УДК 004.65:004.678

**К.т.н., доц., доцент кафедри ПЗКС Саяпіна І.О.,**
**магістрант Шевченко Г.О.**

**Національний технічний університет України**
**«Київський політехнічний інститут імені Ігоря Сікорського»**

# МЕТОД АВТОМАТИЗАЦІЇ УПРАВЛІННЯ ДАНИМИ В ГЕТЕРОГЕННИХ БАЗАХ

## Abstract

**Inna Saiapina, Ph.D., Associate Professor; Hlib Shevchenko, student**
***Method for Automation of Data Management in Heterogeneous Databases***

*This paper presents an automated method for integrating and managing heterogeneous data. The approach formalizes syntactic and semantic transformations, deduplication, and real-time quality control using a unified target model. The method improves data consistency and reliability while reducing update latency and manual interventions in corporate and scientific environments.*

## Вступ

Проблема узгодження даних із різнорідних джерел є однією з ключових у сучасних інформаційних технологіях. У більшості організацій дані зберігаються у вигляді реляційних таблиць, XML-документів, JSON-структур, потокових логів або неструктурованих файлів. Такі дані утворюють гетерогенні бази, що потребують інтеграції та управління. Без автоматизації ці процеси потребують значних людських ресурсів і є джерелом помилок. Розроблення універсального методу автоматизації управління гетерогенними даними дозволить підвищити ефективність, якість і узгодженість інформації у складних обчислювальних середовищах.

## Постановка задачі

Метою роботи є створення методу автоматизації управління даними в гетерогенних базах, який забезпечує узгоджене об’єднання різнотипних джерел у єдину логічну структуру. Необхідно розробити алгоритмічну модель інтеграції, визначити критерії оцінки якості даних і способи автоматичного оновлення [1]. Задача передбачає мінімізацію часу інтеграції $\Delta T$, підвищення

комплексного показника якості Q та масштабованість системи для великих потоків даних [2].

**Аналіз існуючих підходів та алгоритмів**

На сучасному етапі автоматизація управління даними реалізується різними концепціями. Класичний підхід ETL (Extract–Transform–Load) полягає у вилученні, перетворенні та завантаженні даних у централізоване сховище. Його перевага — стабільність, недолік — жорстка структура та складність адаптації до змін джерел.

Підхід Data Lake передбачає зберігання великих обсягів неструктурованих даних без попередньої обробки, що полегшує масштабування, але призводить до «ефекту болотного озера» — втрати структурованості. Data Fabric надає рівень віртуалізації поверх наявних джерел, створюючи єдиний логічний шар доступу. Data Mesh розподіляє відповідальність за дані між командами-доменами, зосереджуючись на самоврядуванні, проте потребує складної координації [3].

Системи MDM (Master Data Management) підтримують еталонні сутності, узгоджуючи дублікати й конфлікти, але вимагають значної попередньої роботи аналітиків. Концепція Knowledge Graph дозволяє описувати зв'язки між сутностями, однак має високу вартість побудови. Federated Query — підхід, що об'єднує запити до різних джерел у реальному часі, забезпечуючи динамічну інтеграцію, проте він обмежений швидкодією мережевих транзакцій [4].

Аналіз показує, що жоден з підходів не забезпечує повної автоматизації інтеграції. Тому необхідний метод, який поєднує універсальність, масштабованість і автономність, забезпечуючи синтаксичне, семантичне та часово узгоджене управління даними.

**Пропонований метод**

Розроблений метод базується на формалізованій моделі інтеграції даних. Нехай множина джерел $S = \{s_1, s_2, \ldots, s_n\}$, де кожне $s_i$ має власну структуру $F_i$. Процес інтеграції описується трансформацією $\tau_i: F_i \rightarrow U$, де U — уніфікована модель. Послідовність етапів інтеграції формалізується як композиція функцій (1):

$$\Omega = \rho_i \circ \sigma_i \circ \nu_i \circ \varphi_i \circ \psi_i \qquad (1)$$

