Issue 16, 2024

УДК 004.6

СИСТЕМА КРИТЕРІЇВ ОЦІНКИ ЯКОСТІ ДАНИХ В РОЗПОДІЛЕНИХ ІНФОРМАЦІЙНИХ СИСТЕМАХ

Юрій Геряк[1], Андрій Берко[2]

[1,2] Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж, Львів, Україна
[1] E-mail: yurii.m.heriak@lpnu.ua, ORCID: 0009-0008-3251-2007
[2] E-mail: andrii.y.berko@lpnu.ua, ORCID: 0000-0003-2892-9519

Автори розробили систему критеріїв оцінки якості даних в контексті розподілених інформаційних систем. Стаття описує набір вимірів якості даних, сформований на основі проблем зберігання та обробки даних у розподілених середовищах. Основна мета дослідження полягає в уточненні основних вимог та викликів, які виникають перед розподіленими інформаційними ресурсами, а також в їх відповідності вибраним критеріям якості даних. Здійснено вичерпний аналіз літературних джерел, що дозволило визначити ключові виміри якості даних, виявлені у більшості досліджень, включаючи повноту, точність, узгодженість та актуальність. Описано основні проблеми, що виникають при роботі з даними у розподілених інформаційних системах. На основі літературного огляду автори розробили підхід до сформування уніфікованого набору критеріїв оцінки якості даних, який включає точність, узгодженість, повноту, актуальність, доступність та інші специфічні властивості даних. Підкреслено, що критерії якості даних прямо залежать від призначення інформаційної системи та базуються на конкретних вимогах, тому дане рішення є лише мінімальним набором характеристик, за якими можна оцінити якість даних у розподілених інформаційних системах.

Ключові слова: оцінка якості даних, розподілені інформаційні системи, виміри якості даних, повнота, точність, узгодженість, актуальність.

Вступ та постановка проблеми

Сьогодні, в більшості застосувань інформаційних технологій значну роль відіграє ресурс даних, від можливостей та властивостей якого залежить правильність, достовірність і надійність результатів розв'язання визначених задач. Характерними рисами сучасних ресурсів даних є, насамперед, розподілене середовище опрацювання, значні обсяги, різноманіття структур даних, динаміка оновлень, відсутність централізовапних засобів контролю і керування. Особливою рисою інформаційних ресурсів розподілених систем є застосування не лише розподілених середовищ створення і підтримання таких даних, а й розподілених процесів їх опрацювання. Способи та механізми їх застосування суттєво відрізняються від традиційних баз даних, які використовують модель даних SQL та механізм виконання транзакцій ACID для підтримання ресурсу у стані, що відповідає наперед визначеному набору вимог. Рівень відповідності показників інформаційного ресурсу параметрам, які встановлено такими вимогами характеризує якість даних – їх здатність забезпечити ефективне виконання завдань інформаційної системи чи сервісу. Для розподілених, різнорідних, динамічних ресурсів даних застосування принципів і технологій SQL в управлінні якістю, загалом, є проблемним або неможливим.

В контексті цієї проблеми, побудова системи критеріїв оцінки якості даних набуває особливої актуальності, оскільки вона спрямована на розробку методологічних засад та інструментарію для об'єктивного визначення рівня достовірності, точності, повноти та інших вимог до даних, які опрацьовують в розподілених інформаційних системах. Забезпечення високої якості даних у таких системах є запорукою їх успішного функціонування, а також є важливою передумовою для прийняття обґрунтованих управлінських рішень та забезпечення надійності обміну інформацією між різними підсистемами.

В напрямі визначення та підтримання належного рівня показників якості даних, на відміну від програмного забезпечення, немає чітко визначених норм, стандартів чи регламентів. На практиці цю проблему вирішують виходячи з установлених практик, досвіду, бачення виконавців та специфіки проєктів. Це часто призводить до виникнення суперечностей, неузгодженості оцінок і, як наслідок, некоректного застосування тих чи інших даних та отримання ненадійних результатів.

Тому, дослідження та розробка системи критеріїв оцінки якості даних стає важливим напрямком наукових досліджень, спрямованим на вдосконалення функціональних можливостей розподілених інформаційних систем (РІС) та підвищення рівня їх ефективності. В даній роботі виконано завдання систематизування та аналізу відомих підходів до побудови критеріїв оцінки якості даних, а також розроблення нових засад та підходів розв'язання таких задач з урахуванням сучасних вимог до розподілених інформаційних систем.

Загалом, забезпечення належного рівня якості даних потребує розробки методів і моделей, які враховують специфіку роботи РІС, їх розподіленість, гетерогенність та складність. Недоліки в системі оцінки якості даних можуть призвести до серйозних наслідків, таких як невірне прийняття рішень, втрати даних або порушення безпеки. Отже, дослідження в напрямі розроблення системи критеріїв оцінки якості даних, яка б враховувала усі аспекти роботи РІС та сприяла підвищенню їхньої ефективності та надійності є достатньо актуальними.

Враховуючи ці чинники, загалом, проблему досліджень, викладених у роботі може бути сформульовано таки чином.

Мета досліджень: ефективне та продуктивне застосування ресурсів даних розподілених інформаційних систем у різноманітних сферах та застосуваннях.

Об'єкт досліджень: процеси управління якістю ресурсів даних розподілених інформаційних систем.

Предмет досліджень: розроблення методів і засобів оцінювання та контролю якості даних інформаційних ресурсів розподілених систем.

