Економічні науки

УДК 004.6: 004.9

**Жмуркевич Андрій Євгенович**

*кандидат економічних наук, доцент*

*Львівський національний університет імені Івана Франка*

**Жмуркевич Андрей Евгеньевич**

*кандидат экономических наук, доцент*

*Львовский национальный университет имени Ивана Франко*

**Zhmurkevych Andriy**

*PhD in Economics, Associate Professor*

*Ivan Franko National University of Lviv*

**Рожок Ірина Валеріївна**

*студент магістратури*

*Львівського національного університету імені Івана Франка*

**Рожок Ирина Валериевна**

*студент магистратуры*

*Львовского национального университета имени Ивана Франко*

**Rozhok Iryna**

*Student of the*

*Ivan Franko National University of Lviv*

# МЕТОДИКА ТЕХНОЛОГІЇ ETL ЯК ПІДХІД ІНТЕГРАЦІЇ ДАНИХ
# МЕТОДИКА ТЕХНОЛОГИИ ETL КАК ПОДХОД ИНТЕГРАЦИИ ДАННЫХ
# ETL TECHNOLOGY AS AN APPROACH TO DATA INTEGRATION

***Анотація.*** *Дана стаття присвячена дослідженню проблеми інтеграції корпоративних даних. Метою роботи є аналіз існуючих способів організації інтеграції великих обсягів даних, і реалізація програмної системи, що дозволяє реалізувати цілісну систему повного циклу виймання, перетворення та завантаження даних (Extract, Transform, Load, ETL).*

*Велика частина інформації необхідної для аналітичних додатків, знаходиться в системах управління операційною діяльністю фірми, базах даних і окремих файлах користувачів. Засоби інтеграції даних забезпечують необхідну інфраструктуру для перетворення розрізнених початкових даних в єдиний ресурс.*

*Коли фірма починає створювати інтегроване корпоративне сховище даних (Enterprise Data Warehouse, EDW), кількість потоків даних у багато разів зростає. Для ефективного використання EDW необхідна наявність у ньому даних про всі аспекти функціонування організації і, відповідно, передача великих об'ємів детальних даних з кожної початкової системи в сховище.*

*Визначимо основні проблеми інтеграції даних: часто для вирішення завдань інтеграції в організаціях використовується велика кількість інструментів і технологій. Проте ефективність таких рішень різко падає із збільшенням кількості потоків даних. Для забезпечення ефективної інтеграції необхідні інструменти, які зможуть надати можливості загального огляду і управління всіма потоками даних. Інструменти для інтеграції повинні однаково ефективно працювати зі всіма використовуваними в фірмі СУБД, джерелами даних, системами обміну повідомленнями і т.д.*

*Розроблений та представлений у статті метод інтеграції призначений для невеликих підприємств, які працюють з великими обсягами даних і потребують маловитратних, легких в розгортанні і масштабуванні рішень.*

*Були досліджені особливості роботи з даними для малих підприємств. На їх основі сформульовані вимоги до вхідних даних, а також методи їх фільтрації та очистки.*

*Результатом роботи є розроблений алгоритм, спрямований на задоволення цих вимог і загрузки даних в заздалегідь підготовлене сховище.*

***Ключові слова:*** *сховище даних, інтеграція даних, програмна система*

***Аннотация.*** *Данная статья посвящена исследованию проблемы интеграции корпоративных данных. Целью работы является анализ существующих способов организации интеграции больших объемов данных, и реализация программной системы, которая позволяет реализовать целостную систему полного цикла извлечения, преобразования и загрузки данных (Extract, Transform, Load, ETL).*

*Большая часть информации необходимой для аналитических дополнений, находится в системах управления операционной деятельностью фирмы, базах данных и отдельных файлах пользователей. Средства интеграции данных обеспечивают необходимую инфраструктуру для превращения разрозненных начальных данных в единственный ресурс.*

*Когда фирма начинает создавать интегрированное корпоративное хранилище данных (Enterprise Data Warehouse, EDW), количество потоков данных во много раз растет. Для эффективного использования EDW необходимо наличие в нем данных обо всех аспектах функционирования организации и, соответственно, передача больших объемов детальных данных из каждой начальной системы в хранилище.*

