменшення навантаження на локальну мережу запропоновано два режими виконання пакетів: за заданим розкладом та за необхідності отримання оперативних даних. Розклад виконання задано для ряду пакетів виробничого контуру, а саме: дані про роботу апаратно-технічного комплексу оновлювати кожну годину протягом доби; сформовані заявки на виробництво фіксувати при їх виникненні; дані зі змінного журналу та про передачу готової продукції на склад вносити на початку наступної зміни. Всі інші пакети завантаження доцільно виконувати за наступним розкладом: на початку робочого дня – після завантаження даних про роботу 3-ї зміни; в обідню перерву; перед закінченням робочого дня – за 30 хвилин до закінчення першої зміни. Крім цього передбачена можливість оперативного виконання певної низки пакетів поза розкладом, що може бути викликано необхідністю невідкладного отримання чи коригування техніко-економічних показників діяльності як всього підприємства так і певного підрозділу.

На рівні завантаження стандартного архіву даних усі пакети поєднуються в групи, кожна з яких призначена для завантаження даних з певного джерела. Таким чином, послідовність виконання пакетів в кожній групі визначається алгоритмом заповнення спочатку довідників, а потім таблиць, що їх використовують. За таким самим алгоритмом виконуються групи пакетів. Наприклад, спочатку завантажується інформація про роботу зміни, потім про передачу готової продукції на склад, після чого дані про продаж та відвантаження продукції. Послідовність завантаження даних сприяє збереженню їх цілісності та хронології.

При розробленні пакетів в якості інформаційних джерел використано:

бази даних бухгалтерії, відділу збуту й постачання, змінного журналу виробництва, складу готової продукції, сировини й матеріалів, відділу кадрів, системи управління автоматизованою лінією Pavan;

репозиторії файлів статистики системи управління автоматизованими лініями Buhler_1, Buhler_2, технологічної лабораторії й головного технолога.

Реляційна структура зазначених інформаційних джерел забезпечила відносну простоту розроблення