Завдання досліджень: побудова системи критеріїв оцінки якості даних в розподілених інформаційних системах.

Формулювання цілі статті

Мета даної роботи полягає у розробленні системи об’єктивних, достовірних та надійних критеріїв оцінки якості даних для розподілених інформаційних систем, що сприятиме підвищенню ефективності та надійності управління даними у таких середовищах. Основними завданнями, які забезпечують досягнення визначеної мети визначено, зокрема, такі.

- Аналіз та узагальнення результатів досліджень, публікацій і практики у галузі оцінки якості даних. Виконання цього завдання дасть змогу узагальнити досвід та практичні напрацювання у даному напрямі та сформувати бачення шляхів розв’язання задачі визначення критеріїв оцінки якості даних для розподілених інформаційних систем.
- Класифікація критеріїв оцінювання якості даних, які використовують у сьогоденній практиці побудови розподілених інформаційних систем. Розподіл критеріїв на попередньо визначені категорії забезпечить можливість розроблення окремих процедур для кожного типу задач оцінювання якості з врахуванням їх специфіки.

- Розроблення загальних принципів формування системи критеріїв якості ресурсів даних розподілених інформаційних систем. Використання таких принципів дасть змогу розробити методи і технології контролю якості ресурсів даних розподілених інформаційних систем.

Результатами виконання цих завдань є формалізована та структурована система критеріїв якості ресурсів даних, її теоретичне обґрунтування та визначення шляхів практичного застосування у процесах управління якістю даних на всіх етапах їх життєвого циклу розподілених інформаційних систем критерії.

Аналіз останніх досліджень та публікацій

Аналіз досліджень на тему забезпечення якості даних варто почати із першої згадки про якість інформаційного продукту, яка отримала місце серед дослідників у 1950 році. У результаті досліджень були впровадженні такі поняття, як “ступінь, до якої набір властивих характеристик відповідає вимогам” [2], “придатність до використання” [3], “відповідність вимогам” [4].

У міру швидкого розвитку інформаційних технологій дослідження перемістились у бік вивчення якості даних. У 1990-х роках вченими були запропоновані різні визначення поняття “якості даних” та різні набори параметрів оцінки такої якості. Група Total Data Quality Management в Університеті Массачусетського технологічного інституту визначила “якість даних”, як “придатність до використання” та підкреслила, що оцінка якості даних напряму залежить від їх споживачів [1].

Більш детально заглиблюються у багатовимірну природу якості даних і проблем, пов’язаних з її визначенням, вимірюванням та покращенням, автори статті “Data Quality at a Glance” М. Сканнап’єко, П. Місьє та К. Батіні [5]. Автори підкреслюють важливість розгляду різних параметрів якості, окрім простої точності, включаючи повноту, узгодженість і пов’язані з часом параметри, щоб забезпечити цілісне уявлення про якість даних. У контексті статті автори вважають особливо важливими та розглядають детальніше, з погляду специфіки вимірювання, наступні параметри якості даних:

Точність: забезпечення правильності даних і відсутності помилок, неточностей або друкарських помилок, що має вирішальне значення для прийняття обґрунтованих рішень і проведення надійного аналізу.

Повнота: Перевірка того, що всі необхідні поля даних заповнені та не втрачено жодної важливої інформації, оскільки неповні дані можуть призвести до прогалин в аналізі та процесах прийняття рішень.

Узгодженість: забезпечення узгодженості та узгодженості значень даних у різних джерелах даних або в межах одного набору даних, що важливо для підтримки цілісності та надійності даних.

Виміри, пов’язані з часом: Оцінка своєчасності та актуальності даних для процесів прийняття рішень, оскільки застарілі або нерелевантні дані можуть призвести до неправильних висновків і неефективних дій.

Наведено практичні приклади, щоб проілюструвати, як ці параметри можна виміряти та застосувати в різних контекстах, таких як електронний бізнес та електронний уряд.

У статті зазначено зростання потреб у вирішенні проблем якості даних, особливо в контексті інтеграції інформації з різних джерел даних, оскільки дані низької якості перешкоджають зусиллям інтеграції. Автори обговорюють вплив проблем якості даних на діяльність зі сховищ даних, де значна частина бюджету впровадження часто виділяється на завдання очищення даних. Вони також торкаються проблем, з якими стикаються під час інтеграції віртуальних даних, де неузгодженість даних, що зберігаються на різних сайтах, ускладнює надання інтегрованої інформації у відповідь на запити користувачів.

Хоча основні параметри якості даних, такі як точність, повнота, узгодженість і параметри, пов'язані з часом, широко визнані, у статті визнається, що не існує загального стандарту для визначення параметрів якості даних у різних пропозиціях. Відсутність консенсусу щодо визначень певних параметрів створює проблему для досягнення єдиного розуміння якості даних. Дослідницьке співтовариство продовжує досліджувати найкращі підходи до визначення та вимірювання параметрів якості даних, постійно намагаючись наблизитися до більш стандартизованої системи [5].

Яскравим представником досліджень на тему забезпечення якості даних є стаття "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era" авторів Лі Чаі та Янджоу Жу [1].