*Определим основные проблемы интеграции данных: часто для решения заданий интеграции в организациях используется большое количество инструментов и технологий. Однако эффективность таких решений резко*

*падает с увеличением количества потоков данных. Для обеспечения эффективной интеграции необходимы инструменты, которые смогут предоставить возможности общего обзора и управления всеми потоками данных. Инструменты для интеграции должны одинаково эффективно работать со всеми используемыми в фирме СУБД, источниками данных, системами обмена сообщениями и так далее.*

*Разработанный и представленный в статье метод интеграции предназначен для небольших предприятий, которые работают с большими объемами данных и нуждаются малорасходных, легких в развертывании и масштабировании решениях.*

*Были исследованы особенности работы с данными для малых предприятий. На их основе сформулированы требования к входным данным, а также методы их фильтрации и очистки.*

*Результатом работы является разработанный алгоритм, направленный на удовлетворение этих требований и загрузки данных в предварительно подготовленное хранилище.*

***Ключевые слова:*** *хранилище данных, интеграция данных, программная система*

***Summary.*** *This article is devoted to the study of the corporate data integration problem. The aim of the work is the analysis of the existing ways of big data integration and implementation of a software system that allows the implementation of the full cycle of data extraction, transformation, and load (ETL).*

*Most of the information required for analytical applications stored in the corporate management systems, databases and individual user files. Data integration tools provide the necessary infrastructure for converting disparate data sources into a single resource.*

*When the company starts an integrated enterprise data warehouse (EDW) development, the number of data streams increases many times. For efficient use of EDW, it requires data on all aspects of the enterprise operations and the transfer of large volumes of detailed data from each source system to the warehouse.*

*The main problem of data integration is the use of a large number of tools and technologies for solving integration problems in organizations. However, the effectiveness of such solutions drops with an increase in data streams quantity. In order to ensure effective integration, tools that can provide a general overview and management of all data flows are needed. Tools for integration should work equally effective with all DBMS used by the organization, data sources, messaging systems, etc.*

*The integration method developed and presented in the article is designed for small enterprises that work with large data volumes that require low cost, easy to deploy and scale solutions.*

*The features of work with data for small enterprises were investigated. Based on them, the requirements for the input data, as well as methods of their filtration and cleaning are formulated.*

*The result of the work is a developed algorithm and data loading procedure into a prepared data warehouse.*

***Key words:*** *data warehouse, data integration, software system*

**Постановка проблеми.** Інтеграція даних охоплює практики, архітектурні підходи й програмні інструменти для забезпечення погодженого доступу й доставки даних для всього спектра додатків і бізнес-процесів. Як свідчать дослідження, витрати на програмні засоби інтеграції даних сьогодні неухильно ростуть у різних індустріях і географічних регіонах. Це відбувається через невідповідність існуючих підходів до керування даними й ситуації з

автоматизованою підтримкою операцій бізнесу з боку прикладних систем. Керування наскрізними бізнес-процесами, що охоплюють різні підрозділи фірми та її зовнішніх партнерів і замовників, демонструє свою ефективність і підкріплено цілком зрілими методами й технологіями інтеграції. Для того, щоб оптимізувати використання та цінність інформації, необхідно знайти альтернативну задачу – процес, зробити дані простими, динамічними і зрозумілими для сприйняття, необхідно обробляти значні обсяги інформації. При цьому ситуація змінюється настільки швидко, що практично не залишає часу для якісного опрацьовування управлінських рішень. Основною задачею для даних значних об'ємів є оперативна автоматизація даних. Але як бути, якщо необхідна інформація знаходиться в різних системах, джерелах, файлах, часто не зв'язаних між собою? У цьому випадку необхідно об'єднати всі інформаційні ресурси фірми в єдиний інформаційний простір за допомогою технологій інтеграції. Узгодженість, актуальність і доступність інформації є важливим елементом в процесі ухвалення рішень і подальшого розвитку фірми. Для визначення своєчасних і обґрунтованих рішень будь-якій фірмі необхідна надійна організація оперативних даних. Дані необхідно витягти, очищені (знищити повторення, виправити помилки введення), узгодити, агрегувати і привести до єдиного стандарту, зрозумілого і звичного для сприйняття користувачем. Розв'язуючи ці задачі, стикаються із загальними проблемами, такими як розрізненість даних, відсутність консолідації і взаємозв'язків, помилки в записах і дублювання інформації, множинність джерел даних і багато що інше. Для вирішення цих завдань обробки інформації використовуються різні методи й інструменти інтеграції даних. Правильна постановка завдання, вибір технології інтеграції і реалізація її за допомогою програмних засобів нададуть фірмі можливість контролю над інформацією.