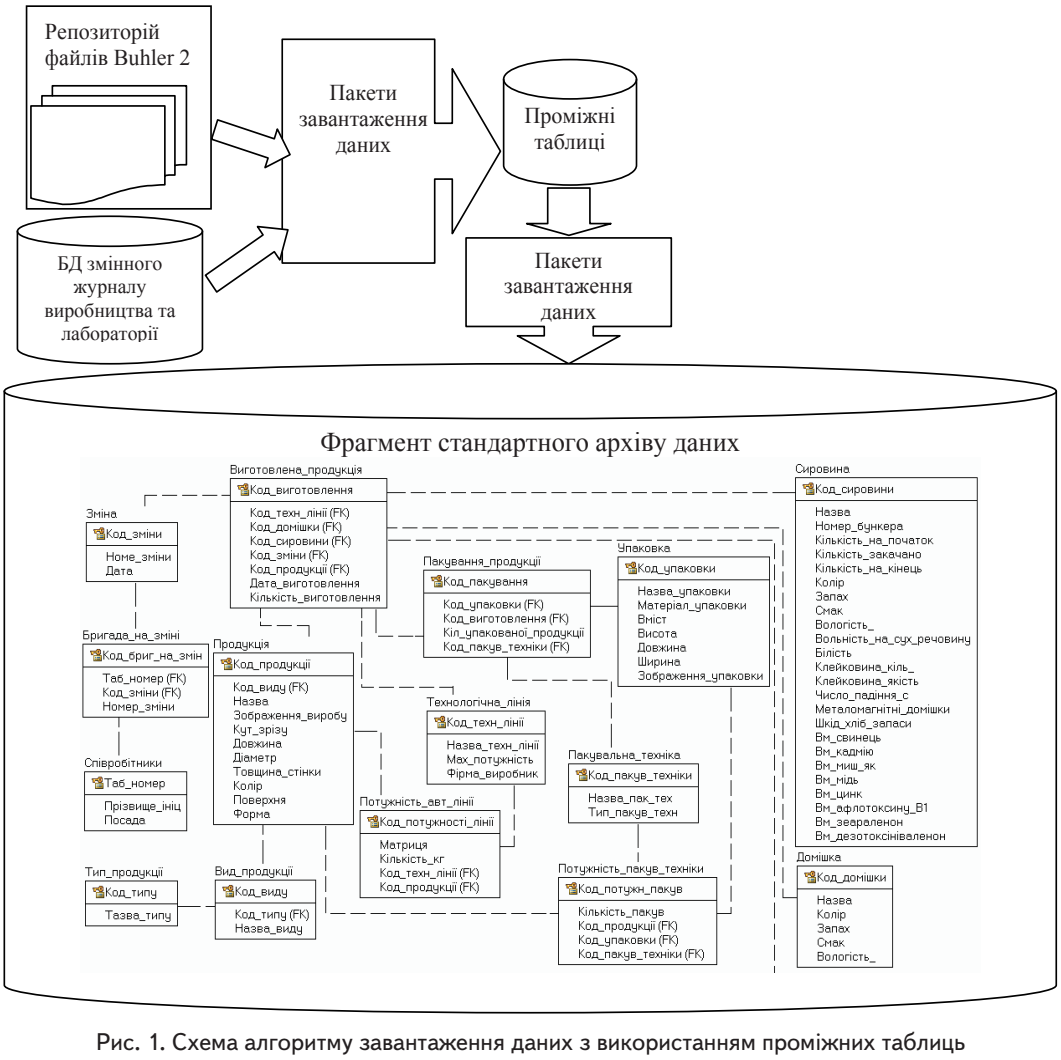

Рис. 1. Схема алгоритму завантаження даних з використанням проміжних таблиць

пакетів. При створенні пакетів завантаження інформації до джерел, що мають нереляційну структуру, розроблено багатокрокові алгоритми з використанням таблиць для проміжного зберігання даних у спеціально створеній БД на сервері. Для зменшення дискового простору такі таблиці очищаються після виконання усіх пакетів. Приклад багатокрокового алгоритму завантаження даних подано на рис. 1.

Даний алгоритм реалізує заповнення певних таблиць СД з джерел, представлених репозиторієм файлів статистичних повідомлень системи управління автоматизованою лінією Buhler_2, БД змінного журналу виробництва та лабораторії.

Складність подібних завантажень полягає в тому, що отримати та в подальшому очистити інформацію, що зберігається в неструктурованому вигляді, без використання проміжних таблиць практично неможливо. Головною вимогою при завантаженні інформації є отримання даних без надлишковості, повторень та збереження цілісності реляційної структури стандартного архіву даних.

При створенні пакетів використано принцип групування за джерелом завантаження. Прикладом такого пакету є *Завантаження повідомлень*, призначеного для оброблення наборів текстових файлів, що зберігають детальну інформацію про експлуатацію автоматизованої лінії Buhler_1. Кожен запис файлу складається з повної дати, типу повідомлення та тексту повідомлення. Основні кроки алгоритму потоку дій даного пакету:

зчитування покажчика останнього запису з таблиці статистики виконання пакетів;

формування та передача значень параметрів потоку даних;

виконання циклу по усіх файлах починаючи із заданого;

формування строкової змінної із значень параметрів останнього зчитаного файлу.

Алгоритм оброблення потоку даних цього пакету подано на рис. 2.

Представлений алгоритм потоку даних складається з наступних кроків:

зчитується кожний запис файлу та стандартними засобами виділяється три основні поля: *Дата, Тип повідомлення, Текст повідомлення*;

перетворення значення поля *Дата* символьного типу у тип дати, що включає дату та час;

перевірка записів за полем *Дата* та вибір тих записів, що мають більше значення від вхідного параметра *Дата та час*;

в залежності від значення поля *Тип повідомлення* далі оброблення проходить по одній з чотирьох гілок: *Message, Alarm, Error та Неідентифіковані;*

в залежності від типу повідомлення інформація проходить оброблення та записується у відповідну таблицю.

Для наведеного прикладу результуючою таблицею у стандартному архіві даних буде *Виготовлена продукція на виробництві, що складається з основних полів: Код технологічної лінії, Код домішки, Код сировини, Код зміни, Код продукції, Дата виготовлення, Кількість виготовленої, Час початку, Час закінчення*. Для логічного закінчення та пояснення призначення наведеного пакету необхідно сказати, що в подальшому використовуючи створені пакети, дані з проміжних таблиць приводяться до відповідності реляційної структури стандартного архіву даних та завантажуються до нього. Таке складне на перший погляд перетворення забезпечить завантаженим даним повну відповідність та хронологію.

Пакети завантаження даних до тематичних областей кардинально відрізняються від попередньої групи пакетів. Основною відмінністю є те, що вони працюють з даними стандартного архіву, що має реляційну структуру та приведений до третьої нормальної форми. У зв'язку з цим при виконанні таких пакетів не виникає потреб у проведенні очищення даних та їх конвертації до певних типів. Такі пакети групуються за тематичними областями, які вони заповнюватимуть. При проектуванні алгоритмів обробки потоків даних для цих пакетів враховуються бізнес-процеси відповідних тематичних областей. Таблиці тематичних областей являють собою набори даних для адаптації моделей Data Mining, побудови звітів та візуального відображення інформації відповідно до бізнес-процесів, для яких вони призначені. За рахунок інструментарію та функцій Integration Services є можливість побудови пакетів, що забезпечують виконання створених моделей аналітичної обробки даних. При формуванні наборів даних для завантаження до тематичних областей враховуються