Стаття досліджує проблеми якості даних в епоху великих даних, наголошуючи на швидкому зростанні та складності глобальних даних. Вона підкреслює, як кількість інформації подвоювалася кожні кілька років після промислової революції, а глобальні дані досягли 1,8 ZB у 2011 році, що створювало труднощі під час збору, очищення та інтеграції високоякісних даних. Експоненціальне збільшення обсягу даних у поєднанні з переважанням неструктурованих даних у великих даних створює значні проблеми для ефективного перетворення та опрацювання даних. Відсутність програмного забезпечення для опрацювання та аналізу великих даних у реальному часі ще більше ускладнює своєчасне вилучення цінної інформації з даних, що швидко змінюються.

З погляду критеріїв якості, у статті обговорюються 4V великих даних – обсяг, швидкість, різноманітність і цінність – які сприяють унікальним проблемам, з якими стикаються під час забезпечення якості даних. Різноманітні джерела великих даних, включаючи дані Інтернету, дані Інтернету речей і наукові дані, призводять до складних структур даних і ускладнюють інтеграцію для підприємств. У статті наголошується на необхідності ієрархічного стандарту якості даних з точки зору користувачів, щоб вирішити змінний характер споживання даних в епоху великих даних (рис. 1).

Він також підкреслює важливість стандартів якості даних і методологій для забезпечення відповідності даних відповідним цілям використання та вимогам, враховуючи динамічне бізнес-середовище та різноманітні джерела даних.

Автори вважають наступні параметри якості даних найважливішими в контексті великих даних:

Доступність: цей параметр стосується ступеня зручності для користувачів отримувати дані та пов'язану інформацію. У контексті великих даних забезпечення доступності даних має вирішальне значення через величезний обсяг джерел даних і необхідність легкого доступу до відповідної інформації для аналізу та прийняття рішень.

Рис. 1. Структура ієрархічного стандарту якості даних

Зручність використання: зручність використання стосується того, чи корисні дані та відповідають потребам користувачів. У сфері великих даних забезпечення зручності використання даних має важливе значення для ефективного отримання цінної інформації з різноманітних джерел і типів даних, підвищення ефективності аналізу даних і процесів прийняття рішень.

Надійність: Надійність зосереджена на тому, чи можна довіряти даним, охоплюючи такі елементи, як точність, послідовність, повнота та можливість перевірки. У контексті великих даних забезпечення надійності даних має першочергове значення для підтримки цілісності та достовірності даних, особливо під час роботи з масивними та різноманітними наборами даних.

Релевантність: Релевантність описує ступінь кореляції між вмістом даних і очікуваннями чи вимогами користувачів. В епоху великих даних підкреслення актуальності даних має вирішальне значення для забезпечення того, щоб дані відповідали конкретним потребам і вимогам користувачів, уможливлюючи змістовний аналіз і прийняття рішень на основі відповідної інформації.

Якість презентації: Якість презентації стосується дійсності методу опису даних, що дозволяє користувачам повністю зрозуміти дані. У контексті великих даних забезпечення високої якості презентації має важливе значення для ефективної передачі інформації про складні дані та висновків зацікавленим сторонам, сприяючи процесам прийняття обґрунтованих рішень.

Ці параметри відіграють вирішальну роль у вирішенні завдань ефективного управління та аналізу великих даних, підкреслюючи важливість доступності даних, зручності використання, надійності, актуальності та якості представлення для забезпечення високоякісного керування та використання даних в епоху великих даних.

Загалом у статті підкреслюється критична важливість вирішення проблем із якістю даних у контексті великих даних, де величезний обсяг, швидкість і різноманітність даних створюють безпрецедентні перешкоди для організацій. Зосереджуючись на критеріях якості та стандартах, адаптованих до унікальних характеристик великих даних, дослідники та практики можуть ефективно орієнтуватися в складнощах обробки та аналізу даних у епоху цифрових технологій [1].

У статті “Data Quality Assessment” авторів Лео Л. Піпіно, Ян В. Лі та Річард Ю. Ван досліджується значення створення практичних метрик якості даних для організацій. Наразі метрики якості даних часто є випадковими із відсутніми основними принципами контролю якості даних. Якість даних визнається як багатовимірний концепт, що включає суб’єктивні сприйняття та об’єктивні вимірювання. У компанії Global Consumer Goods, Inc. (GCG) суб’єктивні оцінки відзначили проблеми з узгодженістю та повнотою, які були підтверджені об’єктивними оцінками. GCG використовувала метрику для аналізу цілісності стовпців для покращення якості даних. Аналіз суб’єктивних та об’єктивних оцінок розділений на чотири квадранти, маючи на меті досягнення оптимальної якості даних у Квадранті IV. Досвід GCG підкреслює важливість дослідження кореневих причин та прийняття виправних заходів на основі результатів оцінки.

Стаття вводить три функціональні форми для розробки об’єктивних метрик якості даних і наголошує на необхідності поєднання суб’єктивних та об’єктивних оцінок. Дані та інформація розрізняються, при цьому інформація визначається як оброблені дані. Підхід, описаний у статті, ефективно поєднує суб’єктивні та об’єктивні оцінки. Для організацій важливо систематично контролювати метрики якості даних і використовувати порівняльні оцінки протягом часу. Стаття робить висновок, що підхід “один розмір підходить для всіх” до метрик якості даних є неефективним і вимагає постійних зусиль. Розуміння основних принципів є важливим для розробки суб’єктивних та об’єктивних метрик якості даних. Стаття наводить ілюстративні метрики для ключових вимірів якості даних та показує, як цей підхід може бути реалізований на практиці.