**Аналіз останніх досліджень і публікацій.** За останні роки здійснено багато досліджень та публікацій, які стосуються питань щодо методик інтеграції даних. Розглянемо найважливіші з цих публікацій. У роботах (Medykovskyi, et al., 2015; O'Leary, 2000; Data Warehouse Design, 2013; Lambda Architecture, 2015; Stashevskyi, & Hrytsiuk, 2013; BeyeNETWORK, 2013) проаналізовано архітектури багаторівневих систем інтерграції даних. Особливістю таких архітектур є орієнтація їх на використання сучасних інтернет-технологій та сховища даних на кожному рівні управління. У працях (Bertocco, et al., 2002; Chiu, Yu Hsien, et al., 2014; BeyeNETWORK, 2015; Teslyuk, et al., 2018) значну увагу приділено питанням збирання та трансформації даних. Проте у розглянутих працях недостатньо уваги приділено розробленню засобів збереження даних для багаторівневих систем. У роботах (Data Warehouse Design, 2013; BeyeNETWORK, 2015) показано, що для побудови сховищ даних використовують два підходи: "класичний" – на основі реляційних баз даних та на основі концепції Big Data.

**Постановка завдання.** Розгляд проблеми опрацювання інтеграції дасть змогу вирішити питання використання даних з розрізнених джерел для підвищення достовірності та надійності інформації, а отже, підвищить ефективність використання даних системою. Розглянемо технологію інтеграції на рівні передачі даних. Процес витягання, перетворення і завантаження даних підтримується так званими ETLінструментами (extraction, transformation, loading). Загалом додатки ETL витягують інформацію з джерел даних (наприклад, баз даних, бінарних і текстових файлів, устаткування, електронних таблиць, електронної пошти і т.п.), перетворюють її на формат, підтримуваний базою даних призначення, а потім завантажують до неї перетворену інформацію. ETL-засоби є незамінними для вирішення найрізноманітніших

завдань: для перенесення інформації з успадкованих додатків в нові або для передачі операційних даних в системи сховища даних.

**Виклад основного матеріалу досліджень.** Враховуючи концептуальну структуру схеми технології ETL, встановимо покрокові дії інтеграції.

