Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)
Кафедра комп'ютерних наук
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістр
(назва освітнього ступеня)

на тему: Методи та засоби оброблення даних з врахуванням вимог якості до них

Виконав(ла): студент(ка) 6 курсу, групи СНмн-61
спеціальності 122 Комп’ютерні науки
(шифр і назва спеціальності)

(підпис) Лотоцький Д.В. (прізвище та ініціали)

Керівник (підпис) Голотенко О.С. (прізвище та ініціали)

Нормоконтроль (підпис) Дуда О.М. (прізвище та ініціали)

Завідувач кафедри (підпис) Боднарчук І.О. (прізвище та ініціали)

Рецензент (підпис) (прізвище та ініціали)

Тернопіль
2026

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра комп'ютерних наук
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Боднарчук І.О.
(підпис) (прізвище та ініціали)
"13" квітн 2026 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня Магістр
(назва освітнього ступеня)

за спеціальністю 122 Комп'ютерні науки
(шифр і назва спеціальності)

студенту Лотоцький Дмитро Володимирович
(прізвище, ім'я, по батькові)

1. Тема роботи **Методи та засоби оброблення даних з врахуванням вимог якості до них.**

Керівник роботи **к.т.н., доц. Голотенко О.С.**
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від "10" березня 2026 року № 4/9-150

2. Термін подання студентом завершеної роботи 26 травня 2026 р.

3. Вихідні дані до роботи Літературні джерела з тематики роботи

4. Зміст роботи (перелік питань, які потрібно розробити) ВСТУП; 1 ОГЛЯД А ОГЛЯД ПРОБЛЕМИ ЯКОСТІ ДАНИХ; 1.1 Джерела та наслідки проблем якості даних; 1.2 Співвідношення понять «дані», «інформація» та «якість»; 1.3 Якість даних, як багатовимірне поняття; 1.4 Точність (Accuracy); 1.5 Повнота (Completeness); 1.6 Узгодженість (Consistency); 1.7 Своєчасність (Timeliness); 2 ОПИС ПРОЦЕСУ ОЦІНЮВАННЯ ЯКОСТІ ДАНИХ; 2.1 Управління якістю даних; 2.1.1 Профілювання даних; 2.1.2 Вимірювання якості даних; 2.1.3 Очищення даних; 2.1.4 Моніторинг якості даних; 2.2 Вимоги до інструментів якості даних; 2.3 Методологія дослідження та відбору інструментів; 2.4 Каталог вимог та стратегія оцінювання; 3 ОГЛЯД ІНСТРУМЕНТІВ ЯКОСТІ ДАНИХ; 3.1 Опис досліджених інструментів; 3.2 Порівняння можливостей профілювання даних; 3.3 Порівняння можливостей вимірювання якості даних; 3.4 Порівняння можливостей моніторингу якості даних; 3.5 Узагальнення результатів огляду; 4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ; ВИСНОВКИ; СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ; ДОДАТКИ

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Сенчишин В.С., к.т.н., доц. каф. МТ		
Безпека в надзвичайних ситуаціях	Теслюк В.М., проректор з адміністративно-господарської роботи та будівництва		

7. Дата видачі завдання 13 квітня 2026 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням до кваліфікаційної роботи	13.04.26-18.04.26	Виконано
2.	Підбір наукових джерел за темою роботи	19.04.26-21.04.26	Виконано
3.	Переклад та опрацювання наукових джерел	20.04.26-23.04.26	Виконано
	за темою кваліфікаційної роботи		
4.	Виконання дослідження щодо теми	24.04.26-10.05.26	Виконано
	кваліфікаційної роботи		
5.	Оформлення першого розділу	04.05.26-05.05.26	Виконано
6.	Оформлення другого розділу	05.05.26-13.05.26	Виконано
7.	Оформлення третього розділу	13.05.26-14.05.26	Виконано
8.	Виконання завдання до підрозділу "Охорона праці"	08.05.26-09.05.26	Виконано
9.	Виконання завдання до підрозділу "Безпека в	10.05.26-05.05.26	Виконано
	надзвичайних ситуаціях"		
10.	Оформлення кваліфікаційної роботи	05.05.26-13.05.26	Виконано
11.	Нормоконтроль	14.05.26-15.05.26	Виконано
12.	Перевірка на плагіат	16.05.2026	Виконано
13.	Попередній захист кваліфікаційної роботи	20.05.2026	Виконано
14.	Захист кваліфікаційної роботи	27.05.2026	

Студент _____ Лотоцький Д.В.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Голотенко О.С.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

"Методи та засоби оброблення даних з врахуванням вимог якості до них" // Кваліфікаційна робота освітнього рівня "Магістр" // Лотоцький Дмитро Володимирович // Тернопільський національний технічний університет ім. І. Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра комп'ютерних наук, група СНнм-61 // Тернопіль, 2026 // с. – 74, рис. – 14, табл. – 1, джерел – 36.

Ключові слова: якість даних, профілювання даних, метрики якості даних, моніторинг якості даних, огляд інструментів, управління даними

Робота присвячена систематичному огляду сучасних програмних інструментів вимірювання та моніторингу якості даних і дослідженню розриву між теоретичними концепціями в цій галузі та їх практичною реалізацією.

У роботі розглянуто теоретичні засади якості даних – визначення, основні виміри (точність, повнота, узгодженість, своєчасність) та метрики їх вимірювання. Описано процес управління якістю даних як циклічний процес, що охоплює профілювання, вимірювання, очищення та моніторинг даних.

Для досягнення мети проведено систематичний пошук, у результаті якого ідентифіковано ряд програмних інструментів якості даних. Зокрема це Informatica Data Quality, Experian Pandora, Apache Griffin, DataCleaner, MobyDQ, Oracle EDQ, SAS Data Quality, Talend Open Studio та інші.

Отримані результати свідчать про такі ключові висновки: більшість інструментів підтримують базове профілювання даних, тоді як багатостовпцеве профілювання та виявлення залежностей реалізовані лише в окремих рішеннях; жоден з оглянутих інструментів не реалізує повного спектра теоретично запропонованих метрик якості даних – натомість більшість пропонують механізм користувацьких бізнес-правил; моніторинг якості даних здебільшого є

преміальною платною функцією в комерційних продуктах, а інструменти з відкритим кодом, орієнтовані на моніторинг, позбавлені функцій профілювання.

Результати дослідження мають практичну цінність для фахівців з якості даних при виборі інструментів, а також для наукової спільноти як підґрунтя для розроблення практико-орієнтованих методологій вимірювання якості даних.

ANNOTATION

"Methods and Tools for Data Processing Considering Data Quality Requirements" // Master's degree qualification paper // Lototskyi Dmytro // Ternopil Ivan Puluj National Technical University, Faculty of Computer Information Systems and Software Engineering, Computer Science Department, group СНнм-61 // Ternopil, 2026 // p. – 74, fig. – 14, tables – 1, references – 36.

Key words: data quality, data profiling, data quality metrics, data quality monitoring, tool overview, data management

The work is devoted to a systematic review of modern software tools for measuring and monitoring data quality and to investigate the gap between theoretical concepts in this field and their practical implementation.

The paper considers the theoretical foundations of data quality - definitions, basic dimensions (accuracy, completeness, consistency, timeliness) and metrics for their measurement. The data quality management process is described as a cyclical process that includes profiling, measuring, cleaning and monitoring data.

To achieve the goal, a systematic search was conducted, as a result of which a number of software tools for data quality were identified. In particular, these are Informatica Data Quality, Experian Pandora, Apache Griffin, DataCleaner, MobyDQ, Oracle EDQ, SAS Data Quality, Talend Open Studio and others.

The results obtained indicate the following key conclusions: most tools support basic data profiling, while multi-column profiling and dependency detection are implemented only in individual solutions; none of the reviewed tools implements the full range of theoretically proposed data quality metrics – instead, most offer a mechanism for custom business rules; data quality monitoring is mostly a premium paid feature in commercial products, and open source tools focused on monitoring lack profiling features.

The results of the study have practical value for data quality professionals when choosing tools, as well as for the scientific community as a basis for developing practice-oriented methodologies for measuring data quality.

ЗМІСТ

ВСТУП

Актуальність задачі.

Якісні дані є ключовою передумовою для отримання достовірних та інтерпретованих результатів аналітики й слугують основою для обґрунтованих рішень, що приймаються на підставі даних. У сучасних інформаційних системах від якості даних безпосередньо залежить ефективність діяльності підприємств, надійність автоматизованих систем підтримки прийняття рішень, а також коректність роботи алгоритмів машинного навчання та штучного інтелекту.

Вимірювання якості даних (англ. data quality, DQ) є фундаментальним будівельним блоком для оцінювання релевантності рішень, що приймаються на основі даних. Такі рішення супроводжують повсякденне життя людини: це й машинні рішення в алгоритмах ранжування, і робота промислових роботів, і функціонування безпілотних автомобілів у новій галузі штучного інтелекту. Негативний вплив низькоякісних даних на рівень помилок моделей машинного навчання неодноразово демонструвався в наукових дослідженнях [9, 10]. Рішення, що приймаються людиною, також спираються на якісні дані: наприклад, рішення щодо просування або припинення виробництва певного продукту зазвичай ґрунтується на даних про продажі.

Попри очевидну кореляцію між якістю даних та якістю рішень, проблема залишається гострою. За даними міжнародних опитувань, переважна більшість керівників вищої ланки занепокоєні якістю даних своїх організацій, а збитки підприємств, спричинені низькою якістю даних, оцінюються в середньому у мільйони доларів на рік [18]. Таким чином, якість даних перестала бути лише питанням «гігієни» даних і перетворилася на критичний чинник операційної досконалості, що сприймається як одне з найбільших завдань корпоративного управління даними [20].

Водночас на практиці склалася парадоксальна ситуація. Згідно з низкою галузевих опитувань, значна частина компаній досі використовує електронні таблиці або прості рішення на кшталт баз даних для перевірки якості даних, а

їхнє оцінювання здебільшого виконується вручну та ситуативно, без довгострокової стратегії управління якістю даних. З огляду на постійне зростання обсягів даних, що підлягають обробці, існує очевидна потреба в інтенсивних дослідженнях, спрямованих на автоматизацію завдань управління якістю даних. Як зазначають фахівці, без автоматизації швидкість та обсяг даних швидко переважать навіть найбільш наполегливі зусилля з їх вимірювання [23].

Дослідження якості даних проводяться з 1980-х років, і відтоді якість даних найчастіше асоціюється з принципом «придатності для використання» (fitness for use), який відображає суб'єктивність та контекстну залежність цього поняття [5, 24]. Якість даних зазвичай розглядається як багатовимірне поняття, у якому окремі аспекти описуються вимірами якості даних (наприклад, точність, повнота, своєчасність), а ступінь виконання кожного виміру може бути кількісно оцінений за допомогою однієї або кількох метрик. Паралельно з науковими розробками з тих часів було створено широкий спектр комерційних, відкритих та академічних інструментів якості даних із різними фокусами. Однак, попри велику кількість публікацій, інструментів та концепцій, досі не завжди зрозуміло, як співвіднести теоретичні концепції (виміри й метрики) з їх практичною реалізацією (інструментами). Тому питання про те, як автоматично вимірювати й моніторити якість даних, дотепер не отримало достатньої відповіді [23].

Саме розрив між науковими дослідженнями якості даних та їх практичними реалізаціями в сучасних інструментах зумовлює актуальність цієї роботи.

Практична значущість дослідження полягає в наданні фахівцям з якості даних обґрунтованих орієнтирів для вибору інструментів під конкретні завдання, а також у систематизації функціональних вимог, що дозволяє об'єктивно порівнювати можливості різних рішень. Для наукової спільноти результати дослідження мають цінність як емпіричне підтвердження того, які з теоретично запропонованих концепцій справді знаходять застосування на практиці, а які залишаються переважно теоретичними. Це створює основу для перегляду

наявних підходів та формування більш практико-орієнтованих методологій вимірювання якості даних.

Мета роботи.

Таким чином метою роботи є детальне вивчення того, як концепції вимірювання та моніторингу якості даних реалізовані в сучасних програмних інструментах, а також виявлення розриву між теоретичними напрацюваннями у сфері якості даних та їх практичним втіленням.

Для досягнення поставленої мети в роботі сформульовано та розв'язано такі основні завдання:

1) розглянути та систематизувати теоретичні засади поняття якості даних, зокрема визначення, виміри та метрики якості даних, що використовуються в науковій літературі;

2) описати процес управління якістю даних як циклічний процес та виокремити його основні етапи – профілювання даних, вимірювання якості даних, очищення даних та безперервний моніторинг якості даних;

3) проаналізувати методологію систематичного пошуку та відбору інструментів якості даних, а також сформувати каталог вимог для їх оцінювання за трьома функціональними напрямами: профілювання даних, вимірювання якості даних у термінах метрик та автоматизований моніторинг якості даних;

4) виконати огляд та порівняльний аналіз сучасних інструментів якості даних щодо їх можливостей профілювання, вимірювання та моніторингу;

5) сформулювати висновки щодо поточного стану інструментів якості даних, виявити їх обмеження та окреслити потенціал для функціонального вдосконалення.

Об’єкт дослідження: є процеси вимірювання та моніторингу якості даних в інформаційних системах.

Предмет дослідження: є методи, метрики та програмні інструменти, призначені для автоматизованого вимірювання й моніторингу якості даних.

Структуру роботи побудовано таким чином. У першому розділі розкрито поняття якості даних, його визначення, виміри та метрики. Другий розділ

присвячено опису процесу управління якістю даних та методології оцінювання інструментів. У третьому розділі наведено огляд сучасних інструментів якості даних та порівняльний аналіз їх можливостей. У висновках узагальнено основні результати дослідження та окреслено напрями подальшої роботи.

Наукова новизна отриманих результатів.

Вперше сформовано структурований каталог із 43 вимог до інструментів якості даних, розподілених за трьома категоріями (профілювання, вимірювання, моніторинг), який дозволяє об'єктивно та відтворювано порівнювати функціональні можливості різних рішень.

Проведено найповніший на момент дослідження систематичний пошук інструментів якості даних, більшість із яких раніше не включалися до жодного наукового огляду, що дає актуальне уявлення про реальний стан ринку.

Вперше на основі практичного оцінювання 13 інструментів емпірично підтверджено, що теоретично запропоновані метрики якості даних (точність, повнота з агрегацією, своєчасність) майже не реалізовані в сучасних програмних продуктах, а виміри якості даних на практиці використовуються лише як концептуальне групування бізнес-правил.

Запропоновано чітке розмежування між "моніторингом даних" (перевірка правил) та "моніторингом якості даних" (безперервне вимірювання метрик у часі), що усуває термінологічну неоднозначність, характерну для попередніх досліджень.

Практичне значення отриманих результатів.

Результати порівняльного аналізу дозволяють фахівцям із якості даних обґрунтовано обирати інструмент під конкретний сценарій використання (профілювання, вимірювання або моніторинг) без необхідності самостійно тестувати кожне рішення.

Дослідження попереджає практиків про конкретні функціональні прогалини інструментів: відсутність загальновикористовуваних метрик точності без еталонного набору даних, обмеженість виявлення залежностей, недоступність моніторингу у безкоштовних версіях.

Сформований каталог із 43 вимог може безпосередньо використовуватися організаціями, як основа для технічного завдання при закупівлі або розробці власного інструменту якості даних.

Для практиків якості даних дослідження обґрунтовує доцільність зосередження не на абстрактних вимірах, а на безпосередньо вимірюваних аспектах (відсутні значення, дублікати, порушення бізнес-правил), що є реалістичнішою стратегією автоматизованого контролю якості.

Апробація результатів та особистий внесок здобувача.

Основні положення роботи доповідались, розглядались та обговорювались на науковій конференції Тернопільського національного технічного університету імені Івана Пулюя у тезах студентської науково-технічної конференції "Природничі та гуманітарні науки. Актуальні питання – 2026", яка проводилась у ТНТУ.

1 ОГЛЯД ПРОБЛЕМИ ЯКОСТІ ДАНИХ

Попри існування різних інтерпретацій, термін «якість даних» найчастіше описують як «придатність для використання» (fitness for use) [5, 24], що відображає високу суб'єктивність та контекстну залежність цього поняття. Якість інформації нерідко вживають як синонім якості даних. Хоча обидва терміни можна чітко розрізнити (адже «дані» позначають прості факти, а «інформація» описує розширення цих фактів контекстом і семантикою), у літературі з якості даних вони часто вживаються як взаємозамінні [22]. У цій роботі переважно використовується термін «якість даних», оскільки фокус зосереджено на обробці об'єктивно та автоматично отримуваних фактів, тобто на внутрішніх (intrinsic) характеристиках даних.