Раніше було визначено, що якість даних є не одновимірною концепцією, а багатовимірною [7] [8]. Загальновідомі метрики якості даних включають доступність, обсяг даних, вірогідність, повноту, зрозуміле представлення, послідовне представлення, зручність у маніпулюванні, відсутність помилок, інтерпретованість, об’єктивність, актуальність, репутацію, безпеку, своєчас-

ність та зрозумілість [8] [9]. У своїх напрацюваннях, [10], винесли на загал думку, що критерії якості даних можна комбінувати між собою, конструюючи таким чином набір характеристик даних, які будуть важливими для конкретного завдання або вимоги користувача. Виміри якості даних вивчаються трьома відомим на сьогодні способами: інтуїтивним, теоретичним та емпіричним [3].

Інтуїтивний підхід до вивчення вимірів якості даних ґрунтується на особистому досвіді дослідників та їхньому інтуїтивному розумінні сутності цих вимірів [3]. Замість використання формальних методів чи стандартів, дослідники використовують свої знання та експертну інтуїцію для визначення важливих аспектів якості даних. Цей підхід дозволяє дослідникам ідентифікувати та вибирати ті виміри якості даних, які вони вважають найбільш важливими в контексті своїх досліджень [7]; [11]; [12]; [13]; [14]; [15]. Такий підхід особливо корисний у випадках, коли немає загальноприйнятих стандартів або коли виміри якості даних потрібно налаштовувати під конкретні потреби дослідження. Підхід також дозволяє враховувати унікальні аспекти кожного дослідження та адаптувати виміри якості даних з урахуванням конкретного контексту [10]. Таким чином, інтуїтивний підхід надає гнучкість та можливість врахування індивідуальних особливостей кожного дослідження при визначенні вимірів якості даних.

Теоретичний підхід до вивчення вимірів якості даних базується на розумінні процесу породження даних та їхньої цінності для споживачів [16]. Згідно з цим підходом, інформаційні системи розглядаються як системи виробництва даних, що обробляють вхідні дані для створення вихідних даних або продуктів даних [16]. Даний підхід підкреслює значення розглядання даних як продукту, який має цінність для користувачів, незалежно від того, чи є ці користувачі в межах підприємства, чи зовнішніми споживачами [16]; [17].

За теоретичним підходом, якість даних визначається споживачами, що використовують ці дані [18]. Тобто, якість даних визначається на основі фактичного використання даних, і відсутність неоднозначності та спірних питань, які можуть виникнути при їхньому використанні, є важливими аспектами цього підходу [18]. Відповідно до цього підходу, аналіз вимірів якості даних повинен ґрунтуватися на припущеннях щодо того, як інформаційні системи відображають реальні системи та як користувачі можуть інтерпретувати цю інформацію для прийняття рішень [18]. Такий підхід стимулює дослідників розробляти внутрішні виміри якості даних, які відображають властивості даних, такі як повнота, однозначність, змістовність та правильність [18].

Теоретичний підхід надає можливість розуміти якість даних з точки зору їхньої цінності для користувачів та розвивати більш детальні та повні підходи до вимірювання та забезпечення якості даних.

Емпіричний підхід до вивчення вимірів якості даних зосереджується на аналізі якості даних з точки зору їхніх споживачів. Одним із ключових принципів цього підходу є переконання в тому, що якість продукту даних визначається його споживачами [3]. Представницьким дослідженням у цій області був проведений Вангом та Стронгом (1996), які визначили виміри та оцінку якості даних, збираючи інформацію від користувачів даних [3].

Згідно з емпіричним підходом, дані мають якість самі по собі, що відображено у внутрішній якості даних. Один із чотирьох вимірів цієї категорії – точність. Контекстуальна якість даних підкреслює необхідність розглядати якість даних як складову частину завдання, забезпечуючи їхню актуальність, своєчасність, повноту та прийнятний обсяг для надання цінності. Значення систем виокремлене у представницькій та доступній якості даних.

Щоб бути ефективним, інформаційна система повинна відображати дані у формі, зрозумілій, легко сприйманій та однаково вираженій [3]; [19]; [20]. Зазначене дослідження аргументує, що попередня концептуальна рамка для якості даних повинна включати чотири аспекти: доступність, інтерпретованість, актуальність та точність. Вони також докладно вдосконалили свою модель, визначивши чотири виміри якості даних, які містять внутрішню, контекстуальну, представницьку та доступну якість даних [3].

Враховуючи вище сказане, можна зробити висновок, що інтуїтивні та теоретичні підходи до дослідження вимірів якості даних, хоч і корисні, але не враховують користувача, який є основним суддею якості даних. Емпіричний підхід, хоча й має свої обмеження, забезпечує більш об'єктивну оцінку якості даних з погляду їх користувачів. Проте необхідно враховувати, що емпіричний підхід може бути складним у доведенні повноти та точності результатів через відсутність фундаментальних принципів.

Варто визначити, що комплексне порівняння підходів дослідження вимірів якості даних може допомогти в розумінні їх ефективності та застосуванні в практиці. Автори Ю. Ван [16] та Р. Прифти з Г. Алімехметі [21] вказують на ці обмеження та потребу подальших досліджень у цьому напрямку.

Проведений огляд публікацій і досліджень з тематики якості даних дає змогу зробити такі висновки:

- сьогодні у напрямі розроблення інформаційних систем різного типу і спрямування не вироблено сталої системи норм і стандартів оцінювання і контролю якості ресурсів даних;
- процеси і процедури управління якістю даних в інформаційних системах будують як правило на основі усталених практик досвіду та внутрішньокорпоративних регламентах та правилах;
- вироблення уніфікованого підходу до побудови системи критеріїв оцінювання, що ґрунтується на передовому досвіді, науковому підґрунті та кращих практиках дасть змогу спростити процеси та процедури управління якістю ресурсів даних та підвищити їх ефективність.