Кожна функція відповідає конкретній операції:
$\psi_i$ — синтаксичне узгодження (приведення структур до стандарту),
$\varphi_i$ — нормалізація одиниць вимірювання,
$\nu_i$ — семантичне зіставлення на основі онтологій,
$\sigma_i$ — дедуплікація записів за функцією подібності,
$\rho_i$ — агрегація результатів у спільну модель.
Функція подібності між двома записами визначається виразом (2):

$$\delta(d_i, d_j) = \frac{1-|A(d_i)\cap A(d_j)|}{|A(d_i)\cup A(d_j)|} \quad (2)$$

де $A(d_i)$ — множина атрибутів запису $d_i$. Якщо $\delta(d_i, d_j) < \varepsilon$, записи вважаються дублікатами.
Рівень узгодженості між схемами обчислюється за формулою (3):

$$C(S_i, S_j) = \frac{|M(S_i,S_j)|}{(|S_i|,|S_j|)} \quad (3)$$

де $M(S_i, S_j)$ — множина співставлених атрибутів.
Якість даних визначається інтегральним показником (4):

$$Q = \sum_{k=1}^{n} \quad w_k q_k \quad (4)$$

де $q_k$ — метрики (повнота κ, точність π, актуальність α, унікальність μ), а $w_k$ — їхні ваги.
Часова деградація даних описується функцією (5):

$$Q(t) = Q_0 e^{(-\lambda t)} \quad (5)$$

де λ — коефіцієнт старіння.
Покращення ефективності автоматизації розраховується як (6):

$$E = \left(\frac{Q_{auto}-Q_{manual}}{Q_{manual}}\right) \times 100\%, \quad (6)$$

що показує відсоткове зростання якості після впровадження системи.
Запропонований метод реалізується у вигляді модульної системи з компонентами збору, нормалізації, семантичного узгодження, контролю якості та моніторингу. Дані з джерел надходять у центральний модуль, де виконуються етапи $\psi_i$–$\rho_i$. Контроль якості здійснюється автоматично, а система моніторингу реєструє показники Q та E у реальному часі. Впровадження у запропонованій архітектурі системи дозволяє забезпечити стійкість і масштабованість процесів інтеграції [5].

**Результати експериментальних досліджень**

Моделювання методу проводилося на тестових наборах даних із різними форматами (CSV, XML, JSON). Результати показали скорочення часу оновлення ΔT на 5–8% і підвищення комплексного показника Q у середньому на 3-5%. Система залишалася стабільною при збільшенні кількості джерел до 20, що підтверджує її масштабованість. Порівняння з класичними ETL-процесами показало зменшення кількості помилок і дублікатів, а автоматичний контроль якості дозволив знизити частоту ручних втручань на понад 10-12%.

**Висновки**

Запропонований метод автоматизації управління гетерогенними даними забезпечує комплексне узгодження, очищення і контроль якості в реальному часі. Він поєднує формальні математичні моделі з гнучкою архітектурою, що дозволяє інтегрувати дані з різних джерел без ручної участі. Метод може бути впроваджений у системах корпоративного рівня, аналітичних центрах і наукових платформах для підвищення достовірності та оперативності даних.

**Література**


1. Batini, C.; Scannapieco, M. Data and Information Quality: Dimensions, Principles and Techniques [Text]. — Springer, 2017. — 492 p.
2. Gubbi, J.; Buyya, R. Internet of Things and Big Data Integration Frameworks [Text] / J. Gubbi, R. Buyya // Future Generation Computer Systems. — 2018. — Vol. 86. — P. 601 – 614.
3. Gartner Research. Data Fabric Architecture: A Unified Data Management Solution [Електронний ресурс] / Gartner Research. — 2021. — Режим доступу: https://www.gartner.com/en/data-analytics/topics/data-fabric?
4. Dehghani, Z. Data Mesh: Delivering Data-Driven Value at Scale [Text]. — Sebastopol: O'Reilly Media, 2022. — 350 p.
5. Hovorushchenko, T.; Medzatyi, D. Software Quality Forecasting Based on Data Integration Attributes [Text] / T. Hovorushchenko, D. Medzatyi // Journal of Intelligent & Fuzzy Systems. — 2023. — Vol. 44, № 3. — P. 3891 – 3905.