Якість даних традиційно розглядається як багатовимірне поняття, де кожен окремий вимір (dimension) відповідає певному аспекту якості даних. Протягом років було запропоновано широкий спектр вимірів та їх класифікацій [3, 24]. Незважаючи на інтенсивні дослідження та тривалу дискусію щодо вимірів якості даних, досі немає консенсусу щодо того, які саме виміри є визначальними для вимірювання якості даних [23]. Тому в подальшому розгляді ми зосереджуємося на чотирьох найчастіше використовуваних вимірах — точності, повноті, узгодженості та своєчасності.

1.1 Джерела та наслідки проблем якості даних

Проблеми якості даних виникають на різних етапах життєвого циклу даних – від їх збирання та введення до зберігання, інтеграції та використання. До типових джерел погіршення якості даних належать: помилки ручного введення даних оператором; неузгодженість форматів під час інтеграції даних з різних інформаційних систем; застарівання даних з плином часу; дублювання записів унаслідок повторного введення тих самих сутностей; порушення обмежень цілісності; а також помилки автоматичних процесів обробки та перетворення

даних. Через те, що дані постійно змінюються й оновлюються, навіть одноразово очищений набір даних з часом знову накопичує дефекти, що робить якість даних динамічною характеристикою, яку необхідно підтримувати безперервно (див. рис. 1.1).

Рисунок 1.1 – Джерела проблем з якістю даних на етапах їх життєвого циклу

Наслідки низької якості даних виявляються на кількох рівнях. На операційному рівні це призводить до помилок у бізнес-процесах, дублювання зусиль та зростання витрат на виправлення помилок. На тактичному рівні низька якість даних знижує довіру до аналітичних звітів та ускладнює прийняття обґрунтованих рішень. На стратегічному рівні вона підриває довіру до даних як до активу організації загалом. Особливої гостроти ця проблема набуває в контексті машинного навчання, де дефекти у навчальних даних безпосередньо погіршують точність та надійність побудованих моделей [9, 10]. Саме тому вимірювання якості даних розглядається як необхідна передумова надійної та достовірної аналітики.

Варто розрізняти внутрішні (intrinsic) та контекстні (contextual) аспекти якості даних. Внутрішні аспекти, як-от точність чи повнота, можна оцінити

об'єктивно, незалежно від конкретного завдання. Контекстні аспекти, навпаки, залежать від мети використання даних: ті самі дані можуть бути якісними для одного завдання й непридатними для іншого. Цей поділ безпосередньо пов'язаний з принципом «придатності для використання» та пояснює, чому повністю автоматизоване вимірювання можливе лише для частини аспектів якості даних, тоді як інші потребують участі експерта чи кінцевого користувача.

1.2 Співвідношення понять «дані», «інформація» та «якість»

Для коректного розуміння якості даних важливо розмежувати базові поняття. Дані (data) – це необроблені факти, представлені у формалізованому вигляді, придатному для зберігання та обробки. Інформація (information) – це дані, наділені контекстом та семантикою, що робить їх осмисленими для отримувача. Хоча в літературі терміни «якість даних» та «якість інформації» часто вживаються взаємозамінно [22], їх розмежування має практичне значення: вимірювання внутрішніх характеристик даних (наприклад, частки порожніх значень) може бути повністю автоматизованим, тоді як оцінювання якості інформації нерідко потребує врахування контексту та залучення користувача.

У дослідженнях якості даних усталилося розуміння того, що якість є не абсолютною, а відносною характеристикою. Один і той самий набір даних може вважатися високоякісним для одного застосування й непридатним для іншого. Саме тому центральним для всієї галузі залишається принцип «придатності для використання» (fitness for use), сформульований ще в ранніх роботах [5, 24]. Цей принцип наголошує, що якість даних слід оцінювати не саму по собі, а у зв'язку з конкретним завданням, для якого ці дані призначені. Подальше розгортання цього принципу в практичні метрики й становить основну складність, яку покликані подолати інструменти якості даних.

1.3 Якість даних, як багатовимірне поняття

Дослідник Piro [21] розрізняє «жорсткі виміри» (hard dimensions), до яких належать, зокрема, точність, повнота та своєчасність і які можна вимірювати об'єктивно за допомогою перевірочних процедур, та «м'які виміри» (soft dimensions), які можна оцінити лише за допомогою суб'єктивного оцінювання. Однак навіть об'єктивні перевірочні процедури вимагають попереднього суб'єктивного та предметно-залежного визначення об'єктів даних, які підлягають вимірюванню, щоб послідовно дотримуватися підходу «придатності для використання» [21].

У зв'язку з обговоренням вимірів якості даних часто наголошується, що для застосування цих вимірів на практиці необхідне визначення конкретних метрик якості даних. Метрика – це функція, яка відображає вимір якості у числове значення, що дозволяє інтерпретувати ступінь виконання відповідного виміру [13]. Згідно зі стандартом IEEE [13], метрика є формулою, що дає числове значення.

Метрику якості даних можна вимірювати на різних рівнях агрегації: на рівні значення (value-level), на рівні стовпця або атрибута (column/attribute-level), на рівні кортежу або запису (tuple/record-level), на рівні таблиці або відношення (table/relation-level), а також на рівні бази даних (database-level) [12]. Агрегацію можна виконати, наприклад, за допомогою зваженого середнього арифметичного результатів метрики, обчислених на попередньому рівні (наприклад, результати рівня запису використовуються для обчислення метрики на рівні таблиці) [11].

Дослідники Heinrich та інші [11] запропонували п'ять вимог до метрик якості даних, що забезпечують надійне прийняття рішень:

– наявність мінімального та максимального значень метрики (R1); інтервальне шкалювання значень метрики (R2);

– якість конфігураційних параметрів та визначення значень метрики (R3);

– коректну агрегацію значень метрики (R4);

– економічну ефективність метрики (R5).

Водночас інші дослідники стверджують, що для оцінювання корисності та валідності метрики якості даних потрібен загальніший підхід.

Розглянемо ці вимоги докладніше. Наявність мінімального та максимального значень (R1) забезпечує однозначну інтерпретацію крайніх станів якості: значення метрики повинно мати чітко визначені межі, що відповідають ідеальній та найгіршій якості. Інтервальне шкалювання (R2) означає, що однакові різниці у значеннях метрики повинні відповідати однаковим різницям у якості, що дозволяє коректно порівнювати результати. Якість конфігураційних параметрів (R3) вимагає, щоб параметри, які користувач задає під час налаштування метрики, мали зрозумілу інтерпретацію та обґрунтований вплив на результат. Коректна агрегація (R4) гарантує, що значення метрики, обчислені на нижчих рівнях (значення, запис), можна осмислено об'єднати у значення на вищих рівнях (таблиця, база даних). Економічна ефективність (R5) наголошує, що витрати на обчислення метрики не повинні перевищувати вигоду від отриманої інформації про якість. Як показує подальший аналіз, більшість реалізацій метрик у сучасних інструментах не задовольняють усіх цих вимог одночасно.

Далі розглянуто чотири найпоширеніші виміри якості даних разом з типовими метриками для їх обчислення. Наведений перелік метрик не є вичерпним, проте дає уявлення про дослідження, проведені в цій галузі, оскільки наявність таких або подібних метрик спостерігається під час оцінювання інструментів якості даних.

Окрім розглянутих далі чотирьох вимірів, у літературі описано й інші, серед яких унікальність (uniqueness – відсутність небажаних дублікатів), валідність (validity – відповідність значень визначеному домену чи формату), доступність (accessibility), інтерпретованість (interpretability), репутація джерела (reputation) та релевантність (relevancy). Класифікації вимірів також різняться: деякі дослідники групують виміри за категоріями внутрішньої, контекстної, репрезентаційної та доступнісної якості, інші – за способом вимірювання на

об'єктивні й суб'єктивні. Така різноманітність підходів є однією з причин відсутності єдиного стандартизованого переліку вимірів, що, своєю чергою, ускладнює зіставлення теоретичних концепцій з реалізаціями в інструментах.

1.4 Точність (Accuracy)

Хоча точність іноді описують як найважливіший вимір якості даних, існує низка різних її визначень [3]. У літературі з якості даних точність можна описати як близькість між інформаційною системою та тією частиною реального світу, яку вона має моделювати [3]. З погляду природничих наук точність зазвичай визначають як «величину похибки». Дослідник Redman визначає точність на рівні поля (field-level) та на рівні запису (record-level) так.

Точність на рівні поля визначається як відношення кількості полів, визнаних «коректними», до загальної кількості перевірених полів:

$$\text{field lvl acc.} = \frac{\text{number of fields judged "correct"}}{\text{number of fields tested}}. \quad (1.1)$$

Точність на рівні запису визначається як відношення кількості записів, визнаних «повністю коректними», до загальної кількості перевірених записів:

$$\text{record lvl acc.} = \frac{\text{number of records judged "completely correct"}}{\text{number of records tested}}. \quad (1.2)$$

Цю метрику також використовує асоціація DAMA UK [4], узагальнюючи поняття «полів» та «записів» до «об'єктів». Деякі дослідники застосовують обернену метрику (одиниця мінус відношення кількості одиниць даних з помилками до загальної кількості одиниць даних), а інші додатково враховують випадковість виникнення помилки (ROE) та розподіл імовірності виникнення помилки (PDOE):

$$\text{acc.} = \left(\frac{\text{NrOfCorrectValues}}{\text{TotalNrOfValues}}, \text{ROE}, \text{PDOE}\right). \quad (1.3)$$

В роботі [11] автор запропонував метрику точності, яку можна агрегувати на різних рівнях. На рівні значення атрибута метрика точності визначається відношенням між «арністю» (кількістю розрядів) значення та його оптимальною арністю для числових значень. Для числового атрибута A нехай оптимальна кількість цифр і десяткових знаків для A позначається відповідним чином, w – значення атрибута A, а фактична кількість цифр і десяткових знаків для w. Оскільки оптимальне значення не обов'язково є максимальним, метрику необхідно нормувати в межах інтервалу [0, 1]:

$$Q_{G\,en}(w, A) = min\left(\frac{s(w)}{s_{opt}(A)}, 1\right). \quad (1.4)$$

Для нечислових атрибутів пропонується відносити значення w до площини i в межах класифікації K з n площинами та заміняти фактичну кількість розрядів на i. Для кортежу t точність обчислюється згідно з наведеним нижче співвідношенням, де ваговий коефіцієнт відображає відносну важливість атрибута стосовно всього кортежу і задається експертом:

$$Q_{G\,en}(t) = \frac{\sum_{j=1}^{n} Q_{G\,en}(t.A_j, A_j) g_j}{\sum_{j=1}^{n} g_j}. \quad (1.5)$$

Точність на рівні таблиці обчислюється як середнє арифметичне результатів вимірювання точності кортежів, а точність на рівні бази даних – як середнє арифметичне результатів вимірювання точності на рівні таблиць.

1.5 Повнота (Completeness)

Повноту дуже узагальнено описують як «широту, глибину та обсяг інформації, що міститься в даних» [3, 24], і вона охоплює умову існування даних. З урахуванням наявних робіт найзагальнішу метрику повноти можна визначити як відношення кількості повних елементів до загальної кількості елементів:

$$\text{Completeness} = \frac{|e_c|}{|e|}, \tag{1.6}$$

де у чисельнику число повних елементів, а в знаменнику – загальна кількість всіх елементів.

Тут узагальнений термін «елемент» може позначати будь-яку одиницю даних, наприклад атрибут, запис або таблицю. Деякі дослідники використовують обернену метрику, а інші пропонують порівнювати кількість повних елементів із загальною кількістю елементів у ідеальному еталонному наборі даних. Детальнішу специфікацію способу обчислення повноти надає робота [11], в якій присвоюється значення 0,0 полю, що містить null або еквівалент, та значення 1,0 – у протилежному випадку. На основі цього припущення повноту можна обчислювати аналогічно метриці точності на різних рівнях агрегації за допомогою зваженого середнього арифметичного. Наприклад, повнота на рівні таблиці визначається як середнє арифметичне повноти всіх записів таблиці:

$$Q_{Voll}(T) = \frac{\sum_{i=1}^{|T|} Q_{Voll}(t_i)}{|T|} \tag{1.7}$$

Варто зазначити, що окрім підходу, за яким враховуються лише справді відсутні значення (null), можливий також суворіший підхід до повноти, за якого як неповні значення розглядаються значення за замовчуванням або текстові записи на кшталт «NaN» (не число).

Хоча автор [11] не пропонує метрику повноти на рівні атрибута (стовпця), а інші роботи описують повноту на рівні атрибута лише текстово, таку метрику

можна вивести з опису як відношення кількості повних (не порожніх) значень у стовпці до загальної кількості значень у цьому стовпці:

$$C_{att} = \frac{|v_c|}{|v|}. \tag{1.8}$$

1.6 Узгодженість (Consistency)

Для виміру узгодженості також існують різні визначення. Згідно з одним із поширених визначень [3], узгодженість фіксує порушення семантичних правил, визначених над елементами даних, де елементами можуть бути кортежі реляційних таблиць або записи у файлі. Прикладом таких правил є обмеження цілісності з реляційної теорії. В [11] автор припускає для своєї метрики узгодженості, що предметні знання закодовані у вигляді правил, та виключає суперечності всередині правил, а також нечіткі чи ймовірнісні припущення. Відповідно, узгодженість значення атрибута w визначається так:

$$Q_{Kon}(w) = \frac{1}{\sum_{j=1}^{n} r_j(w) g_j + 1}, \tag{1.9}$$

де ваговий коефіцієнт відображає ступінь серйозності порушення відповідного правила узгодженості, а функція порушення правила узгодженості, застосована до значення атрибута w (у межах множини з n правил узгодженості), визначається як індикаторна функція.

Вона набуває значення 0, якщо w задовольняє правило, і 1 – у протилежному випадку:

$$r_j(w) \begin{cases} 0 & \text{if } w \text{ satisfies } r_j \\ 1 & \text{otherwise.} \end{cases} \tag{1.10}$$

Правила узгодженості можна визначати не лише на рівні значення атрибута, а й на рівні кортежу. Обчислення узгодженості на рівні таблиці або бази даних виконується аналогічно метрикам точності й повноти – як середнє арифметичне узгодженості на рівні кортежів. Крім того, узгодженість можна вимірювати в часі, порівнюючи розподіл кількості записів за значеннями (профіль стовпця) з попередніми екземплярами даних, що заповнювали те саме поле [23].

1.7 Своєчасність (Timeliness)

Своєчасність описує, «наскільки актуальними є дані для поставленого завдання» [3], і тісно пов'язана з поняттями актуальності (currency – частоти оновлення даних) та волатильності (volatility – швидкості, з якою дані стають нерелевантними). За іншим визначенням, своєчасність можна інтерпретувати як імовірність того, що значення атрибута досі є актуальним. Перелік різних метрик для обчислення своєчасності наведено авторами в [11], які пропонують обчислювати своєчасність на основі експоненційної функції згасання:

$$Q^{\omega}_{Time.}(t) := \exp(-\operatorname{decline}(A) \cdot t). \tag{1.11}$$

де розглядається значення ω атрибута, а коефіцієнт згасання (decline rate) визначає середню кількість атрибутів, що застарівають протягом певного періоду часу. Наведений перелік метрик для вимірів точності, повноти, своєчасності та узгодженості жодним чином не є вичерпним, проте він демонструє, що література пропонує низку конкретно сформульованих метрик для вимірювання вимірів якості даних. Подальше дослідження спостерігає реалізацію цих метрик у сучасних інструментах якості даних.

Підсумовуючи, варто наголосити, що широке узгодження щодо вимірів та метрик якості даних у науковій спільноті співіснує з відсутністю стандартизованого переліку вимірів і метрик для вимірювання якості даних. Це

створює труднощі під час зіставлення теоретичних концепцій з їх практичною реалізацією в програмних інструментах, що детально розглядається в наступних розділах.

2 ОПИС ПРОЦЕСУ ОЦІНЮВАННЯ ЯКОСТІ ДАНИХ

2.1 Управління якістю даних

Для підвищення довіри до рішень, що приймаються на основі даних, необхідно вимірювати, знати та покращувати якість використовуваних даних за допомогою відповідних інструментів [9]. Покращення якості даних (тобто очищення даних), що ґрунтується на вимірюванні якості даних, є частиною комплексного управління якістю даних. Більшість наявних методологій описують управління якістю даних як циклічний процес, який виконується безперервно. Цей розділ присвячено опису процесу управління якістю даних, його основних етапів, а також методології оцінювання інструментів якості даних.

Асоціація управління даними (DAMA) визначає «управління якістю даних» як аналіз, покращення та забезпечення якості даних [20]. Протягом років було запропоновано низку різних методологій якості даних (також відомих як «фреймворки», «програми» чи «методи»), наприклад методологію загального управління якістю даних TDQM (Total Data Quality Management) [22], методологію оцінювання якості інформації AIMQ, а також методи оцінювання якості даних, запропоновані іншими дослідниками. Комплексне порівняння методологій якості даних було проведено у відповідних оглядових роботах.