Основні результати дослідження

Одним з результатів літературного огляду наявних досліджень у сфері визначення критеріїв якості даних є статистична показники, щодо частоти використання конкретних показників якості у публікаціях. Графік частотного розподілу таких показників подано на рис. 2.

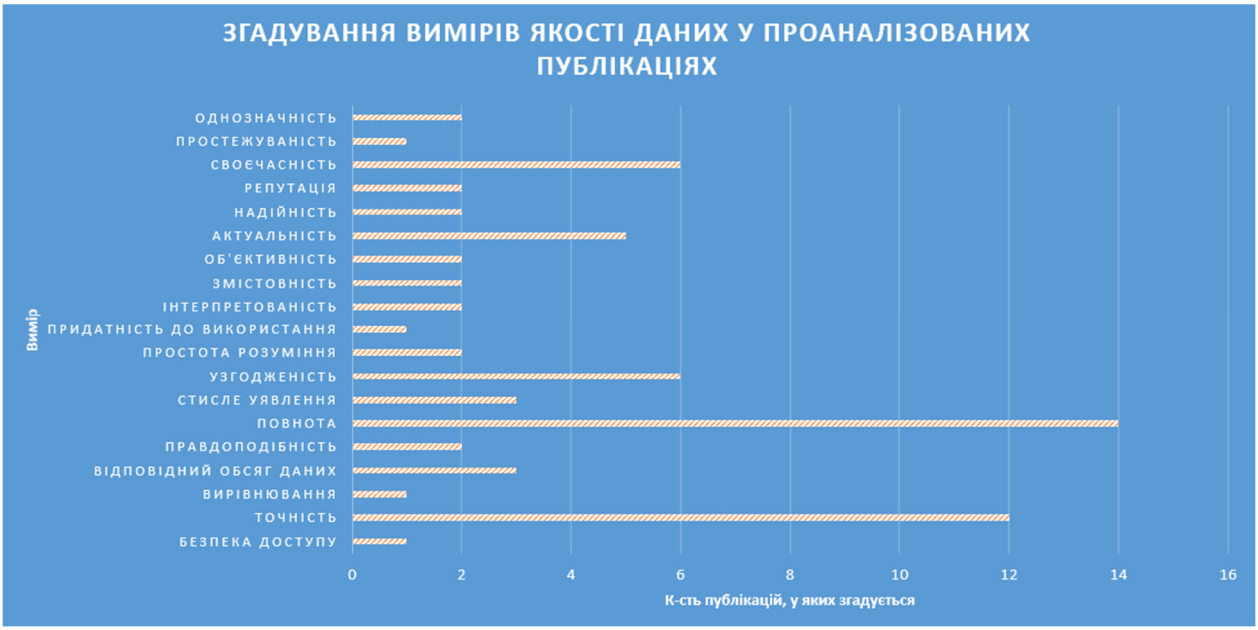

Рис. 2. Частота згадування вимірів даних у наукових публікаціях

Як показує аналіз частотності, серед усіх показників якості можна виділити п'ять найбільш популярних: *повнота*, *точність*, *своєчасність*, *узгодженість та актуальність*. У розподілених

інформаційних системах ці показники якості мають певну специфіку, обумовлену як розподіленим середовищем формування та зберігання, так і розподіленими процесами їх опрацювання і застосування [22].

Повнота (Completeness). У контексті розподілених інформаційних систем, повнота даних виступає як ключовий вимір якості, що визначається наявністю та доступністю всієї необхідної інформації для задоволення потреб користувачів та виконання бізнес-процесів. Важливість повноти полягає у забезпеченні того, щоб усі аспекти та атрибути даних, необхідні для прийняття рішень та аналітичних операцій, були присутні та доступні у всій системі. Недоліки у повноті даних можуть призвести до виникнення прогалин у знаннях або недостовірних результатів аналізу, що у свою чергу може підірвати довіру до системи та призвести до прийняття неправильних рішень.

У зв'язку з тим, що розподілені інформаційні системи мають складну структуру та велику кількість джерел даних, забезпечення повноти вимагає встановлення ефективних механізмів збору, інтеграції та реплікації даних, а також розробки стратегій забезпечення доступності інформації у всіх частинах системи. Цей вимір якості має критичне значення для забезпечення об'єктивності та компетентності при прийнятті рішень на основі даних та для успішного функціонування розподілених інформаційних систем.

Точність (Accuracy). Є іншою ключовою вимогою якості, яка характеризує міру відповідності між даними та показниками у реальності. Помилки у процесах застосування і перетворення можуть призвести до неточностей, що знижують цінність результатів та ефект застосування даних. Для управління якістю даних важливим є дотримання визначених мір відповідності та похибок точності. Особливістю контролю точності даних в розподілених середовищах є ризики використання різних мір і оцінок точності в різних сегментах розподіленого середовища. Це може призводити до суперечностей та конфліктів у процесах опрацювання даних, що негативно відображається на точності кінцевих результатів.