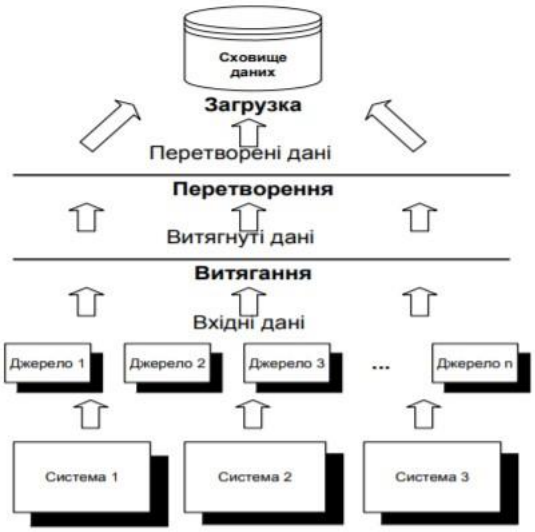
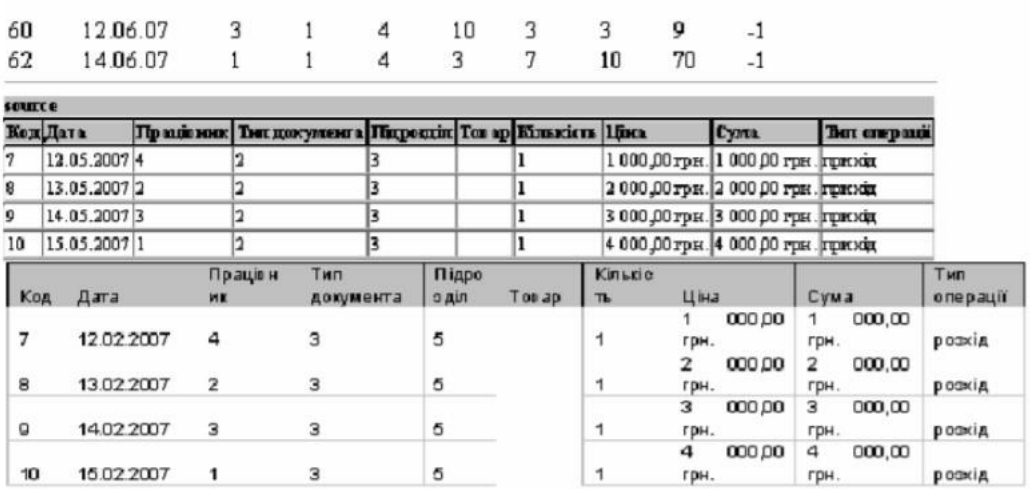



**Рис. 1. Концептуальна схема ETL – процесу**

**Підготовка і отримання вхідних даних з декількох джерел**

На даному етапі вибираємо таблиці (документи), які містять однотипні дані отримані з різних джерел інформації, зображені в різних форматах та структурах (MS Access, Visual Foxpro, dBase, Ms Excel, структурований текст тощо), наприклад, текстовий файл (файл з розділювачами), таблиця з MS Access та електронна таблиця.



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 60 | 12.06.07 | 3 | 1 | 4 | 10 | 3 | 3 | 9 | -1 |
| 62 | 14.06.07 | 1 | 1 | 4 | 3 | 7 | 10 | 70 | -1 |

source

| Код | Дата | Працівник | Тип документа | Підрозділ | Товар | Кількість | Ціна | Сума | Тип операції |
|---|---|---|---|---|---|---|---|---|---|
| 7 | 12.05.2007 | 4 | 2 | 3 | | 1 | 1 000,00 грн. | 1 000,00 грн. | прихід |
| 8 | 13.05.2007 | 2 | 2 | 3 | | 1 | 2 000,00 грн. | 2 000,00 грн. | прихід |
| 9 | 14.05.2007 | 3 | 2 | 3 | | 1 | 3 000,00 грн. | 3 000,00 грн. | прихід |
| 10 | 15.05.2007 | 1 | 2 | 3 | | 1 | 4 000,00 грн. | 4 000,00 грн. | прихід |

| Код | Дата | Працівник | Тип документа | Підрозділ | Товар | Кількість | Ціна | Сума | Тип операції |
|---|---|---|---|---|---|---|---|---|---|
| 7 | 12.02.2007 | 4 | 3 | 5 | | 1 | 1 000,00 грн. | 1 000,00 грн. | розхід |
| 8 | 13.02.2007 | 2 | 3 | 5 | | 1 | 2 000,00 грн. | 2 000,00 грн. | розхід |
| 9 | 14.02.2007 | 3 | 3 | 5 | | 1 | 3 000,00 грн. | 3 000,00 грн. | розхід |
| 10 | 15.02.2007 | 1 | 3 | 5 | | 1 | 4 000,00 грн. | 4 000,00 грн. | розхід |

**Рис. 2. Дані, створені в текстовому редакторі Word, Access та Excel**

Як видно з таблиці метаданих, джерела містять однотипну інформацію, яка має спільний характер, але відрізняється складом, способами подання і форматами. На основі цих метаданих може бути утворена інтегрована таблиця, яка буде виконувати функції оперативного сховища даних.