Хоча ці методології мають різні характеристики та акценти, з них можна виокремити чотири основні види діяльності:

1) реконструкція стану (state reconstruction);

2) вимірювання або оцінювання якості даних (DQ measurement or assessment);

3) очищення або покращення даних (data cleansing or improvement);

4) встановлення безперервного моніторингу якості даних (continuous DQ monitoring).

Не всі методології включають усі ці етапи. Наприклад, у деяких методологіях опускається етап реконструкції стану, а в окремих оглядах

методологій – етап моніторингу. Крім того, деякі методології включають додаткові види діяльності, як-от моніторинг інтерфейсів інтеграції даних, що тут не розглядаються через їх вузьку спеціалізацію.

Етап реконструкції стану описує збирання контекстної інформації про спостережувані дані, а також про організацію, у якій реалізується проєкт з якості даних. Оскільки фокус цієї роботи зосереджено на функціональності інструментів якості даних, етап реконструкції стану в подальшому обмежується частиною, що стосується даних (тобто профілюванням даних), без детального розгляду збирання контекстної інформації про організацію. Нижче детально описано чотири основні етапи методології якості даних, щоб прояснити різницю між вимірюванням якості даних, моніторингом якості даних та очищенням даних.

2.1.1 Профілювання даних

Профілювання даних (data profiling) описують як процес аналізу набору даних для збирання даних про дані (тобто метаданих) із застосуванням широкого спектра методів [1, 19]. Таким чином, це важливе завдання, що передує будь-якій діяльності з вимірювання чи моніторингу якості даних, оскільки воно дає змогу отримати уявлення про наявний набір даних. Прикладами інформації, що збирається під час профілювання даних, є кількість різних або відсутніх (null) значень у стовпці, типи даних атрибутів, а також наявні шаблони та частота їх появи в наборі даних (наприклад, формат запису телефонних номерів) [1]. Згідно з результатами галузевих оглядів та власного дослідження, більшість інструментів якості даних загального призначення тією чи іншою мірою пропонують можливості профілювання даних.

Завдання профілювання даних зазвичай класифікують за рівнем складності та об'єктом аналізу [1, 2]. Одностовпцеве (single-column) профілювання охоплює аналіз окремих атрибутів і поділяється на кілька підкатегорій. Потужності (cardinalities) включають кількість рядків, кількість та відсоток порожніх

значень, кількість різних значень (cardinality) та відношення кількості різних значень до кількості рядків. Розподіли значень (value distributions) охоплюють частотні гістограми, мінімальне й максимальне значення в числовому стовпці, сталість (відношення частоти найчастішого значення до кількості рядків), квартилі та аналіз розподілу першої цифри для перевірки закону Бенфорда. Категорія шаблонів, типів даних і доменів охоплює розпізнавання базових та специфічних для СКБД типів даних, вимірювання довжини значень, виявлення шаблонів значень, а також узагальнених семантичних типів даних і семантичних доменів.

Багатостовпцеве (multi-column) профілювання та виявлення залежностей є складнішими завданнями. До виявлення залежностей належать унікальні комбінації стовпців (unique column combinations, UCC) як кандидати на ключі, залежності включення (inclusion dependencies) як основа для виявлення зовнішніх ключів, а також функціональні залежності (functional dependencies). Розширене багатостовпцеве профілювання охоплює аналіз кореляцій, кластеризацію, виявлення викидів, точне та наближене виявлення дублікатів, а також аналіз асоціативних правил. Саме ці складніші завдання, як показує подальший аналіз, підтримуються в сучасних інструментах найгірше, що становить значний потенціал для вдосконалення.

2.1.2 Вимірювання якості даних

Одним із найбільших викликів для практиків якості даних є відповідь на питання про те, як саме слід вимірювати якість даних [23]. Те саме стосується синонімічно вживаного терміна «оцінювання» (assessment): одним із головних питань досліджень якості даних є питання «Як оцінити якість даних?» [10]. Термін «вимірювати» (measure) описує визначення розміру, кількості чи ступеня чогось за допомогою інструмента або шляхом порівняння з об'єктом відомого розміру.

Хоча термін «оцінювання» часто вживають як синонім вимірювання, у літературі з якості даних існує чітке розмежування між цими термінами. Оцінювання – це «оцінка чи визначення природи, здатності або якості чогось», що розширює поняття вимірювання шляхом інтерпретації результатів вимірювання та формування висновку про об'єкт оцінювання [23]. У цій роботі переважно використовується термін «вимірювання», оскільки фокус зосереджено на можливостях вимірювання інструментів якості даних, незалежно від інтерпретації результатів користувачем.

Окрім наукових публікацій, консенсус практиків і дослідників мають відображати стандарти. У сфері якості даних значну роботу проведено підкомітетом 7 спільного технічного комітету ISO/IEC JTC 1. Робоча група цього підкомітету опублікувала стандарти ISO/IEC 25012:2008 [14], ISO/IEC 25024:2015 та ISO/IEC 25040:2011. Паралельно підкомітет SC 4 технічного комітету ISO/TC 184 опублікував стандарт ISO 8000-8:2015 [15]. Якщо стандарт ISO 8000-8:2015 визначає передумови для вимірювання та звітування про якість інформації й даних на дуже загальному рівні, то стандарт ISO/IEC 25012:2008 надає конкретніші заходи з якості даних, а також пояснення щодо їх застосування.

Згідно зі стандартом ISO 8000-8:2015 [15], дані можна вимірювати на дуже загальному рівні за такими аспектами: синтаксична якість (syntactic quality), що описує ступінь відповідності даних заданому синтаксису; семантична якість (semantic quality), тобто ступінь відповідності даних їх реальному представленню; прагматична якість (pragmatic quality), тобто ступінь придатності даних для конкретної мети. Стандарт ISO/IEC 25012:2008 [14] визначає вимірювання (якості даних) як «набір операцій, метою яких є визначення значення міри», і задає набір нормалізованих мір якості (у межах від 0 до 1).

Поділ якості даних на набір вимірів, які можна вимірювати за допомогою метрик, є широко прийнятим у дослідженнях якості даних [12, 24]. Заходи з якості, що пропонуються стандартом ISO/IEC 25012:2008 [14], відповідають

найпопулярнішим метрикам у літературі (наприклад, точність, повнота, узгодженість). Однак, попри широке узгодження щодо вимірів і метрик якості даних загалом та тривалі дослідження протягом останніх десятиліть, досі немає консенсусу щодо стандартизованого переліку вимірів і метрик для вимірювання якості даних [23].

2.1.3 Очищення даних

Очищення даних (data cleansing) описує процес виправлення помилкових даних або збоїв у даних [8]. На практиці до автоматизованих завдань очищення належать стандартизація даних про клієнтів, усунення дублікатів та зіставлення (matching). Інші зусилля щодо покращення якості даних зазвичай виконуються вручну. Хоча автоматизовані методи очищення даних є дуже цінними для великих обсягів даних, вони несуть ризик внесення нових помилок, які рідко бувають добре зрозумілими [17]. У межах цього дослідження функціональність очищення даних навмисно не розглядається, оскільки фокус зосереджено на виявленні проблем якості даних. Проте варто зазначити, що алгоритми очищення даних зазвичай ґрунтуються на вимірюванні якості даних, адже для підвищення якості наявного набору даних спершу необхідно виявити проблеми якості.

2.1.4 Моніторинг якості даних

Термін «моніторинг якості даних» (DQ monitoring) здебільшого вживається в літературі неявно, без усталеного визначення та спільного розуміння. Це призводить до різних інтерпретацій, коли цей термін згадується в наукових публікаціях або компаніями, що просувають і описують свої інструменти якості даних. Існує різниця між «моніторингом даних» (data monitoring), що описує безперервну перевірку правил, та «моніторингом якості даних» (DQ monitoring), що є безперервним вимірюванням якості даних [9]. Метою цього дослідження є спостереження не лише за функціональністю

сучасних інструментів якості даних щодо профілювання та вимірювання, а й за справжнім моніторингом якості даних.

Попередні дослідження вказували на те, що жоден з оглянутих інструментів не мав функціональності моніторингу. Однак, оскільки на низці вебсайтів інструментів якості даних є свідчення того, що вони пропонують функції моніторингу, цей критерій було включено до каталогу вимог. Згідно зі стандартом ISO 8000-8:2015 [15], прагматичне вимірювання якості даних вимагає взаємодії з відповідними користувачами, які перевіряють дані. Відповідно, повністю автоматизований моніторинг якості даних обмежується синтаксичними та семантичними аспектами якості даних.

2.2 Вимоги до інструментів якості даних

Загалом існує дуже мало наукових праць, що досліджують функціональний обсяг інструментів якості даних, і ще менше праць, які пропонують спеціальний каталог вимог для їх оцінювання. Окрім вимог, визначених дослідниками, існує кілька орієнтованих на практиків та постачальників оглядів, зокрема щорічні звіти аналітичної компанії Gartner Inc. [6, 25], які розглядають інструменти якості даних за такими можливостями: підключення (connectivity), профілювання даних, вимірювання та візуалізація, моніторинг, парсинг, стандартизація та очищення, зіставлення, зв'язування та об'єднання, підтримка кількох предметних областей, перевірка адрес, курування та збагачення даних, вирішення проблем та робочі процеси, управління метаданими, середовище розгортання, архітектура та інтеграція, а також зручність використання.

Подібним чином у літературі визначають вісім вимог, яким має відповідати інструмент якості даних: профілювання даних, парсинг, стандартизація, вирішення ідентичності (identity resolution), зв'язування та об'єднання записів, очищення даних, збагачення даних, а також інспектування й моніторинг даних. Однак такі переліки вимог виявилися надто узагальненими для мети детального

спостереження за функціональністю профілювання даних, вимірювання якості даних та моніторингу якості даних.

На основі наявних досліджень у цій роботі використовується каталог вимог для оцінювання інструментів якості даних, що складається з трьох категорій: профілювання даних, вимірювання якості даних у термінах вимірів і метрик та безперервний моніторинг якості даних. Каталог вимог підсумовує та класифікує завдання, необхідні для автоматизованого й безперервного вимірювання якості даних, у новий спосіб.

2.3 Методологія дослідження та відбору інструментів

Систематичний огляд зазвичай розпочинається з визначення протоколу, що задає досліджувані питання та методи, які будуть застосовані [16]. Структуру протоколу дослідження було виведено з методології систематичних оглядів у комп'ютерних науках. Оскільки оригінальна методологія зосереджена на оцінюванні первинних наукових праць, а не конкретних реалізацій, деякі етапи (зокрема оцінювання якості, стратегію вилучення даних та синтез вилучених даних з оригінальних наукових праць) було опущено.

2.3.1 Дослідницькі питання

Головне дослідницьке питання роботи формулюється так: як концепції вимірювання та моніторингу якості даних реалізовані в сучасних інструментах якості даних? Це питання деталізовано трьома підпитаннями:

1) Які можливості профілювання даних підтримуються сучасними інструментами якості даних?
2) Які виміри та метрики якості даних можна вимірювати за допомогою сучасних інструментів якості даних?
3) Чи дозволяють інструменти якості даних автоматизований моніторинг якості даних у часі?

2.3.2 Систематичний пошук інструментів

Для відповіді на дослідницькі питання було проведено систематичний пошук, у межах якого ідентифіковано 667 програмних інструментів, присвячених темі «якість даних». Пошук виконувався за допомогою онлайн-пошукової системи наукових публікацій на додаток до основних вебсайтів видавців, із спеціальним розглядом посилань з наявних оглядів інструментів якості даних, а також з паралельним ручним пошуком. Загалом через систематичний пошук було виявлено понад тисячу наукових праць, які посилаються на сотні інструментів якості даних. Після об'єднання всіх знайдених інструментів в один файл та видалення дублікатів отримано загальну кількість у 667 унікальних інструментів якості даних.

Приблизно половина (50,82 %) виявлених інструментів якості даних виявилися предметно-залежними (domain specific), тобто вони були або призначені для конкретних типів даних, або створені для вимірювання якості даних певного пропрієтарного інструмента. Близько 16,67 % інструментів зосереджувалися на очищенні даних без належної стратегії вимірювання якості даних.

2.3.3 Критерії відбору інструментів

Відповідно до загальної стратегії пошуку, було визначено три критерії включення: інструмент був включений до одного з попередніх оглядів; інструмент був ідентифікований у систематичному пошуку; інструмент був ідентифікований у випадковому (ручному) пошуку. Після формування переліку інструментів-кандидатів було проведено перегляд для виключення всіх інструментів, що відповідали принаймні одному з таких критеріїв виключення:

– (EC1) інструмент є предметно-залежним (наприклад, лише для веб-даних або конкретних реалізацій);

– (EC2) інструмент призначений для конкретних завдань управління даними без явного надання можливості вимірювання якості даних (зокрема для очищення даних, інтеграції даних або інших завдань, як-от візуалізація даних);

– (EC3) інструмент не є загальнодоступним (наприклад, описаний лише в науковій праці);

– (EC4) інструмент вважається застарілим (постачальник більше не існує або останній комміт у репозиторії був до 1 січня 2016 року);

– (EC5) інструмент знайдено у відкритому репозиторії без будь-якої додаткової доступної інформації;

– (EC6) інструмент потребує оплати, і пробна версія не надається на запит.

Більшість інструментів було виключено через те, що вони є предметно-залежними (EC1) та/або зосереджені на конкретних завданнях управління даними (EC2). У результаті застосування критеріїв виключення для глибшого дослідження було відібрано 17 інструментів якості даних, з яких вдалося оцінити 13: три з них базувалися на платформі SAP, де не було доступної інсталяції, а ще один не вдалося успішно встановити протягом часу проєкту.

2.3.4 Обмеження дослідження

Фаза відбору є критичною, оскільки рішення, прийняті на цьому етапі, безсумнівно мають значний вплив на валідність результатів огляду літератури. Основними загрозами валідності дослідження вважаються проведення процесу відбору та притаманні йому обмеження. Ризик пропуску важливої наукової праці та відповідного інструмента якості даних було зменшено такими заходами: використанням онлайн-пошукової системи на додаток до основних вебсайтів видавців; спеціальним розглядом посилань з наявних оглядів інструментів якості даних; включенням ручного пошуку паралельно з систематичним пошуком.

З огляду на співвідношення між кількістю інструментів, відібраних для глибшого дослідження, та загальною кількістю ідентифікованих інструментів, критерії виключення можуть видаватися дуже суворими. Проте вони були обрані

адекватно з кількох причин. По-перше, існує велика кількість інструментів, які є лише простими скриптами для очищення конкретних наборів даних. По-друге, навмисно виключено інструменти, обмежені конкретними завданнями управління даними (наприклад, очищенням даних), оскільки вони не сприяють відповіді на загальне дослідницьке питання. По-третє, час, витрачений на кожен інструмент, становив близько одного людино-місяця, що було зумовлено детальним каталогом вимог, а також складністю інсталяції деяких інструментів.

2.4 Каталог вимог та стратегія оцінювання

На основі наявних досліджень було сформовано каталог вимог для оцінювання інструментів якості даних, що містить 43 вимоги, розподілені за трьома категоріями. Категорія профілювання даних охоплює 30 вимог, що ґрунтуються здебільшого на класифікації профілювання даних [1], та поділяється на підкатегорії: потужності (cardinalities), розподіли значень (value distributions), шаблони, типи даних і домени (patterns, data types, and domains), залежності (dependencies) та розширене багатостовпцеве профілювання (advanced multi-column profiling). Категорія вимірювання якості даних охоплює виміри точності, повноти, узгодженості та своєчасності, інші метрики якості даних, а також бізнес-правила. Категорія моніторингу якості даних охоплює можливості безперервного вимірювання якості даних у часі.

Для кожної вимоги під час оцінювання можливі три варіанти: вимога виконується (✓); вимога не виконується (−); вимога виконується частково (p). Покриття кожної вимоги описується в текстовій формі з акцентом на обґрунтуванні часткового виконання. Така стратегія оцінювання забезпечує однорідне та відтворюване порівняння функціональних можливостей різних інструментів якості даних.

3 ОГЛЯД ІНСТРУМЕНТІВ ЯКОСТІ ДАНИХ

У цьому розділі описано інструменти якості даних, відібрані для глибшого дослідження, та наведено порівняльний аналіз їх можливостей. Загалом було оцінено 13 інструментів (8 комерційних та 5 з відкритим кодом), що надають досліджувані функції та не обмежені конкретною предметною областю. Серед оцінених інструментів – Aggregate Profiler, Apache Griffin, Ataccama ONE, DataCleaner, Datamartist, Experian Pandora, Informatica Data Quality, InfoZoom & IZDQ, MobyDQ, OpenRefine & MetricDoc, Oracle Enterprise Data Quality, SAS Data Quality та Talend Open Studio for Data Quality. Опис інструментів подано в алфавітному порядку.