Узгодженість (Consistency). У розподілених інформаційних системах, узгодженість є критичним виміром якості даних, оскільки вона визначає ступінь єдності та спільності інформації між різними джерелами, базами даних або вузлами системи. У контексті розподілених середовищ, де дані можуть бути реплікованими та зберігатися на різних вузлах, необхідно забезпечити, щоб усі екземпляри даних були узгодженими і мали однакове значення. Узгодженість даних сприяє уникненню конфліктів, розбіжностей та невідповідностей між різними джерелами. Це особливо важливо для забезпечення однозначного та правильного розуміння стану даних і прийняття коректних рішень на основі цих даних.

Забезпечення узгодженості вимагає ретельного контролю за процесами реплікації, синхронізації та консистентності даних усередині розподілених систем. Відсутність узгодженості може призвести до виникнення проблем виконання операцій та аналізу даних, що, в свою чергу, може призвести до спотворення результатів та втрати довіри до системи в цілому. Узгодженість даних у розподілених інформаційних системах є ключовою для забезпечення надійності та цілісності даних, що є фундаментальним для успішної експлуатації системи та досягнення бізнесових цілей.

Своєчасність (Timeliness). Вимого своєчасності даних в інформаційних системах полягає насамперед, у швидкості і регулярності надходження, оновлення та забезпечення доступу до них для користувачів та процесів. Своєчасність є критично важливим показником, оскільки затримки чи перешкоди у доступі до актуальних даних можуть призводити до неефективних дій чи неправильних результатів.

Проблема своєчасності в розподілених інформаційних системах є особливо актуальною і складною, оскільки надходження та оновлення даних може відбуватись на різних вузлах мережі з можливою затримкою доступності. Це вимагає застосування спеціальних механізмів синхронізації і реплікації.

Актуальність (Currentness). У контексті розподілених інформаційних систем, актуальність даних виступає як ключовий вимір якості, який визначається актуальністю інформації з моменту її збору чи оновлення до моменту використання або аналізу. Важливість актуальності полягає в

забезпеченні того, щоб дані відображали поточний стан справ та відповідали реальному часові. Невідповідність даних актуальності може призвести до прийняття неправильних рішень або зниження ефективності бізнес-процесів, оскільки вони ґрунтуються на застарілих чи неправдивих даних.

Для забезпечення актуальності важливо мати ефективні механізми моніторингу та оновлення даних у реальному часі, а також стратегії управління життєвим циклом інформації від збору до використання. Крім того, у розподілених системах, де дані можуть бути реплікованими та оновлюватися на різних вузлах, необхідно також забезпечувати синхронізацію даних між усіма компонентами системи для збереження їх актуальності та уникнення конфліктів. Цей вимір якості є ключовим для забезпечення надійності та ефективності розподілених інформаційних систем, особливо в областях, де швидке прийняття рішень та реакція на зміни є критичними.

Загалом, перелічені показники можна вважати *базисним набором*, який є характерним для різних класів розподілених інформаційних систем. Але окрім такого універсального набору показників, у процесах розроблення та застосування розподілених інформаційних систем доцільним є визначення певних специфічних вимог щодо якості ресурсів даних. Метою таких вимог є врахування особливостей предметної області, процесів які в ній виконують, змісту завдань інформаційної системи, характеру, складу та форми кінцевих результатів, вимог і потреб користувачів.

Виходячи з такого бачення, повна система критеріїв якості даних розподілених інформаційних систем буде складатись з критеріїв трьох типів:

- *загальні (базові)* – критерії які може бути застосовано для ресурсів даних різного спрямування і застосування, що є обов'язковими для всіх категорій розподілених інформаційних систем;
- *спеціальні* – критерії, які визначають виходячи з особливостей предметної області, категорії задач, змісту і специфіки ресурсів даних, особливостей виконання процесів в середовищі інформаційної системи;
- *додаткові* – критерії, які враховують вимоги користувачів, характеристики платформи реалізації інформаційної системи, режим формування і використання результатів тощо.

Таким чином систему критеріїв якості даних конкретної розподіленої інформаційної системи можна подати у вигляді формальної моделі

$$Q=<Q^g, Q^s, Q^a>,$$

де, Q – множина критеріїв якості, Q^g – множина загальних критеріїв, Q^g – множина спеціальних критеріїв, Q^a – множина додаткових критеріїв, елементом кожної з таких множин є деякий предикат виду $q^c{}_i(d_j)$, який формулює вимогу виду c щодо деякого набору даних d_j зі складу ресурсу даних інформаційної системи. Кожен такий предикат формулюють шляхом застосування вимірів відповідності визначеного елемента даних змісту вимоги, визначеної щодо елемента даних такого типу. У процесі застосування, в залежності від виконання чи невиконання такої вимоги предикат приймає одне з можливих значень – *True/False*, на основі якого виробляються рішення щодо дій над елементом даних.

В загальному вигляді, процес формування системи критеріїв якості даних може бути описано відображенням виду

$$Q:R \rightarrow D$$

деякого набору вимог якості $R=\{r_i\}$ на множину елементів даних деякого інформаційного ресурсу розподіленої системи $D=\{ d_j{}^j\}$, при цьому $R=\{R^g, R^s, R^a\}$, де R^g – загальні вимоги якості даних, R^s – спеціальні вимоги, R^a – додаткові вимоги якості. Така модель дає змогу звести процес формування критеріїв якості даних до виконання наступної послідовності кроків (рис. 3):

1. визначення набору загальних вимог щодо якості даних розподіленої інформаційної системи;

2. визначення набору спеціальних вимог щодо якості даних розподіленої інформаційної системи;

3. визначення набору додаткових вимог щодо якості даних розподіленої інформаційної системи;

4. формування набору критеріїв якості даних у складі трьох наборів – загальних, спеціальних,

додаткових;

5. подання сформованого набору критеріїв у визначеному форматі;

6. передавання для використання засобами управління якістю даних.