**Визначення вимог до вхідних даних і критеріїв перетворення**

Для попередньо отриманих і збережених у оперативному сховищі даних розробимо систему норм і критеріїв перетворення їх до форми, придатної для збереження у сховищі даних. Приклади критеріїв та вимог: – обмеження часового періоду фактів "від ... – до ...";

Приклад: вибір даних за останній місяць

*SELECT doc.**

*FROM doc*

*WHERE (((doc.evdate)>=DateDiff(„d", NOW(), 30)));*

– обов'язкове значення виміру чи показника факту;

*SELECT doc.**

*FROM doc*

*WHERE (((doc.documenttype) Is Not Null));*

– входження значень виміру у попередньо визначену множину чи область визначення;

*SELECT doc.**

*From doc*

*WHERE doc.documenttype IN (SELECT id FROM doc)*

– унікальність значень.

SELECT DISTINCT doc.evdate, doc.*

FROM doc;

**Аналіз, фільтрація та перетворення вхідних даних**

Сформуємо та виконаємо стосовно таблиці оперативного сховища даних запити на вилучення або зміну даних, які не відповідають критеріям, побудованим у пункті 2. Наприклад:

– вилучення даних, які не містять обов'язкових значень і внаслідок цього не можуть бути застосовані у сховищі даних:

*DELETE **

*FROM documents1*

*WHERE product_id Is Null;*

– виправлення некоректних чи помилкових значень (наприклад, коли у джерела даних внесено не коди підрозділів, а назви:

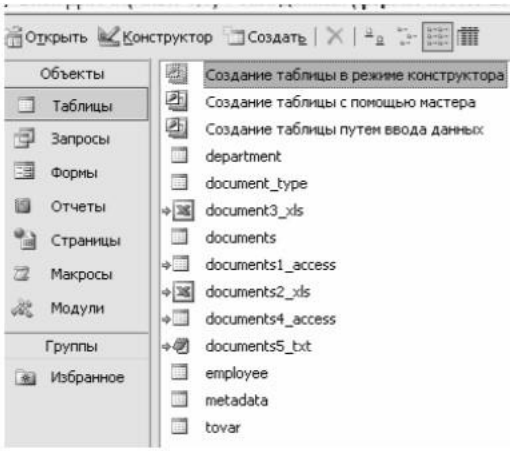
```
Public Sub Correct_Emplid(Source As String)
Dim str As String
Dim dept As Recordset

Dim str1 As String
Dim rs1 As Recordset
Dim str2 As String
Dim rs2 As Recordset

'вибираємо всі значення id документів із source
Set rs = CurrentDb.OpenRecordset("select [emploeeid] from [" & Source & "] ")
While Not rs.EOF
'MsgBox (rs!emploeeid)
'визначаємо код документу за назвою
str1 = "select id from Emploee_Id where name='" & rs!emploeeid & "'"
Set rs1 = CurrentDb.OpenRecordset(str1)
If Not rs1.EOF Then 'якщо знайдено текст то робимо заміну
  rs.Edit
  rs!emploeeid = rs1!id
  rs.Update
End If
rs1.Close
rs.MoveNext
Wend
rs.Close
End Sub
```

**Рис. 3. Завантаження таблиці фактів**

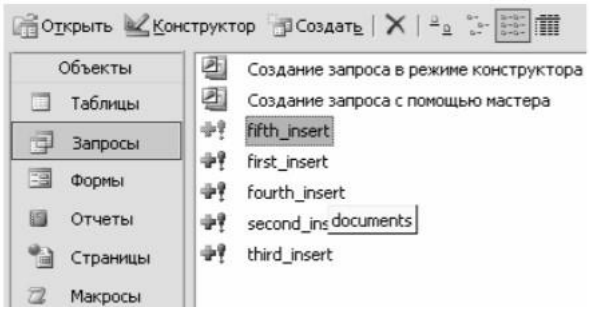
Створимо та виконаємо запит (процедуру) переміщення значень з підготованої таблиці вхідних даних сховища оперативних даних у таблицю фактів та проаналізуємо виконані дії. – підготовка вхідних даних: внесемо зміни в склад, структуру і зміст джерел даних (за потреби доповнимо новими стовпчиками, вилучимо зайві або змінимо параметри стовпчика, внесемо

відсутні значення), узгодивши їх відповідно до таблиці метаданих та структури оперативного сховища даних (ОСД); – створимо таблицю ОСД; – приєднаємо визначені джерела даних як зовнішні таблиці; – сформуємо та реалізуємо запити (процедури) для перенесення даних зовнішніх джерел в ОСД.