3.1 Опис досліджених інструментів

3.1.1 Aggregate Profiler

Aggregate Profiler (AP) – це безкоштовно доступний інструмент якості даних, призначений для профілювання даних (див. рис. 3.1).

Окрім можливостей профілювання, як-от статистичний аналіз та зіставлення шаблонів, Aggregate Profiler можна також використовувати для підготовки та очищення даних, наприклад для виправлення адрес або усунення дублікатів. Крім того, у ньому можна визначати та планувати бізнес-правила на задані користувачем періоди.

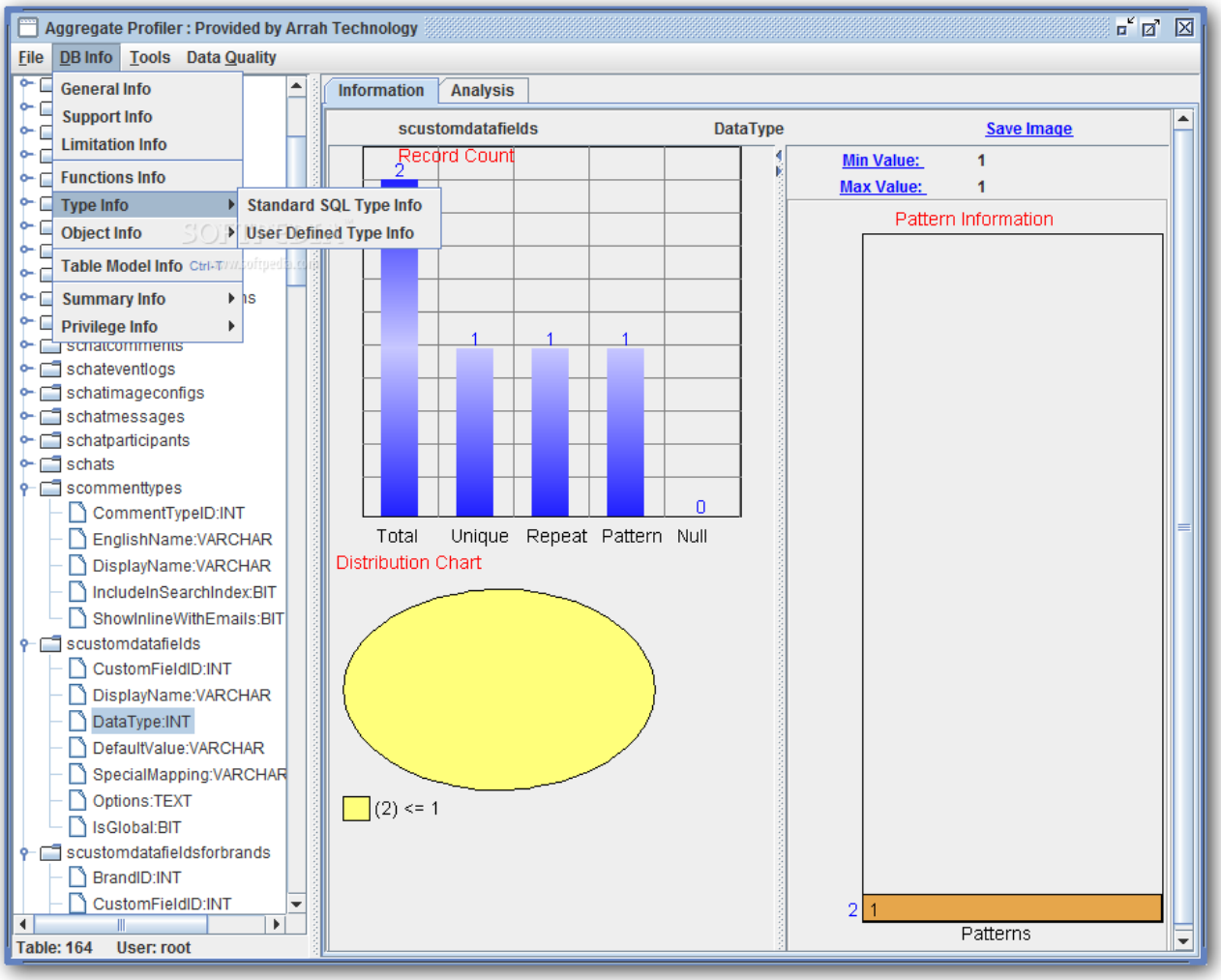

Рисунок 3.1 – Приклад інтерфейсу Aggregate Profiler

Користувацький інтерфейс інструмента сприймався як менш досконалий порівняно з іншими інструментами, оскільки навігація та застосування функцій профілювання не були інтуїтивно зрозумілими.

3.1.2 Apache Griffin

Apache Griffin (AG) суттєво відрізняється від інших інструментів цього огляду, оскільки не пропонує жодної функціональності профілювання даних і не є комплексним рішенням якості даних.

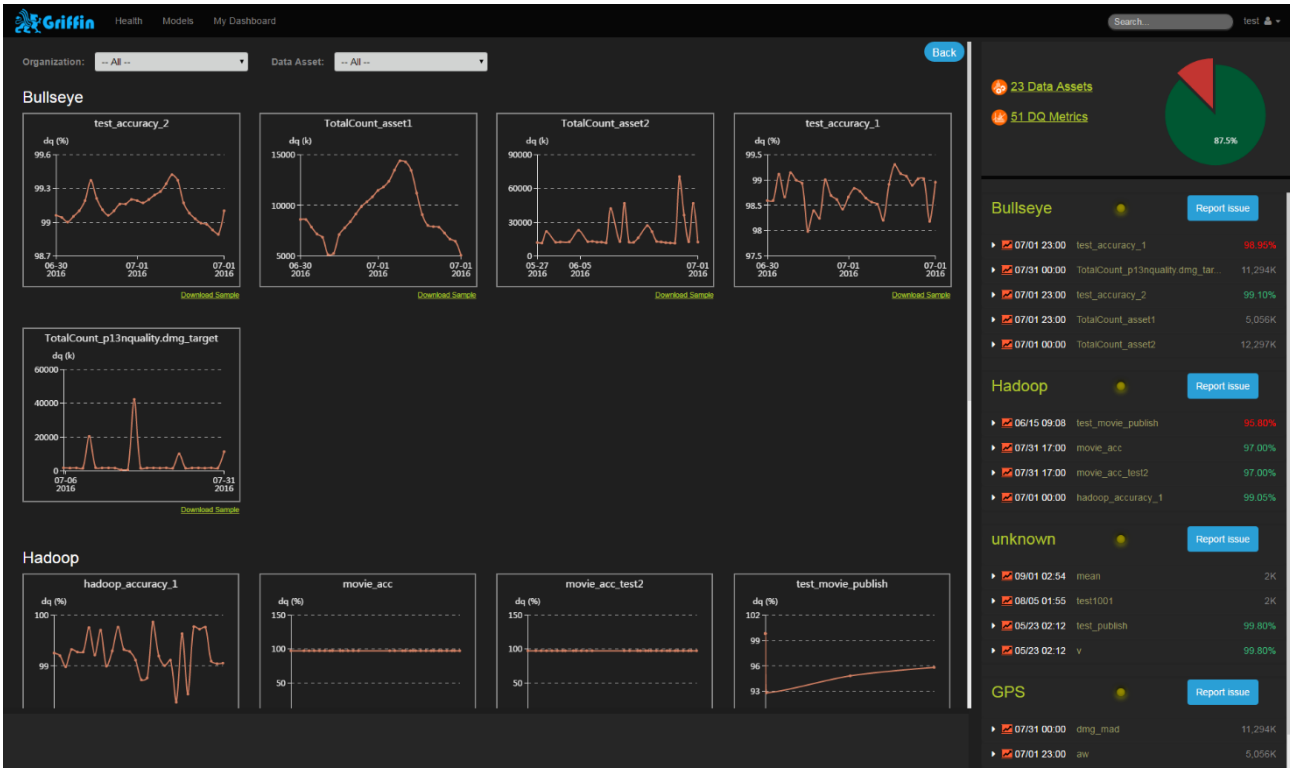

Рисунок 3.2 – Приклад інтерфейсу Apache Griffin

Однак, оскільки частиною оцінювання є спостереження за тим, наскільки сучасні інструменти підтримують безперервне вимірювання якості даних, Apache Griffin було включено до огляду, адже він призначений для безперервного вимірювання якості великих даних – як пакетних, так і потокових. Інструмент потребує низки залежностей (зокрема JDK, MySQL, Hadoop, Spark, Hive та інших), через що його інсталяція виявилася дуже складною: двом досвідченим фахівцям знадобилося понад тиждень для повного встановлення. Після встановлення інтерфейс є інтуїтивно зрозумілим і підтримує предметно-залежне визначення метрик точності, а також планування та моніторинг цих метрик (див. рис. 3.2).

3.1.3 Ataccama ONE

Компанія Ataccama пропонує кілька продуктів якості даних. З 2017 року окремі рішення були консолідовані в продукт «Ataccama ONE» (A-ONE). Хоча

ліцензія повного рішення є платною, модуль профілювання даних Ataccama ONE доступний безкоштовно (див. рис. 3.3).

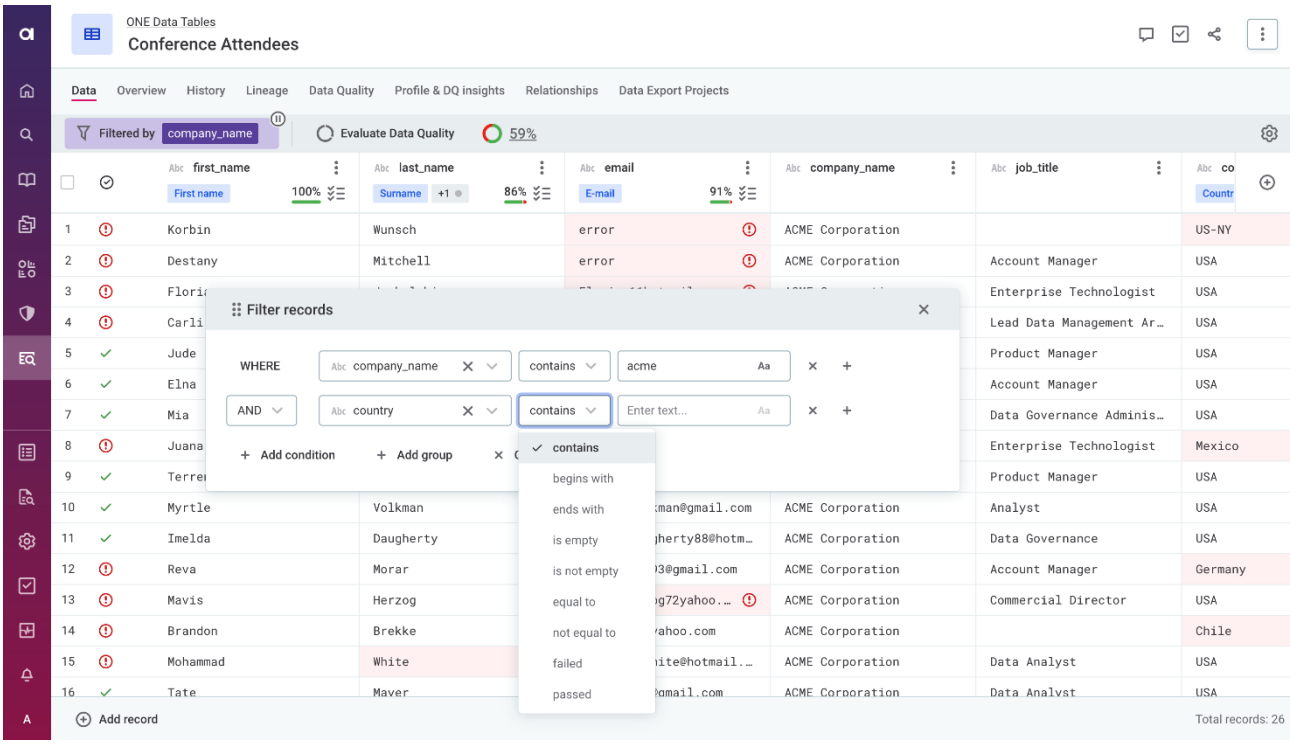

Рисунок 3.3 – Вигляд інтерфейсу Ataccama ONE

У межах дослідження вдалося вивчити лише безкоштовний профайлер «Ataccama ONE profiler», фокус якого зосереджено на профілюванні даних і який не надає функціональності моніторингу. Модуль профілювання даних був дуже інтуїтивним та простим у використанні, зокрема й для бізнес-користувачів. Згідно з інформацією постачальника, повне рішення надавало б значно багатший набір функцій, зокрема моніторинг якості даних.

3.1.4 DataCleaner

Продукти якості даних «DataCleaner» (DC) та «DataHub» спершу розроблялися компанією Human Inference. DataCleaner пропонує спеціальну та незалежну функціональність вимірювання якості даних, хоча через назву можна було б очікувати лише функцій очищення даних.

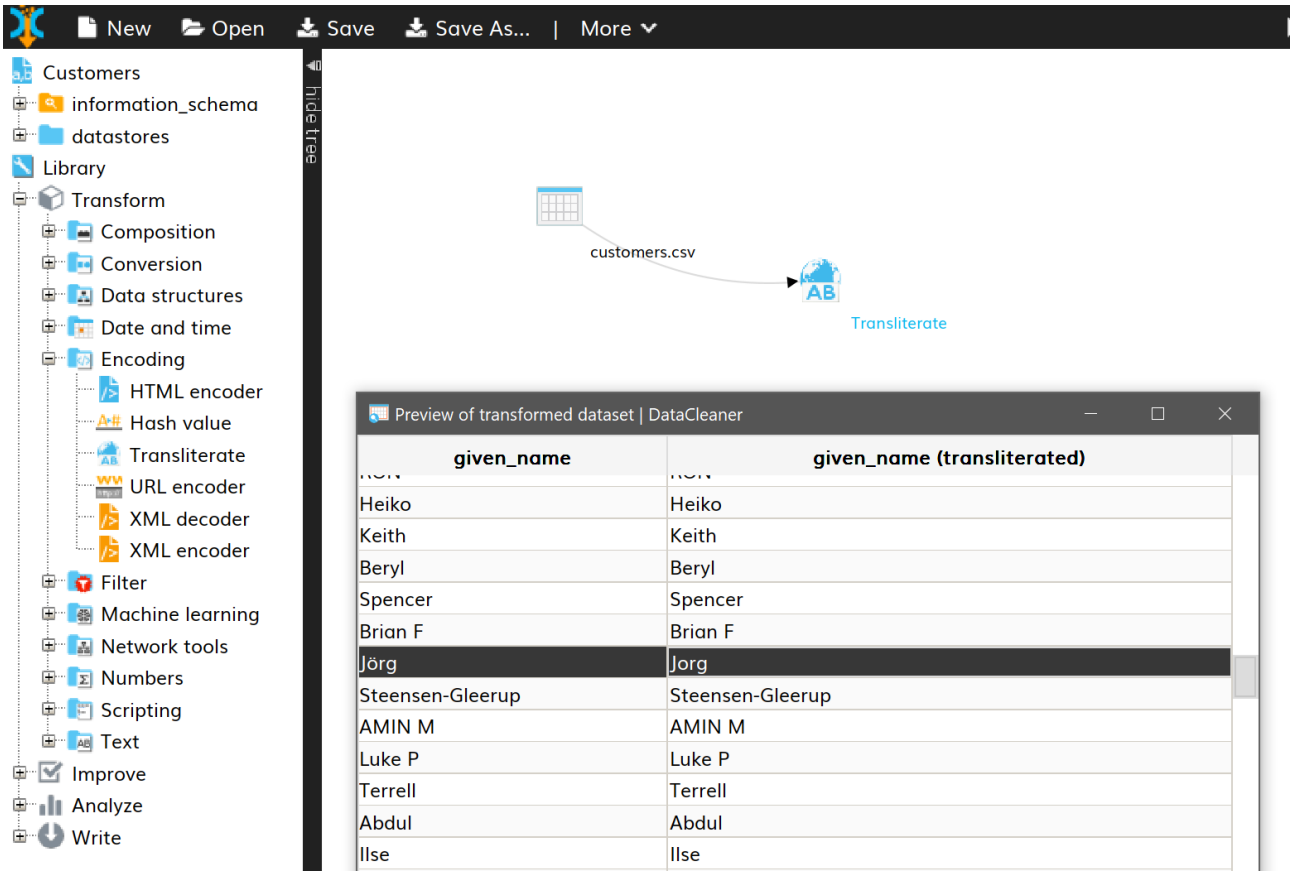

Рисунок 3.4 – Інтерфейс програми DataCleaner

Професійна версія DataCleaner надає ті самі функції вимірювання якості даних, що й DataHub, відрізняючись лише зручністю використання, інтерфейсом та функціями інтеграції даних (див. рис. 3.4). Компанія робить акцент на даних про клієнтів, що відображається в спеціальних алгоритмах виявлення дублікатів, зіставлення адрес та очищення даних. Попри орієнтацію на технічних користувачів, інтерфейс сприймався як дуже інтуїтивний.