Рис. 3. Алгоритм формування системи критеріїв якості даних

Результатом виконання описаних кроків алгоритму є формування повного і узгодженого набору критеріїв оцінювання якості даних, які може бути застосовано в системах управління якістю інформаційного ресурсу розподіленої системи на всіх етапах його життєвого циклу – формування, підтримання, застосування.

Висновки

У статті викладено систематичний підхід до визначення та оцінки якості даних у розподілених середовищах. В ході написання було виявлено ключові аспекти, які необхідно враховувати при розробці системи критеріїв оцінки якості даних. Було проаналізовано проблеми, які виникають у розподілених інформаційних системах, і визначено їх вплив на якість даних.

Визначено шляхи і способи формування уніфікованого набору критеріїв оцінки якості даних, що дає змогу створення та запровадження засобів автоматизації управління якістю даних. Це може бути корисним як для практиків, які розробляють розподілені інформаційні системи, так і для науковців, які вивчають цю тему.

Результати роботи може бути використана для оцінки якості даних у РІС; розроблення методів та інструментів для очищення та покращення якості даних; підвищення ефективності та надійності РІС.

Подальші дослідження буде спрямовано на розроблення методів та інструментів для практичної реалізації підтримання якості даних у РІС з урахуванням запропонованої системи критеріїв.

Очікується, що розробка та апробація системи критеріїв оцінки якості даних для РІС дозволить підвищити ефективність та надійність РІС через покращення якості даних; удосконалити процеси прийняття рішень на основі даних; зменшити ризики, пов’язані з використанням неякісних даних.

Список літератури

1. *Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, 2. https://doi.org/10.5334/dsj-2015-002*

2. *General Administration of Quality Supervision (2008) Inspection and Quarantine of the People’s Republic of China. Quality management systems-Fundamentals and vocabulary (GB/T19000–2008/ISO9000:2005), Beijing*

3. *Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099.*

4. *Crosby, P. B. (1980). Quality is free: The art of making quality certain. New American Library.*

5. *Scannapieco M., Missier P., Batini C. (2005) Data Quality at a Glance.*

6. *Геряк, Ю. М., & Берко, А. Ю. (2024). Проблеми контролю якості даних в розподілених інформаційних системах. У Стан, досягнення та перспективи інформаційних систем і технологій (с. 98–100). Видавництво ОНТУ.*

7. *Ballou, D. P., & Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. Management Science, 31(2), 150–162. https://doi.org/10.1287/mnsc.31.2.150*

8. *Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211–218. https://doi.org/10.1145/505248.506010.*

9. *Abdouli, M., & Omri, A. (2021). Exploring the nexus among FDI infows, environmental quality, human capital, and economic growth in the Mediterranean region. Journal of the Knowledge Economy, 12(2), 788–810.*

10. *Cho, S., Weng, C., Kahn, M. G., & Natarajan, K. (2021). Identifying Data Quality Dimensions for Person-Generated Wearable Device Data: Multi-Method Study. JMIR mHealth and uHealth, 9(12), Стаття e31618. https://doi.org/10.2196/31618.*

11. *Bailey, J. E., & Pearson, S. W. (1983). Development of a Tool for Measuring and Analyzing Computer User Satisfaction. Management Science, 29(5), 530–545. https://doi.org/10.1287/mnsc.29.5.530.*

12. DeLone, W. H., & McLean, E. R. (1992). Information Systems Success: The Quest for the Dependent Variable. Information Systems Research, 3(1), 60–95. https://doi.org/10.1287/isre.3.1.60.

13. Ives, B., Olson, M. H., & Baroudi, J. J. (1983). The measurement of user information satisfaction. Communications of the ACM, 26(10), 785–793.

14. Laudon, K. C. (1986). Data quality and due process in large interorganizational record systems. Communications of the ACM, 29(1), 4–11. https://doi.org/10.1145/5465.5466.

15. Morey, R. C. (1982). Estimating and improving the quality of information in a MIS. Communications of the ACM, 25(5), 337–342. https://doi.org/10.1145/358506.358520.

16. Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. IEEE Transactions on Knowledge and Data Engineering, 7(4), 623–640.

17. Feki, C., & Mnif, S. (2016). Entrepreneurship, Technological Innovation, and Economic Growth: Empirical Analysis of Panel Data. Journal of the Knowledge Economy, 7(4), 984–999. https://doi.org/10.1007/s13132-016-0413-5.

18. Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), 86–95. https://doi.org/10.1145/240455.240479.

19. Ghasemaghaei, M., Ebrahimi, S., & Hassanein, K. (2018). Data analytics competency for improving firm decision making performance. The Journal of Strategic Information Systems, 27(1), 101–113. https://doi.org/10.1016/j.jsis.2017.10.001.

20. Ouechtati, I. (2022). Financial inclusion, institutional quality, and inequality: An empirical analysis. Journal of the Knowledge Economy, 1–25. https://doi.org/10.1007/s13132-022-00909-y.

21. Prifti, R., & Alimehmeti, G. (2017). Market orientation, innovation, and firm performance—an analysis of Albanian firms. Journal of Innovation and Entrepreneurship, 6(1). https://doi.org/10.1186/s13731-017-0069-9.