Приклад:




Вікно бази даних із запитами на додавання даних із різних джерел:




Запит на додавання даних із текстового файла:

*INSERT INTO documents (evdate, employee_id, document_type, dept_id, tovar_id, [count], price, suma, type)*

*SELECT Поле2, Поле3, Поле4, Поле5, Поле6, Поле7, Поле8, Поле9, Поле10 FROM documents5_txt;*

Запит на додавання даних із файла MS Access:

*INSERT INTO documents ( evdate, employee_id, document_type, dept_id, tovar_id, [count], price, suma, type )*

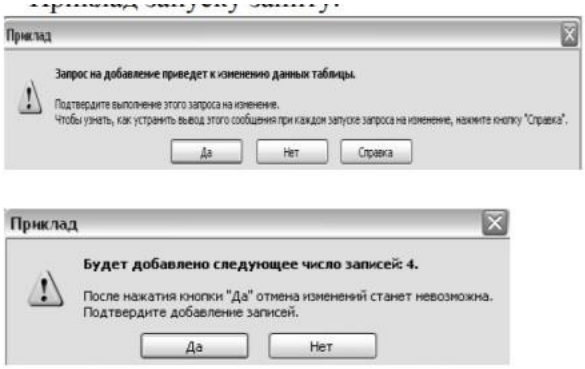
*SELECT documents1_access.evdate, documents1_access.employee_id, documents1_access.document_type, documents1_access.dept_id, documents1_access.tovar_id, documents1_access.count, documents1_access.price, documents1_access.suma, documents1_access.type FROM documents1_access;*

Запит на додавання даних із електронної таблиці:

*INSERT INTO documents ( evdate, employee_id, document_type, dept_id, tovar_id, [count], price, suma, type )*

*SELECT documents1_access.evdate, documents1_access.employee_id, documents1_access.document_type, documents1_access.dept_id, documents1_access.tovar_id, documents1_access.count, documents1_access.price, documents1_access.suma, documents1_access.type FROM documents1_access;*

Приклад запуску запиту:




**Аналіз та верифікація сховища даних**

Перевіримо та обґрунтуємо працездатність сховища даних, для чого перевіримо наявність всіх необхідних значень у стовпчиках таблиці фактів; Приклад: Рахуємо кількість даних у джерелах

*SELECT Count(documents1.id) AS [Count-id]*

*FROM documnets1;*

Додаємо отримані кількості, Рахуємо кількість вставлених у таблицю сховища записів (дата внесення цих записів така, як сьогоднішня)

*SELECT documents.real_date, Count(documnets.id) AS [Count-id]*

*FROM documents*

*GROUP BY documents.real_date*

*HAVING (documents.real_date=Date());*

Якщо суми однакові, то ETL пройшов без перешкод.

**Висновки.** У результаті застосування технології ETL в інтеграції даних створюється основа для єдиного інформаційного простору бізнесу, яка має на меті надати користувачеві узгоджену й надійну інформацію для забезпечення цілісності даних. Науковою новизною є використання методики ETL-технології у сховищах даних бізнесу для інтеграції даних з розрізнених джерел. Практична цінність полягає у розробленні алгоритму, бази даних та оперативного сховища даних.

## Література

1. Уайт К. (Colin White). Интеграция данных: использование технологий ETL, EAI и EII для создания интегрированной корпорации (Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise). Ноябрь 2005.
2. Интеграция данных - ключ к эффективным решениям (Data Integration: The Key to Effective Decisions).

3. Огляд технологій інтеграції інформаційних систем, 2006, http://www.microsoft.com/Ukraine/Government/Analytics/IntegrationTechnologies/Overview.msp.
4. Dan Linstedt. Data Vaulttm overview the next evolution in data modeling. – 2005, http://www.tdan.com/i021hy01.htm/.