3.1.5 Datamartist

Комерційний інструмент Datamartist (DM) потребує операційної системи Microsoft Windows та фреймворку .NET. Datamartist призначений для профілювання та трансформації даних. Досліджена 30-денна пробна версія пропонувала всі функції Pro-видання.

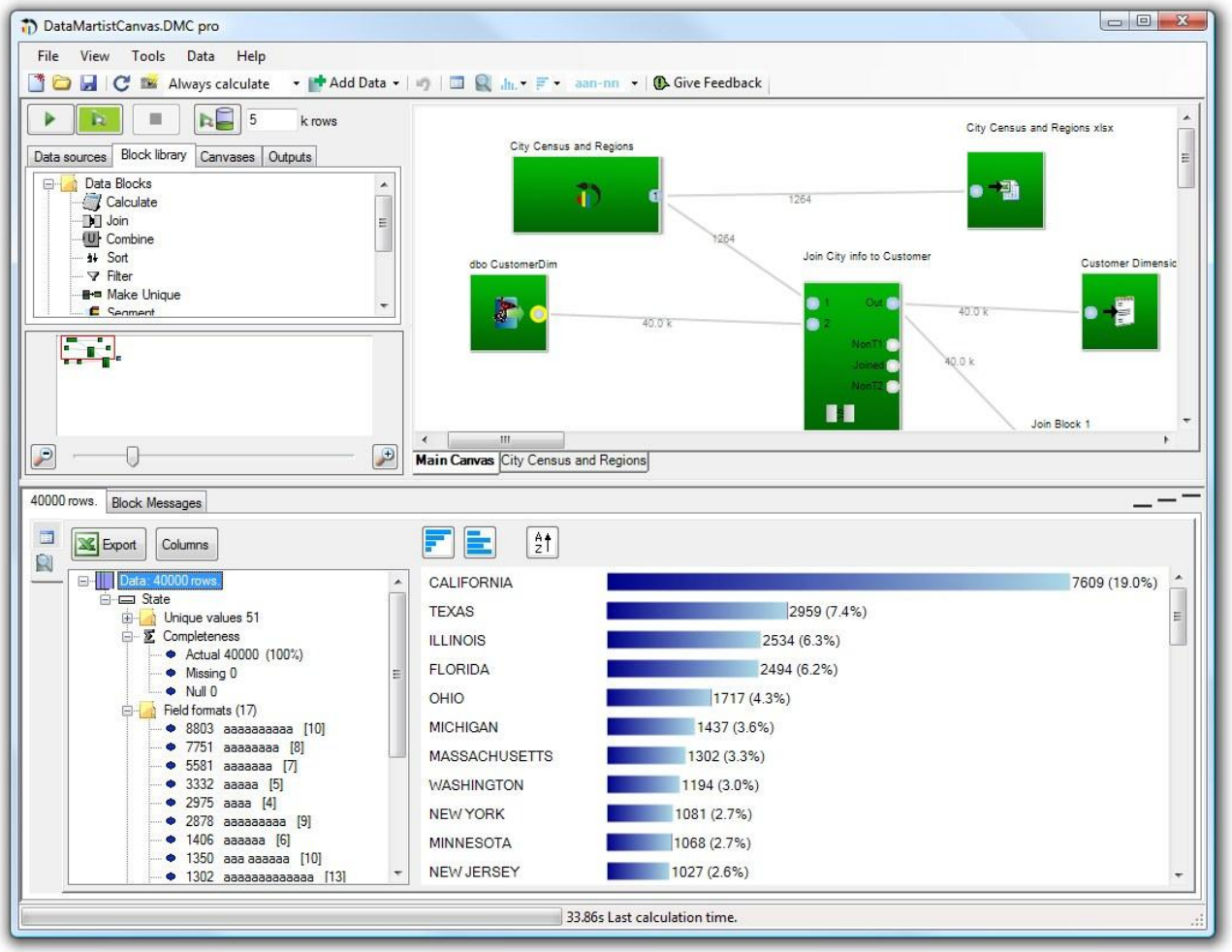

Рисунок 3.5 – Вигляд інтерфейсу Datamartist

Інтерфейс Datamartist (див. рис. 3.5) сприймався як дещо менш досконалий порівняно з іншими комерційними інструментами, оскільки для деяких завдань (наприклад, експорту результатів профілювання) була потрібна командна стрічка.

3.1.6 Experian Pandora

Компанія Experian пропонує два комерційні рішення якості даних: Cleanse та Pandora (EP). У межах дослідження вивчено комплексніший інструмент Pandora. Інструмент сприймався як простий у встановленні та використанні; особливо варто відзначити комплексні можливості профілювання даних загалом і можливості міжтабличного профілювання зокрема (див. рис. 3.6).

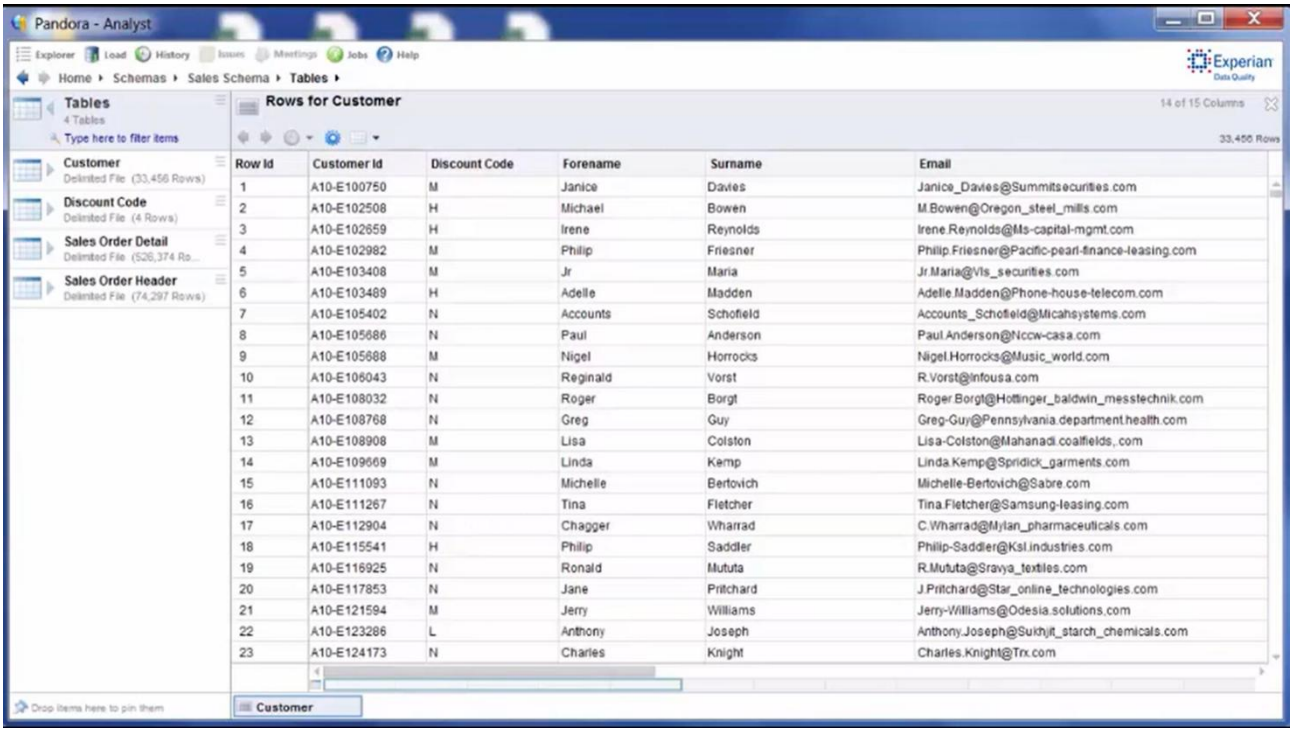

Рисунок 3.6 – Вигляд інтерфейсу Experian Pandora

Крім того, Pandora надає широку можливість розширення наявної палітри функцій за допомогою користувацьких функцій. Загалом Pandora отримав одну з найкращих сумарних оцінок у дослідженні.

3.1.7 Informatica Data Quality

Informatica Data Quality (IDQ) – це один із модулів комерційного рішення для управління даними від компанії Informatica, яка протягом кількох років є лідером у звітах Gartner [25]. Пробна версія включала Informatica Developer (десктопну інсталяцію для розробників), Informatica Analyst (вебплатформу для бізнес-користувачів) та Informatica Administrator (для планування завдань). Щодо вимірювання якості даних, Informatica пропонує, ймовірно, найближчу реалізацію до погляду на виміри й метрики якості даних, що просувається в науковій спільноті.

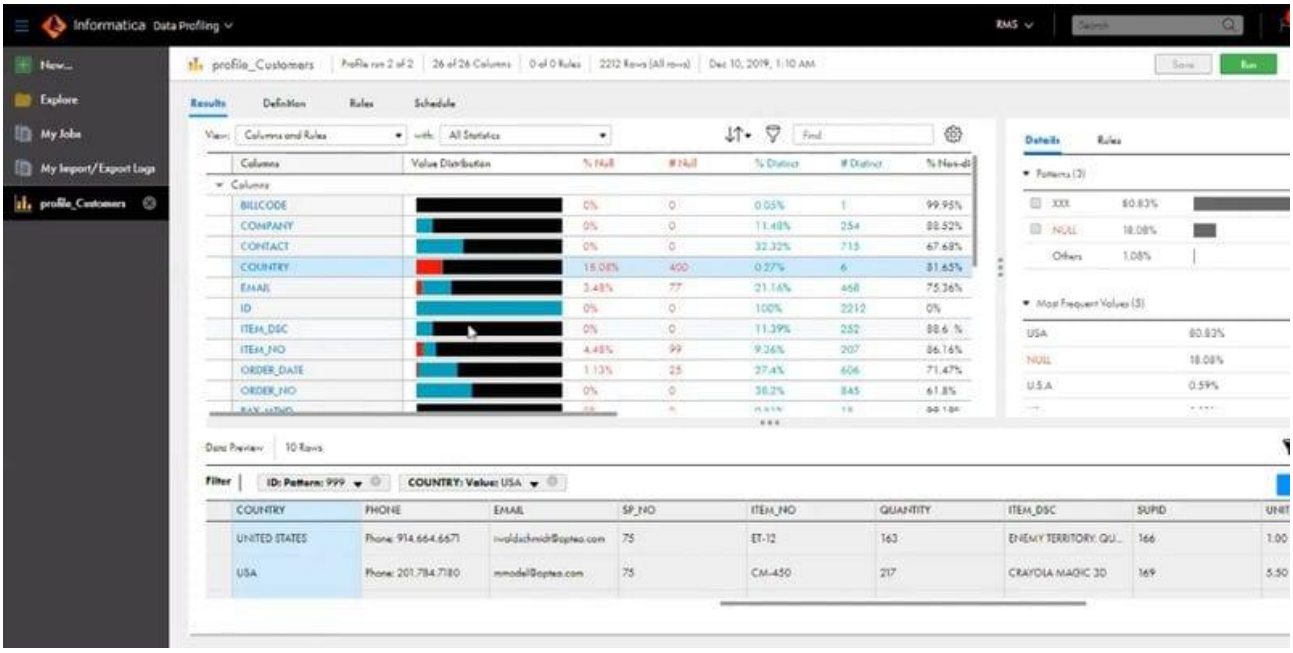

Рисунок 3.7 – Інтерфейс Informatica Analyst

Інтерфейс Informatica Analyst сприймався як простий у використанні, зокрема й для бізнес-користувачів (див. рис. 3.7).

3.1.8 IBM InfoSphere Information Server for Data Quality

Продукт «InfoSphere Information Server for Data Quality» (IBM ISDQ) від компанії IBM (див. рис. 3.8) було знайдено через дослідження Gartner та Fraunhofer IAO. На жаль, оцінити інструмент не вдалося через ранню помилку в процесі інсталяції, що вказувала на відсутність потрібного файлу.

Попри інтенсивне вивчення документації, розв'язати проблему в межах часових рамок проєкту не вдалося, оскільки не було надано ані підтримки з боку IBM, ані конкретних інструкцій з інсталяції. Цей досвід узгоджується із зауваженням про те, що референтні клієнти оцінюють технічну підтримку та документацію IBM нижче середнього рівня.

3.1.9 InfoZoom & IZDQ

InfoZoom – це комерційний інструмент якості даних від німецького постачальника, призначений для профілювання даних з використанням аналітики в оперативній пам'яті (in-memory analytics).

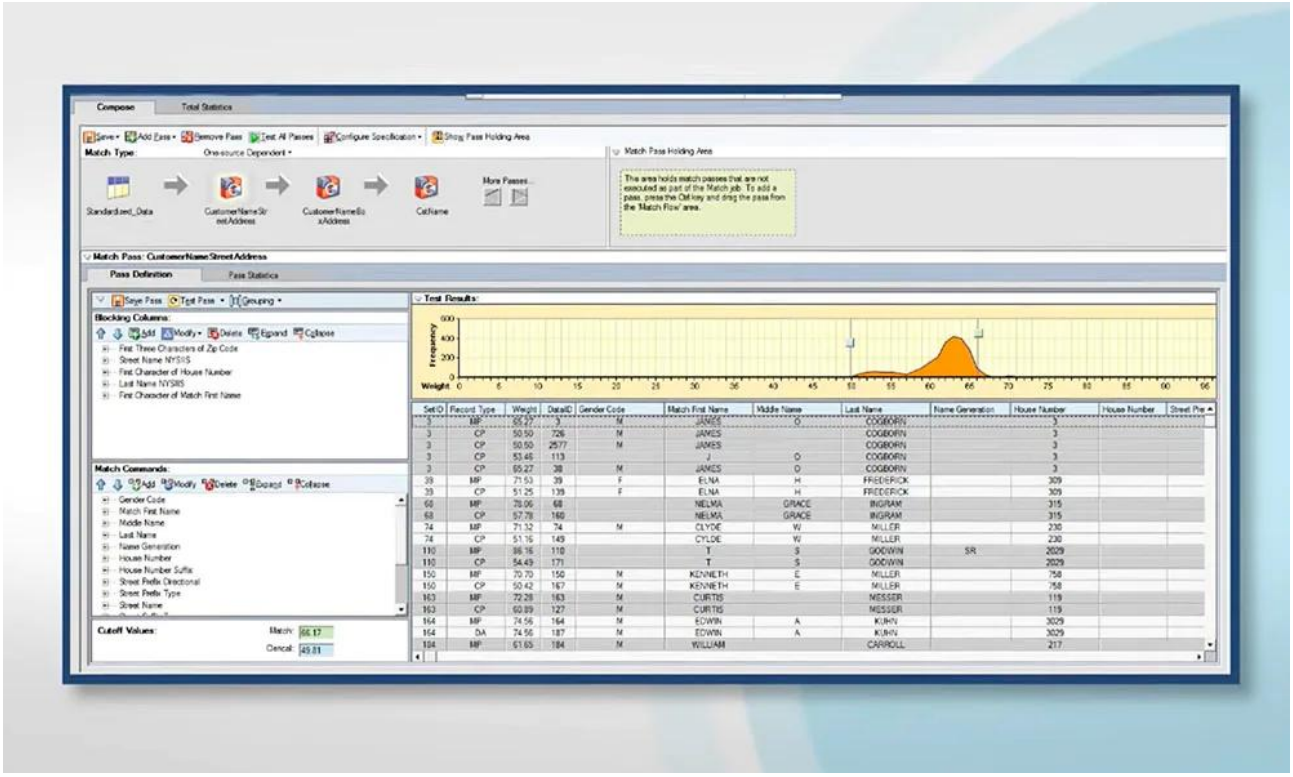

Рисунок 3.8 – Інтерфейс IBM InfoSphere Information Server for Data Quality

Досліджено InfoZoom Desktop Professional з розширенням IZDQ (InfoZoom Data Quality). Якщо InfoZoom Desktop призначений для профілювання та дослідження даних, то розширення IZDQ дозволяє користувачеві визначати правила та завдання для комплексного управління якістю даних. Загалом InfoZoom спрямований на спостереження та розуміння даних, але не підтримує жодної діяльності з очищення.