22. Berko, A. , Alieksieiev, V. , Lytvyn, V. (2018). Knowledge-based big data cleanup method. CEUR Workshop Proceedings, 2019, 2386, pp. 96–106.

References

1. Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, 2. https://doi.org/10.5334/dsj-2015-002

2. General Administration of Quality Supervision (2008) Inspection and Quarantine of the People's Republic of China. Quality management systems-Fundamentals and vocabulary (GB/T19000—2008/ISO9000:2005), Beijing

3. Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099.

4. Crosby, P. B. (1980). Quality is free: The art of making quality certain. New American Library.

5. Scannapieco M., Missier P., Batini C. (2005) Data Quality at a Glance.

6. Heriak, Y. M., & Berko, A. Y. (2024). Problems of data quality control in distributed information systems. In Status, achievements and prospects of information systems and technologies (p. 98–100). ONTU Publishing House..

7. Ballou, D. P., & Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. Management Science, 31(2), 150–162. https://doi.org/10.1287/mnsc.31.2.150

8. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211–218. https://doi.org/10.1145/505248.506010.

9. Abdouli, M., & Omri, A. (2021). Exploring the nexus among FDI infows, environmental quality, human capital, and economic growth in the Mediterranean region. Journal of the Knowledge Economy, 12(2), 788–810.

10. Cho, S., Weng, C., Kahn, M. G., & Natarajan, K. (2021). Identifying Data Quality Dimensions for Person-Generated Wearable Device Data: Multi-Method Study. JMIR mHealth and uHealth, 9(12), Стаття e31618. https://doi.org/10.2196/31618.

11. Bailey, J. E., & Pearson, S. W. (1983). Development of a Tool for Measuring and Analyzing Computer User Satisfaction. Management Science, 29(5), 530–545. https://doi.org/10.1287/mnsc.29.5.530.

12. DeLone, W. H., & McLean, E. R. (1992). Information Systems Success: The Quest for the Dependent Variable. Information Systems Research, 3(1), 60–95. https://doi.org/10.1287/isre.3.1.60.

13. Ives, B., Olson, M. H., & Baroudi, J. J. (1983). The measurement of user information satisfaction. Communications of the ACM, 26(10), 785–793.

14. Laudon, K. C. (1986). Data quality and due process in large interorganizational record systems. Communications of the ACM, 29(1), 4–11. https://doi.org/10.1145/5465.5466.

15. Morey, R. C. (1982). Estimating and improving the quality of information in a MIS. Communications of the ACM, 25(5), 337–342. https://doi.org/10.1145/358506.358520.

16. Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. IEEE Transactions on Knowledge and Data Engineering, 7(4), 623–640.

17. Feki, C., & Mnif, S. (2016). Entrepreneurship, Technological Innovation, and Economic Growth: Empirical Analysis of Panel Data. Journal of the Knowledge Economy, 7(4), 984–999. https://doi.org/10.1007/s13132-016-0413-5.

18. Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), 86–95. https://doi.org/10.1145/240455.240479.

19. Ghasemaghaei, M., Ebrahimi, S., & Hassanein, K. (2018). Data analytics competency for improving firm decision making performance. The Journal of Strategic Information Systems, 27(1), 101–113. https://doi.org/10.1016/j.jsis.2017.10.001.

20. Ouechtati, I. (2022). Financial inclusion, institutional quality, and inequality: An empirical analysis. Journal of the Knowledge Economy, 1–25. https://doi.org/10.1007/s13132-022-00909-y.

21. Prifti, R., & Alimehmeti, G. (2017). Market orientation, innovation, and firm performance—an analysis of Albanian firms. Journal of Innovation and Entrepreneurship, 6(1). https://doi.org/10.1186/s13731-017-0069-9.

22. Berko, A. , Alieksieiev, V. , Lytvyn, V. (2018). Knowledge-based big data cleanup method. CEUR Workshop Proceedings, 2019, 2386, pp. 96–106

THE SYSTEM OF DATA QUALITY ASSESSMENT CRITERIA IN DISTRIBUTED INFORMATION SYSTEMS

Yurii Heriak[1], Andrii Berko[2]

[1,2] Lviv Polytechnic National University,
Department of Information Systems and Networks, Lviv, Ukraine
[1] E-mail: yurii.m.heriak@lpnu.ua, ORCID: 0009-0008-3251-2007
[2] E-mail: andrii.y.berko@lpnu.ua, ORCID: 0000-0003-2892-9519

The authors developed a system of criteria for assessing data quality in the context of distributed information systems. The article describes a set of data quality dimensions formulated based on the challenges of data storage and processing in distributed environments. The main objective of the research is to identify the primary requirements and challenges faced by distributed information resources and to satisfy them with specifically selected data quality criteria. A comprehensive analysis of the literature was conducted to identify key data quality dimensions commonly found in most studies. These dimensions include completeness, accuracy, consistency, and timeliness. The article also outlines the main problems encountered when working with data in distributed information systems. Considering the results of the literature review, an attempt was made to formulate a unified set of data quality assessment criteria, which includes accuracy, consistency, completeness, timeliness, accessibility, and other specific data features. Authors emphasize that data quality criteria depend directly on the purpose of the information system and are based on specific requirements. Therefore, this solution represents only a minimum set of characteristics for evaluating data quality in distributed information systems.

Keywords: data quality assessment, distributed information systems, data quality dimensions, completeness, accuracy, consistency, timeliness.