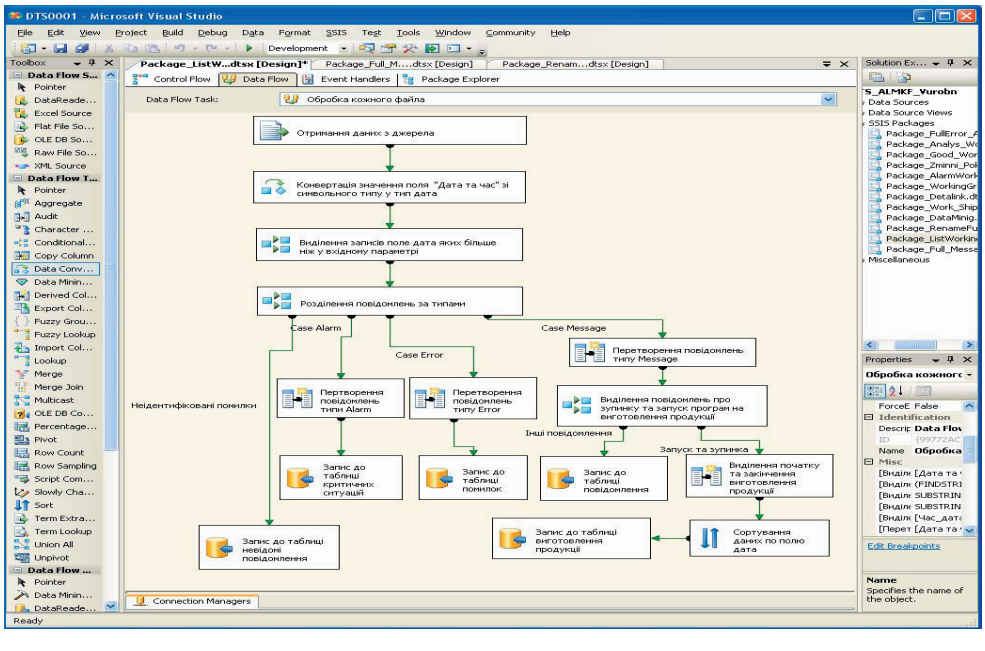

Рис. 2. Схема алгоритму оброблення потоку даних пакету «Завантаження повідомлень»

часові межі накопичення їх до стандартного архіву. Тому для кожного пакету були задані свої часові межі для отримання інформації зі стандартного архіву даних.

Розташування та інсталяцію пакетів завантаження даних здійснено засобами Integration Services. Пакети зберігаються централізовано на сервері MS SQL Server 2005 в папці розташування стандартного архіву даних, при цьому забезпечується їх максимальний захист від пошкодження. Для автоматизованого послідовного виконання пакетів в заданий проміжок часу використано спеціальну службу – SQL Server Agent, що забезпечує виконання локальних та багатосерверних завдань.

Результати та висновки

Розроблені алгоритми та методи завантаження інформації до стандартного архіву та тематичних областей єдиного сховища даних випробувано в умовах діючого підприємства – ВАТ «Макаронна фабрика». Результати випробувань дають підстави стверджувати, що технологія використання СД в якості репозиторію інформаційної інфраструктури підприємства є ефективною та задовольняє сучасні вимоги підрозділів підприємства використовувати спільні набори даних, а також забезпечує узгоджену роботу різноплатформенних засобів автоматизації. Позитивні результати підтверджують функціональну достатність і повноту інструментарію Integration Services SQL Server та інтегрованого середовища Business Intelligence Development Studio як вдалого вибору засобів для розробки.

Література

1. E. F. Codd, S.B.Codd. Providing OLAP. On-line Analytical Processing to User-Analists: An IT Mandate. C. T. Salley, E. F. Codd & Associates, 1993
2. Kamal Hathi, Microsoft Введение в SQL Server 2005 Integration Services Май 2005 http://citforum.univ.kiev.ua/database/mssql/integration_services/#3.2.1
3. А.В. Бергер Microsoft SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных. – СПб.: БХВ-Петербург, 2007. – 928 с.: ил.
4. А.А. Берсегян Технологія аналізу даних: Data Mining, Visual Mining, Text Mining, OLAP. – 2-е вид. перероб. та доп. – СПб.: БХВ-Петербург, 2008. – 384 с.: ил.
5. Л.Г. Загоровська, С.В. Грибков. Використання методів інтелектуального аналізу та добування даних при удосконаленні систем підтримки прийняття рішень / „Вісник Хмельницького національного університету ” №5 . – Х.: ХНУ, 2008.
6. Л.Г. Загоровська, С.В. Грибков. Моделювання структури сховища даних системи підтримки прийняття рішень для підприємств макаронної галузі // Інтелектуальні системи підтримки прийняття рішень та інформаційні технології. Матеріали Міжнародної науково-практичної конференції. – Чернівці: Рута, 2006. – 304 с.