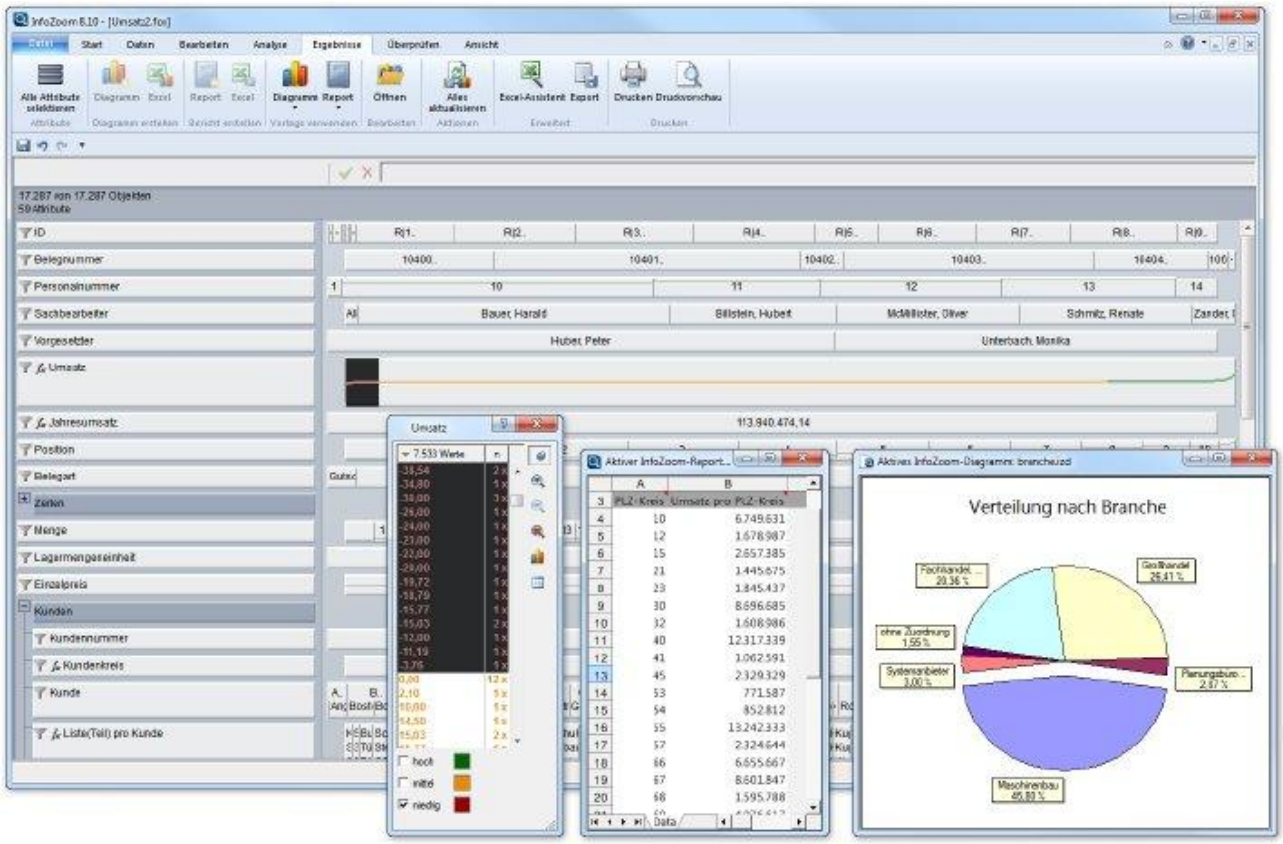

Рисунок 3.9 – Інтерфейс InfoZoom Desktop

Інтерфейс InfoZoom Desktop (див. рис. 3.9) сприймався як простий у використанні, тоді як розширення IZDQ потребує технічних знань.

3.1.10 MobyDQ

MobyDQ (раніше відомий як «Data Quality Framework») – це безкоштовне рішення якості даних з відкритим кодом, що має на меті автоматизацію перевірок якості даних під час обробки даних, збереження результатів вимірювань і метрик та запуск сповіщень у разі аномалії. Інструмент був натхненний внутрішнім проєктом якості даних у компанії-розробникові ігор (див. рис. 3.10). На відміну від Apache Griffin, MobyDQ можна встановити швидко та без ускладнень завдяки детальній документації.

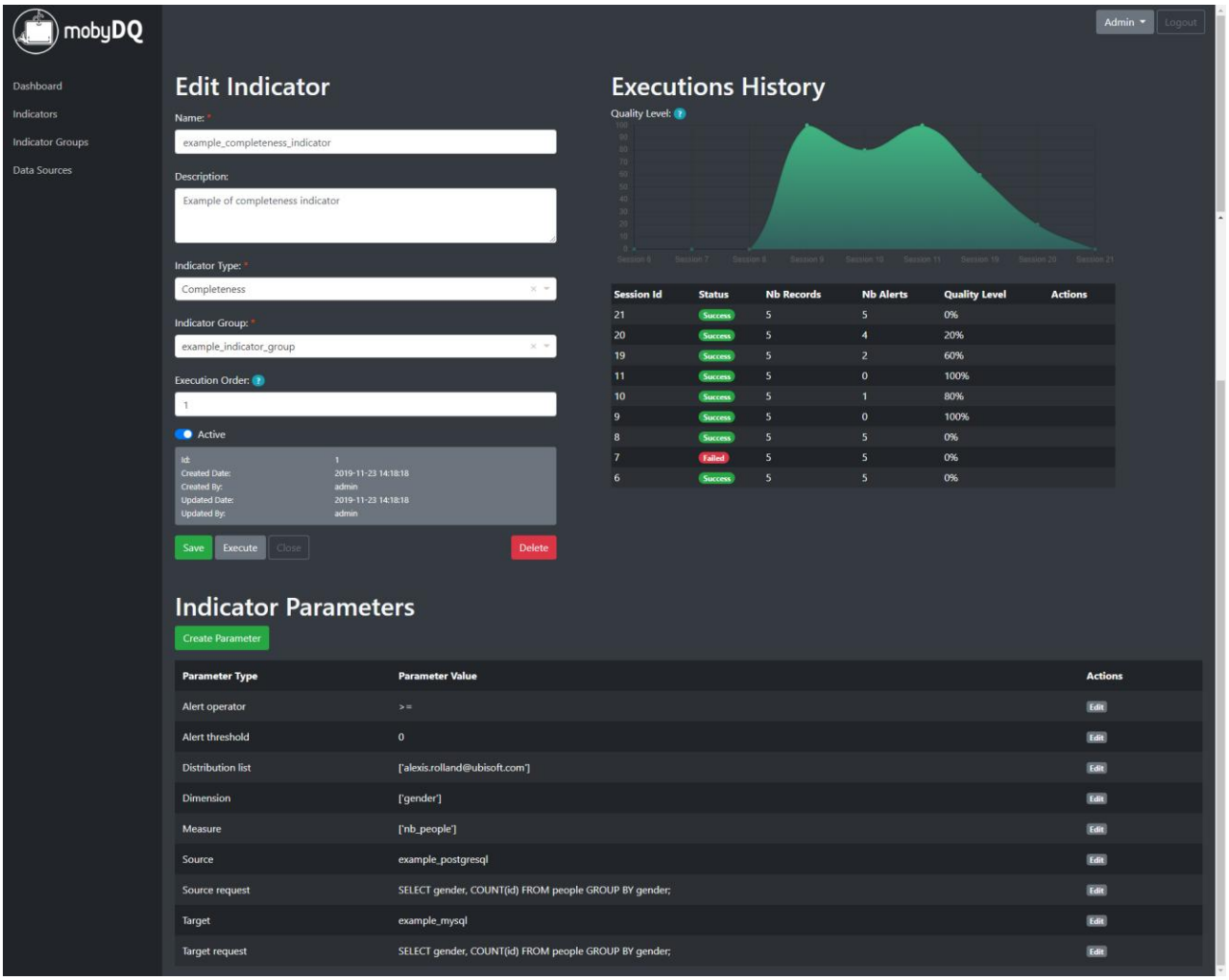

Рисунок 3.10 – Вигляд інтерфейсу MobyDQ

MobyDQ не надає функціональності профілювання даних, оскільки його фокус зосереджено на створенні, застосуванні та автоматизації перевірок якості даних.

3.1.11 OpenRefine & MetricDoc

OpenRefine (раніше Google Refine) – це безкоштовний інструмент якості даних з відкритим кодом, призначений для очищення та трансформації даних. Хоча оригінальна функціональність інструмента не повністю узгоджується з фокусом дослідження, його розширення MetricDoc спеціально спрямоване на оцінювання якості даних за допомогою настроюваних, придатних для

повторного використання метрик якості в поєднанні з негайним візуальним зворотним зв'язком.

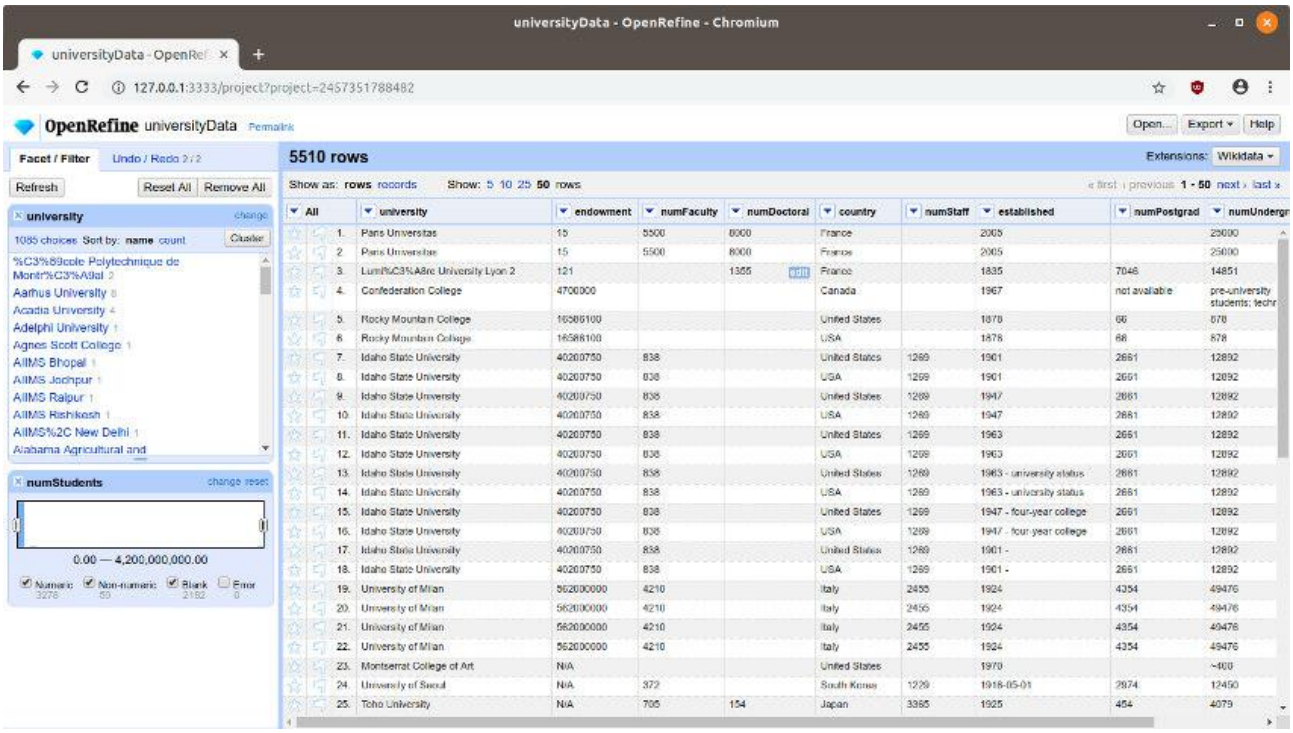

Рисунок 3.11 – Інтерфейс користувача OpenRefine

Зручність використання OpenRefine сприймалася як середня, особливо в розширенні MetricDoc, де зручність низки функцій відображала його стан як актуального дослідницького проєкту (див. рис. 3.11).

3.1.12 Oracle Enterprise Data Quality

Комерційний інструмент Oracle Enterprise Data Quality (EDQ) досліджено на основі безкоштовно доступної попередньо зібраної віртуальної машини. Окрім класичних можливостей профілювання даних, EDQ пропонує очищення даних (парсинг, стандартизацію, зіставлення й об'єднання, перевірку адрес), а також моніторинг якості даних певною мірою. Графічний інтерфейс сприймався як середній, а основним недоліком була негнучка можливість підключення джерел даних (див. рис. 3.12).

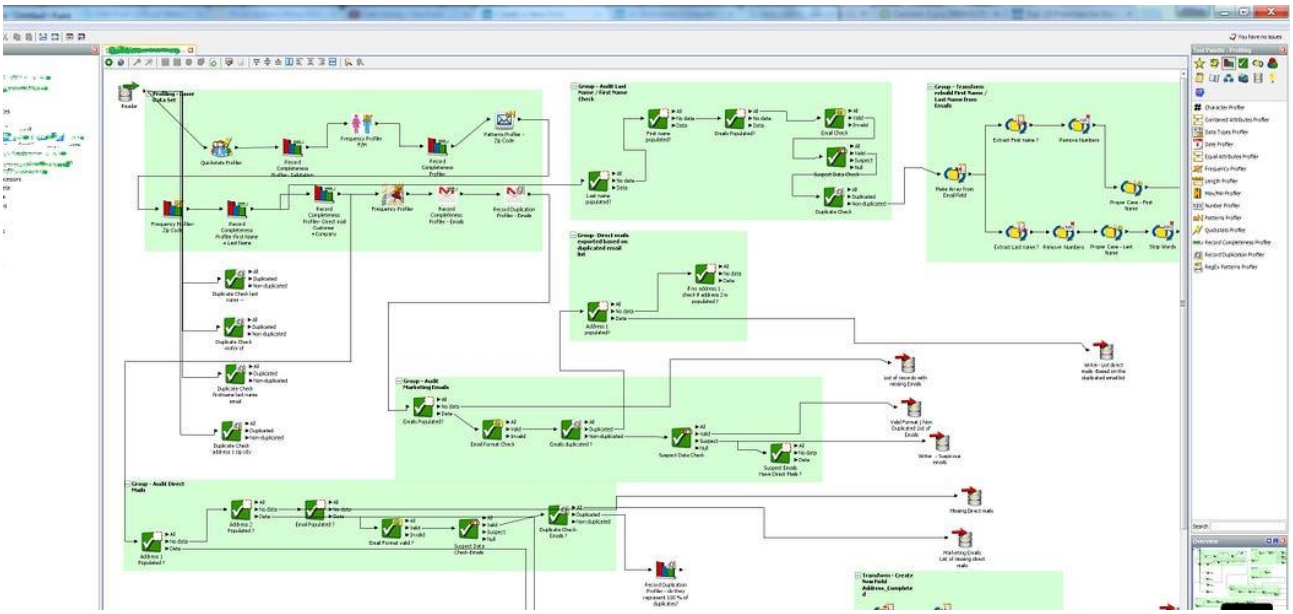

Рисунок 3.12 – Інтерфейс користувача програми Oracle Enterprise Data Quality

На відміну від інших інструментів, Oracle EDQ потребує створення «знімка» (snapshot) фактичного підключення до даних перед будь-яким завданням профілювання або вимірювання якості даних, що перешкоджає автоматичному оновленню джерела даних.

3.1.13 SAS Data Quality та Talend Open Studio

SAS Data Quality – комерційний інструмент від компанії SAS, досліджений у пробній версії, що поєднує SAS Base та SAS Data Quality Desktop. Інструмент надає можливості профілювання, стандартизації та зіставлення даних (див. рис. 3.13). Talend Open Studio for Data Quality – інструмент з відкритим кодом, який підтримує профілювання даних, зокрема аналіз кореляцій (хоча й частково, без обчислення коефіцієнта кореляції), та надає широкі можливості профілювання. Обидва інструменти посилаються на набір вимірів якості даних у своїх методологіях та документації, хоча наявність загальнозастосовних метрик якості даних у них обмежена.

Загальна інформація по властивостях наведених інструментів показана у таблиці 3.1

Таблиця 3.1 – Основні характеристики аналізованих інструментів

Інструмент	**Базове одностовпцеве профілювання**	**Аналіз кореляцій**	**Виявлення викидів (Outliers)**	**Особливості та обмеження за текстом**
1	**2**	**3**	**4**	**5**
Apache Griffin	Не підтримує	Не підтримує	Не згадується	Один із двох інструментів, який взагалі не підтримує профілювання даних.
MobyDQ	Не підтримує	Не підтримує	Не згадується	Один із двох інструментів, який взагалі не підтримує профілювання даних.
Aggregate Profiler	Підтримує	Повна підтримка	Лише візуально	Єдиний інструмент із повною підтримкою аналізу кореляцій. Візуалізація викидів для числових значень через квантильний графік, стовпчасту діаграму або діаграму розмаху.
Talend Open Studio	Підтримує	Часткова підтримка	Не згадується	Інструмент надає можливості профілювання, стандартизації та зіставлення даних
Experian Pandora	Підтримує	Не згадується	Настроювані параметри	Пропонує низку різних типів перевірок викидів із параметрами «поріг рідкісності» та «толерантність до стандартного відхилення».

Продовження таблиці 3.1

1	2	3	4	5
Ataccama ONE	Підтримує	Не згадується	Лише візуально	Візуалізація викидів для числових значень через квантильний графік, стовпчасту діаграму або діаграму розмаху.
Datamartist	Підтримує	Не згадується	Лише візуально	Візуалізація викидів для числових значень через квантильний графік, стовпчасту діаграму або діаграму розмаху.
InfoZoom	Підтримує	Не згадується	Лише візуально	Візуалізація викидів для числових значень через квантильний графік, стовпчасту діаграму або діаграму розмаху.

3.2 Порівняння можливостей профілювання даних

Аналіз можливостей профілювання даних показав, що 11 із 13 досліджених інструментів (усі, крім Apache Griffin та MobyDQ) принаймні частково підтримують профілювання даних. Базове одностовпцеве профілювання, як-от потужності (кількість рядків, кількість і відсоток порожніх значень, кількість різних значень), підтримується всіма 11 інструментами. Проте, з огляду на сучасний стан досліджень, існує потенціал для функціонального вдосконалення щодо багатостовпцевого профілювання та виявлення залежностей.

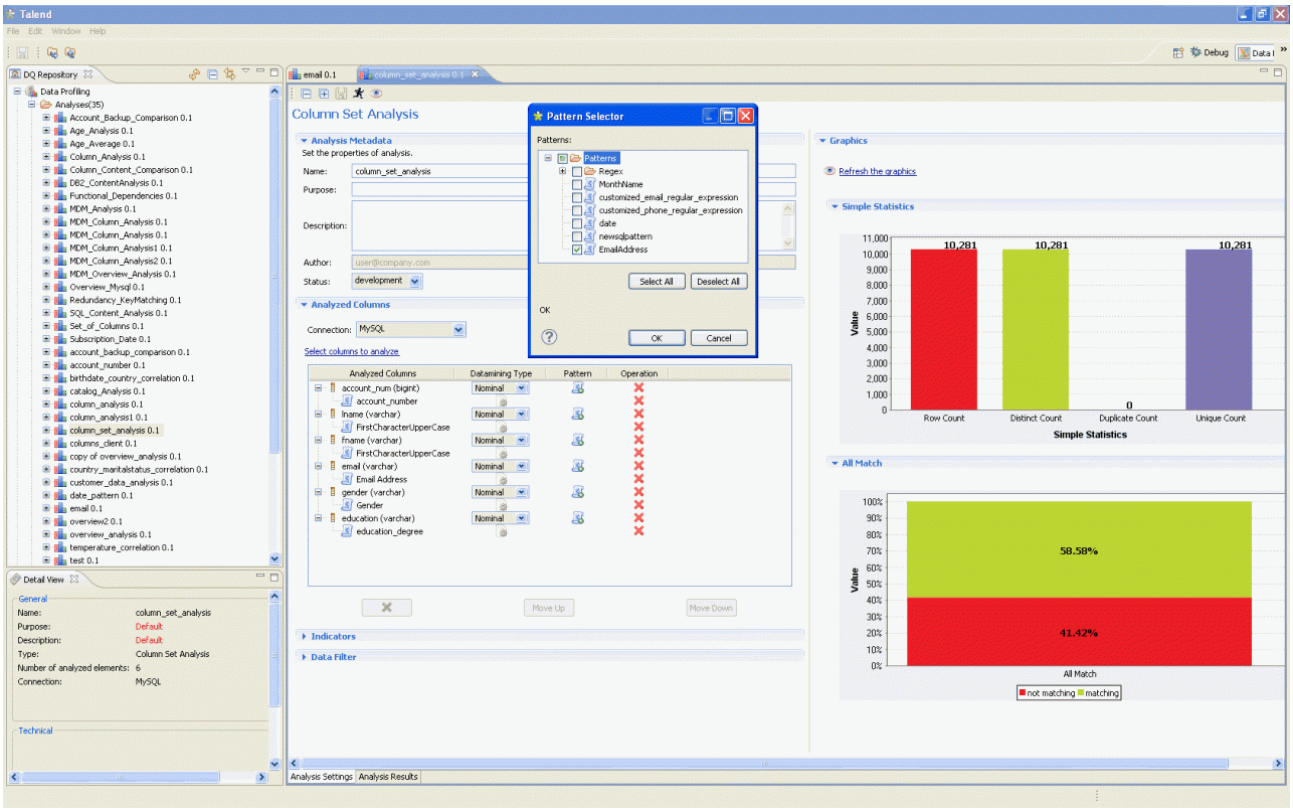

Рисунок 3.13 – Приклад роботи програми Talend Open Studio

Зокрема, виявлення залежностей (унікальні комбінації стовпців, залежності включення, функціональні залежності) комплексно підтримується лише двома інструментами. У групі багатостовпцевого профілювання точне та наближене виявлення дублікатів є дуже поширеною функцією (підтримується принаймні частково десятьма інструментами загалом), тоді як аналіз кореляцій повністю підтримується лише одним інструментом (Aggregate Profiler) і частково ще одним (Talend Open Studio). Аналіз асоціативних правил (association-rule mining) не підтримується жодним інструментом узагалі, і також немає повної підтримки кластеризації в жодному з оглянутих інструментів.

Виявлення викидів (outlier detection) реалізоване в інструментах дуже по-різному, і порівняно з поточним станом досліджень використовуються лише прості методи. Не було знайдено інструмента, що підтримує багатовимірне виявлення викидів або складніші підходи. Кілька інструментів (Aggregate Profiler, Ataccama ONE, Datamartist, InfoZoom) надають виявлення викидів для числових значень лише візуально – у вигляді квантильного графіка, стовпчастої

діаграми або діаграми розмаху. Натомість Experian Pandora пропонує низку різних типів перевірок викидів з настроюваними параметрами, як-от «поріг рідкісності» та «толерантність до стандартного відхилення».

Окремою проблемою є відсутність чіткого розмежування між поняттями «профілювання даних» та «інтелектуальний аналіз даних» (data mining). Згідно з наявними дослідженнями [1], ці теми розрізняють за об'єктом аналізу (фокус на стовпцях у профілюванні даних проти рядків в інтелектуальному аналізі) та за метою завдання (збирання технічних метаданих у профілюванні проти отримання нових знань в інтелектуальному аналізі). За словами референтних клієнтів та постачальників, такі функції, як кластеризація чи аналіз кореляцій, зазвичай не вважаються частиною профілювання даних і реалізуються в окремих аналітичних інструментах, що може пояснювати їх обмежену присутність у досліджених інструментах.

3.2.1 Розбіжності у тлумаченні базових характеристик

Детальне порівняння тестових кейсів виявило істотні розбіжності в тому, як різні інструменти інтерпретують навіть базові характеристики профілювання. Показовим є приклад обчислення квантилів. Хоча більшість інструментів (Aggregate Profiler, DataCleaner, Talend Open Studio) подають класичні квартилі, що ділять значення на чотири рівні групи, інші реалізують це поняття по-різному: SAS Data Quality застосовує деци-децилі (20 блоків), Ataccama ONE – децилі (10 блоків), а InfoZoom використовує обернену функцію, відображаючи відсоткове значення розподілу для кожного значення (функцію кумулятивного розподілу). Унаслідок цього навіть позначення відрізняються: «Q1», «нижній квартиль» та «25 %» можуть вживатися для опису одного й того самого поняття. Це ускладнює зіставлення результатів профілювання, отриманих різними інструментами.

Подібна неоднорідність спостерігається й під час розпізнавання базових типів даних. Для одного й того самого текстового атрибута різні інструменти

повертали позначення «String», «Text», «Alphanumeric», а для числового – «Number», «Decimal», «Numeric», «Long» або шаблонне представлення на кшталт «####.##». Розбіжності виявлено також у вимірюванні довжини значень: мінімальну та максимальну довжину надають майже всі інструменти, проте середню довжину подають не всі, а медіанну довжину – лише Ataccama ONE. Деякі інструменти обмежують цю функцію лише текстовими значеннями. Більше того, для атрибутів спостерігалися відмінності в точності обчислення середніх значень, що свідчить про різні підходи до реалізації навіть простих статистичних показників.

Розпізнавання шаблонів значень та їх візуалізація у вигляді гістограми підтримуються більшістю інструментів, хоча SAS обмежується лише круговими діаграмами. Кількість доступних шаблонів для виявлення семантичних типів даних і доменів суттєво різниться між інструментами: від приблизно 10–50 шаблонів (Pandora, DataCleaner, SAS) до 50–100 (Talend) і навіть 100–300 шаблонів (Informatica, Oracle). Способи подання результатів також відрізняються: одні інструменти відображають відповідні шаблони для кожного атрибута, інші – навпаки, відповідні атрибути для кожного шаблону.

3.2.2 Виявлення залежностей

Категорія виявлення залежностей має найнижче покриття серед усіх завдань профілювання і найкраще підтримується інструментами Experian Pandora та Informatica Data Quality (останній – лише у редакції Developer). Виявлення унікальних комбінацій стовпців (UCC), що є кандидатами на ключі, реалізоване по-різному: Informatica DQ пропонує повне виявлення UCC, тоді як Experian Pandora дозволяє виявляти лише ключі з одного стовпця. Обидва інструменти дають змогу задати порогове значення для наближеного виявлення UCC та ідентифікувати записи, що порушують правило, через деталізацію (drill-down). Під час тестування на одній і тій самій таблиці Informatica DQ виявив п'ять

UCC (зокрема складені комбінації атрибутів з порогом 98 %), тоді як Experian Pandora виявив лише два ключі з одного стовпця.

Виявлення залежностей включення (також відоме як виявлення зовнішніх ключів) підтримується не широко. Найкращу автоматизацію забезпечує Experian Pandora, де спершу виводяться первинні ключі та зв'язки зовнішніх ключів, а потім залежності включення відображаються графічно у вигляді діаграми Венна з можливістю деталізації до записів, що порушують правило. Informatica DQ та SAS Data Quality підтримують виявлення залежностей включення лише частково, оскільки вимагають, щоб користувач сам обрав відповідний первинний ключ та призначив його можливим кандидатам на зовнішній ключ.

Виявлення функціональних залежностей перевірялося на тій самій тестовій таблиці. Experian Pandora та Informatica DQ виявили вісім точних функціональних залежностей та ще дві при послабленні порогового значення. Talend Open Studio виконує виявлення функціональних залежностей лише частково, оскільки потребує втручання користувача для визначення множин атрибутів, і не завжди коректно виключає тривіальний випадок, коли атрибут функціонально визначає сам себе. Усі три інструменти подавали виявлені залежності в табличному форматі, проте з дещо відмінною термінологією для лівої та правої частин залежності.

3.3 Порівняння можливостей вимірювання якості даних

Під час дослідження не було знайдено жодного інструмента, який реалізує ширший спектр метрик якості даних для найважливіших вимірів, як це пропонується в наукових працях [11, 21]. Виявлені реалізації метрик якості даних мають кілька недоліків: деякі застосовні лише на рівні атрибута (без можливості агрегації), деякі потребують еталона (gold standard), який може не існувати, а деякі містять помилки реалізації.

Два інструменти з відкритим кодом, що реалізують метрики для вимірів точності (Apache Griffin) та повноти між двома таблицями (MobyDQ),

покладалися на еталонний набір даних, наданий користувачем. Apache Griffin ґрунтує свою метрику на визначенні DAMA UK [4], згідно з яким точність – це «ступінь, до якого дані правильно описують об'єкт або подію реального світу». MobyDQ спеціально спрямований на автоматизацію перевірок якості даних у конвеєрах даних, тобто на обчислення різниці між джерелом та цільовим джерелом даних, де еталон чітко визначений. Однак у сценаріях, де потрібно оцінити якість єдиного джерела даних, такі метрики не є придатними, оскільки еталон часто відсутній.

Попри відсутність готових метрик якості даних, більшість інструментів посилаються на набір вимірів якості даних у своїх посібниках користувача чи визначених методологіях. Однак перелік згадуваних вимірів і метрик якості даних у різних постачальників є дуже неоднорідним. Подальші запити щодо метрик дали дві різні відповіді від контактних осіб постачальників: одні явно зазначали, що не пропонують загальнозастосовних метрик якості даних, інші не змогли відповісти на питання про те, як саме реалізовані конкретні метрики. Це підтверджує думку про те, що люди часто не можуть сказати, як вимірювати повноту чи точність, що призводить до різних інтерпретацій та реалізацій [23].

Деякі постачальники обґрунтовували відсутність загальнозастосовних метрик якості даних двома причинами: такі метрики не є здійсненними на практиці, і клієнти їх не запитують. Кілька стратегій якості даних також вказують на те, що метрики якості даних мають створюватися користувачем та налаштовуватися під дані. Таке розуміння відповідає принципу «придатності для використання», який підкреслює суб'єктивність якості даних.

3.3.1 Вимірювання за окремими вимірами

Аналіз реалізації окремих вимірів якості даних у досліджених інструментах виявив таку картину. Точність як вимір, що потребує порівняння з реальним світом, реалізована лише в обмеженій кількості інструментів і майже завжди вимагає наявності еталонного набору даних. Через те, що такий еталон у

практичних сценаріях часто відсутній, метрики точності мають обмежену застосовність. Жоден з інструментів не реалізував складних метрик точності з нормуванням та агрегацією на різних рівнях, як це пропонується в наукових працях [11].

Повнота та унікальність на рівні атрибута виявилися єдиними характеристиками, що знаходять широке узгодження в реалізації та визначенні серед досліджених інструментів. Майже всі інструменти здатні обчислити частку порожніх значень у стовпці, що відповідає базовій метриці повноти. Проте обчислення повноти на вищих рівнях агрегації (запис, таблиця, база даних) з використанням зважених коефіцієнтів важливості атрибутів, як це описано в теоретичних метриках, в інструментах не реалізовано. Узгодженість зазвичай реалізується через механізм бізнес-правил, які користувач визначає вручну, а не через готову метрику. Своєчасність як вимір, що потребує врахування часових міток та коефіцієнтів застарівання, практично не представлена готовими метриками в жодному з інструментів.

Підсумовуючи, можна стверджувати, що поза повнотою та унікальністю на рівні атрибута жоден вимір якості даних не знаходить широкого узгодження в реалізації та визначенні на практиці. Це особливо помітно для виміру точності, який часто згадується, але вимагає еталонного набору даних, що зазвичай недоступний. Дослідження дає підстави поставити під сумнів безпосереднє використання абстрактних вимірів якості даних і пропонує зосередитися на безпосередньо вимірюваних аспектах, як-от відсутні дані та дублікати, які справді можна вимірювати автоматично.

3.3.2 Бізнес-правила як механізм вимірювання

Замість готових загальнозастосовних метрик більшість досліджених інструментів пропонують механізм визначення користувацьких бізнес-правил (business rules). Бізнес-правило задає умову, якій повинні відповідати дані, а результатом його застосування є частка записів, що задовольняють або

порушують правило. Цей підхід відповідає принципу «придатності для використання», оскільки дозволяє користувачеві кодувати предметно-залежні вимоги до якості даних. Водночас він перекладає відповідальність за визначення метрик на користувача та вимагає від нього глибокого розуміння як даних, так і предметної області. На практиці виміри якості даних здебільшого використовуються як спосіб групування таких користувацьких правил на вищому концептуальному рівні, а не як безпосередньо вимірювані величини.

3.4 Порівняння можливостей моніторингу якості даних

На відміну від попередніх досліджень, які не знаходили жодного інструмента з підтримкою моніторингу якості даних, у цьому дослідженні було виявлено наявність цієї функції. В інструментах якості даних загального призначення (наприклад, DataCleaner, Informatica, InfoZoom & IZDQ) моніторинг якості даних вважається преміальною функцією, яка є платною та надається лише в професійних версіях. Це також є причиною того, чому моніторинг якості даних досі не вивчався в попередніх роботах, що зосереджувалися на інструментах з відкритим кодом.

Винятком із цього спостереження є спеціалізований інструмент моніторингу якості даних з відкритим кодом Apache Griffin, який підтримує автоматизацію метрик якості даних, але не має попередньо визначених функцій та можливостей профілювання даних. Відкритим залишається питання про те, які саме аспекти даних слід вимірювати під час моніторингу. Згідно зі стандартом ISO 8000-8:2015 [15], повністю автоматизований моніторинг якості даних обмежується синтаксичними та семантичними аспектами, оскільки прагматичне вимірювання якості даних вимагає взаємодії з користувачами.

Важливо розрізняти два рівні безперервної перевірки даних. Моніторинг даних (data monitoring) у вузькому сенсі означає періодичну перевірку заздалегідь визначених правил і сповіщення про їх порушення. Моніторинг якості даних (DQ monitoring) у повному сенсі передбачає безперервне

вимірювання значень метрик якості даних з плином часу, накопичення історії цих значень та аналіз їх динаміки. Саме другий рівень дозволяє виявляти не лише поодинокі порушення, а й поступову деградацію якості даних, сезонні коливання та аномальні зміни, що можуть свідчити про збої в процесах постачання даних [9].

Серед досліджених інструментів можливості моніторингу реалізовані нерівномірно. Інструменти загального призначення зазвичай дозволяють планувати завдання профілювання чи перевірки правил на регулярній основі та зберігати результати, проте повноцінний аналіз історичних трендів та автоматичне виявлення аномалій у динаміці метрик підтримуються рідко й переважно в платних професійних версіях. Спеціалізовані інструменти з відкритим кодом, як-от Apache Griffin та MobyDQ, навпаки, від початку проєктувалися навколо ідеї безперервного вимірювання та сповіщень про аномалії, проте поступаються інструментам загального призначення в багатстві функцій профілювання та готових перевірок. Це свідчить про функціональний розрив між двома класами інструментів, який наразі не подолано жодним рішенням.

Перспективним напрямом розвитку моніторингу якості даних є застосування методів аналітики часових рядів до накопичених значень метрик. Це дозволило б не лише фіксувати поточний стан якості даних, а й прогнозувати тренди та завчасно виявляти раптові зміни [9]. Однак реалізація такого підходу потребує тривалого накопичення історичних даних про якість, що, своєю чергою, вимагає від інструментів сталої та автоматизованої інфраструктури моніторингу. На момент дослідження жоден з оглянутих інструментів не пропонував повноцінних можливостей прогнозної аналітики якості даних.

3.5 Узагальнення результатів огляду

Зведений аналіз досліджених інструментів дозволяє сформулювати кілька загальних спостережень. Інструменти можна умовно поділити на три групи за

основним призначенням. Перша група – комплексні інструменти загального призначення (Informatica DQ, Oracle EDQ, SAS Data Quality, Experian Pandora, DataCleaner, InfoZoom & IZDQ), які поєднують профілювання, вимірювання та частково моніторинг, але здебільшого є комерційними та потребують ліцензії для повного функціоналу. Друга група – інструменти, переважно орієнтовані на профілювання (Aggregate Profiler, Datamartist, Ataccama ONE profiler), які добре виконують аналіз даних, але обмежені у вимірюванні та моніторингу. Третя група – спеціалізовані інструменти моніторингу з відкритим кодом (Apache Griffin, MobyDQ), які зосереджені на безперервному вимірюванні, але не надають профілювання.

Щодо зручності встановлення та використання, досвід дослідження показав значну варіативність. Комерційні інструменти загального призначення зазвичай супроводжувалися якіснішою документацією та підтримкою, хоча й тут траплялися винятки: інсталяцію одного з комерційних продуктів не вдалося завершити взагалі. Серед інструментів з відкритим кодом інсталяція також різнилася за складністю – від швидкої й безпроблемної (MobyDQ) до надзвичайно трудомісткої через велику кількість залежностей (Apache Griffin). Якість користувацьких інтерфейсів коливалася від інтуїтивно зрозумілих, придатних для бізнес-користувачів, до таких, що потребують технічних навичок, зокрема вміння писати SQL-запити.

Загальний висновок огляду полягає в тому, що жоден з досліджених інструментів не реалізує повного спектра можливостей, описаних у науковій літературі. Сильні сторони одних інструментів (наприклад, виявлення залежностей у Experian Pandora чи близькість до наукового погляду на метрики в Informatica DQ) поєднуються зі слабкими сторонами в інших аспектах. Це означає, що вибір інструмента має ґрунтуватися на конкретних потребах сценарію використання, а практикам слід чітко усвідомлювати функціональні обмеження кожного рішення.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1 Питання щодо охорони праці

Тема кваліфікаційної роботи магістра пов’язана із дослідженням методів і засобів тестування програмного забезпечення комп’ютерних систем з використанням алгоритмів машинного навчання. Такі роботи передбачають використання комп’ютерної техніки на етапах формування пояснювальної записки, налаштування засобів автоматизації процесу управління тестуванням з використанням алгоритмів машинного навчання. Тому, при виконанні робіт з тестування програмних складових комп’ютерних систем необхідно враховувати вимоги з охорони праці при експлуатації комп’ютерної техніки.

В Україні діють ряд законів, нормативних документів та актів, які регулюють процеси забезпечення та управління охороною праці у різних галузях народного господарства. До них належать: Конституція України, Закони України "Про охорону праці", "Про охорону здоров'я", "Про пожежну безпеку", "Про забезпечення санітарного та епідемічного благополуччя населення", "Про загальнообов'язкове державне соціальне страхування від нещасного випадку на виробництві та професійного захворювання, які спричинили втрату працездатності", Кодекс законів про працю України (КЗпП).

Однією з основних вимог до приміщень, де робочі місця обладнані комп’ютерною технікою і планується використання програмного комплексу для забезпечення процесу тестування ПЗ, є вимоги щодо площі, яка відводиться на один ПК. При проектуванні автоматизованих робочих місць тестувальників програмного забезпечення необхідно дотримуватись вимог щодо розміщення комп’ютерів. На один ПК передбачено площу 6 м^2 та об’єм 20 м^3.

Однак робота з комп’ютером включає різні завдання, які об'єднуються такими загальними чинниками, як те, що робота проводиться в сидячому положенні і вимагає уважного, неперервного та іноді тривалого спостереження.

Перше правило, якого варто дотримуватись тестувальникам програмного забезпечення стосується правильного облаштування робочого столу. При цьому слід передбачити наступні його параметри: фіксована висота – 720 мм, забезпечення необхідного простору для рук по висоті, ширині і глибині, в області сидіння не повинно бути шухляд.

Друге правило визначає облаштування робочого стільця: можливість регулювання висоти стільця, забезпечення обертання конструкції стільця. У приміщеннях з ПК, на яких планується виконання задач з тестування програмних складових комп'ютерних систем, яскравість знаків і яскравість фону дисплею повинна бути спроектована таким чином, щоб не було великої відмінності з яскравістю навколишнього середовища, але знаки повинні чітко розпізнаватися на відстані читання. Характеристики освітлення, зокрема у приміщеннях, де експлуатується ПК, повинні відповідати ДБН В.2.5-28-2006 "Природнє і штучне освітлення". Основні вимоги даного нормативного документу стосуються забезпечення наступних вимог:

- освітлення з лівої сторони;
- рівномірне освітлення всього робочого простору;
- комп'ютерна техніка встановлюється у місцях, віддалених від вікон;
- встановлення непрямого штучного освітлення;
- світло, що поступає через вікна, «пом'якшують» за допомогою штор;
- робоче місце організовується так, щоб напрям погляду був паралельним фронту вікон.

Ще одне правило, якого слід дотримуватись тестувальникам програмного забезпечення, передбачає оптимальний метод роботи, що полягає у передбачені зміни завдань і навантажень, дотримання перерви в роботі: 5 хвилин через 1 годину роботи біля дисплея або 10 хвилин після 2-х годин роботи біля дисплея. Вимоги цього правила регламентовані нормативним документом «Державні санітарні правила і норми роботи з візуальними дисплейними терміналами електроннообчислювальних машин».

При створенні сприятливих умов для підвищення продуктивності і зменшення напруги значну роль грають чинники, що характеризують стан навколишнього середовища: мікроклімат приміщення, рівень шуму і освітлення.

Рекомендована величина відносної вологості, яка повинна бути забезпечена у приміщеннях з експлуатації програмного комплексу поведінкового тестування програмних складових комп'ютерних систем, повинна відповідати НПАОП 0.007.15-18 і становити 65 – 70%. При цьому робоче місце повинно бути добре вентильованим.

У даний час з погляду шумового навантаження досягнуто значного прогресу. Рівень шуму в приміщеннях (приблизно 40 Дб) не перевищує допустимого рівня, незалежно від кількості використовуваного обладнання. Для приміщень, в яких експлуатується програмний комплекс підтримки запропонованих методів поведінкового тестування, потрібно забезпечити виконання вимог пожежної безпеки, які визначені Правилами пожежної безпеки в Україні, НПАОП 0.00-7.1518 «Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями».

Будівлі і ті їх частини, в яких розташовуються ПК можуть належати до II ступеня вогнестійкості. Над та під приміщеннями, де розташовуються ПК, а також у суміжних з ними приміщеннях не дозволяється розташування приміщень категорій А і Б за вибухопожежною небезпекою. Приміщення категорії В повинні бути відділеними від приміщень з ПК протипожежними стінами.

Таким, чином при дослідженні методів і засобів поведінкового тестування програмних складових комп'ютерних систем, встановлено, що найбільш повним нормативним документом щодо охорони праці користувачів ПК, до яких належать тестувальники програмного забезпечення, є НПАОП 0.00-7.15-18 «Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями». Дотримання вимог, які неведені у цьому документі, сприяє зниженню негативного впливу ПК, його компонентів та інших зовнішніх

пристроїв на тестувальників, які проводять роботи щодо перевірки правильності функціонування програмних складових комп’ютерних систем.

4.2 Підвищення стійкості роботи об'єктів господарської діяльності у воєнний час

На основі вивчення факторів, які впливають на стійкість роботи об'єктів господарської діяльності, та оцінки стійкості елементів і галузей виробництва проти уражаючих факторів ядерної, хімічної і біологічної зброї, стихійних лих і виробничих аварій, необхідно завчасно організувати і провести організаційні, інженерно-технічні й технологічні заходи для підвищення стійкості роботи.

Здійснення організаційних заходів передбачає завчасну підготовку всіх структур цивільного захисту, служб і формувань до надзвичайних ситуацій, в тому числі і військових дій.

Вжиттям технологічних заходів підвищується стійкість роботи об'єктів шляхом змінювання технологічних процесів, режимів, можливих в умовах різних надзвичайних ситуацій.

Інженерно-технічні заходи мають забезпечити підвищену стійкість виробничих споруд, технологічних ліній, устаткування, комунікацій об'єкта до впливу уражаючих факторів під час військових дій.

При проведенні цих заходів необхідно враховувати конкретні умови об'єкта народного господарства. Проте є загальні організаційні інженерно-технічні заходи, які мають проводитись на всіх об'єктах.

Одним з найбільш важливих завдань в умовах воєнного часу і надзвичайних ситуацій є забезпечення захисту людей та їх життєдіяльності.

Для підвищення стійкості об’єктів господарювання та захисту людей необхідно:

– створити на об'єкті надійну систему оповіщення про загрози нападу противника, радіоактивне забруднення, хімічне і біологічне зараження, загрозу стихійного лиха і виробничої аварії.

– організувати розвідку і спостереження за радіоактивним забрудненням, хімічним і біологічним зараженням;

– організувати гідрометеорологічне спостереження за рівнем води, напрямком і швидкістю вітру, рухом і поширенням хмари радіоактивного забруднення, сильнодіючих отруйних речовин і отруйних речовин.

– створити фонд захисних споруд ЦО, запасів засобів індивідуального захисту і забезпечення своєчасної видачі їх населенню.

– завчасно підготуватись до масової санітарної обробки населення і знезаражування одягу;

– організувати взаємодію з установами охорони здоров'я для медичного обслуговування населення в умовах воєнного часу.

Також в умовах воєнного часу необхідно провести підготовку до евакуації населення, розміщеного в зонах можливих руйнувань і катастрофічного затоплення. Це передбачає завчасну підготовку місць евакуації, організацію прийому евакуйованого населення на територію населених пунктів.

Окрім цього, необхідно забезпечити постачання продуктів харчування, питної води, предметів першої необхідності та провести заходи щодо морально-психологічної підготовки населення до виживання в умовах воєнного часу, забезпечити процес чіткого інформування про обстановку та правила дій і поведінки населення в надзвичайних ситуаціях воєнного часу.

Для забезпечення стійкості роботи об'єктів повинні проводитись інженерно-технічні заходи на мережах комунального господарства з метою захисту джерел тепла із заглибленням у ґрунт комунікацій. Котельні слід розміщувати в спеціальному окремо розміщеному приміщенні.

Якщо об'єкт одержує тепло з міської теплоцентралі, необхідно провести заходи для забезпечення стійкості трубопроводів і розподільних пристроїв, підведених до об'єкта.

Теплова мережа має будуватися за кільцевою системою з прокладанням труб у спеціальних каналах зі з'єднанням паралельних ділянок. Для відключення пошкоджених ділянок мають бути встановлені запірно-регулюючі засувки,

вентилі та ін. Ці пристосування необхідно розміщувати в оглядових колодязях, на території, що не завалюється при руйнуванні будівель.

Система каналізації має будуватись окремо: одна для дощових, друга для промислових і господарських вод. На об'єкті має бути не менше двох виводів з підключенням до міських каналізаційних колекторів, а також виводи і колодязі з аварійними засувками на об'єктових колекторах з інтервалом 50 м на території, що не завалюється, для аварійного скидання неочищеної води в найближчі штучні та природні заглиблення.

На деяких промислових об'єктах є системи для забезпечення технології виробництва: для подання кисню, аміаку, стиснутого повітря та інших рідких і газових реактивів. Для цих систем розробляють заходи для попередження виникнення вторинних факторів зброї, стихійних лих та виробничих аварій і катастроф.

Створення резерву енергетичних потужностей за рахунок автономних пересувних електростанцій, а також місцевих джерел електроенергії. Підготовка автономних електростанцій до роботи за спеціальним режимом (графіком) для забезпечення технологічних процесів виробництва, для яких неможливі тривалі перерви в електропостачанні.

З метою попередження аварій на електричних мережах необхідно установити автоматичну систему відключення при виникненні перенапруги. Повітряні лінії електропостачання замінити на підземно-кабельні.

Створення необхідних запасів (резервів) паливно-мастильних матеріалів та інших видів палива й організація їх безпечного зберігання.

Щоб не допустити зупинки підприємства через дефіцит палива, необхідно підготуватись для роботи на різних видах палива: нафта, вугілля, газ.

Для підвищення стійкості забезпечення водою слід провести такі заходи. Необхідно створити основні і резервні джерела водопостачання. Як резервне джерело краще мати артезіанську свердловину, яку необхідно підключити до системи водопостачання. Крім того, воду можна брати з близько розміщеної

природної водойми або спорудити штучну водойму чи резервуари з обладнанням пристроїв для збору і перекачування води.

Всі ділянки водопостачання повинні бути заглиблені в ґрунт з обладнанням пожежних гідрантів і пристроїв для відключення пошкоджених ділянок. Локальні мережі водопостачання окремих великих підприємств варто з'єднати із загальноміською системою водопостачання в єдине кільце.

Підвищенню стійкості забезпечення водою сприяє подавання води безпосередньо в мережу поза водонапірними баштами, спорудження обвідних ліній для подання води поза пошкодженими спорудами.

Завчасне вжиття заходів захисту вододжерел, водопровідних споруд, свердловин і шахтних колодязів від забруднення радіоактивними речовинами, зараження хімічними і біологічними засобами.

Підготовка меліоративних, гідротехнічних та іригаційних споруд і систем до експлуатації в надзвичайних умовах.

Для забезпечення виробництва продукції необхідні електроенергія, паливо, мастила, засоби захисту рослин, міндобрива, профілактичні й лікувальні препарати ветеринарної медицини, запасні частини, сировина та інші матеріально-технічні засоби. Забезпечення об'єктів цими ресурсами дасть можливість випускати необхідну продукцію в надзвичайних умовах мирного і воєнного часу. Тому повинні проводитись такі заходи, які б забезпечили стійкість постачання і сприяли підвищенню захисту мережі електро-, водо-, газопостачання, транспортних комунікацій і джерел постачання всім необхідним для забезпечення функціонування галузей сільського господарства в надзвичайних умовах.

З метою попередження аварій на електричних мережах необхідно встановити автоматичну систему відключення перенапруги. Повітряні лінії електропостачання слід замінити на підземно-кабельні.

Газ використовується як паливо і на хімічних підприємствах у технологічному процесі. Для безперебійного забезпечення газом, газові мережі необхідно підводити до об'єкта з двох напрямків, які мають бути з'єднані в єдине

кільце з обладнанням для можливого дистанційного автоматичного управління й у разі необхідності відключення пошкоджених ділянок.

На великих підприємствах необхідно мати підземні ємності із закачаним резервним газом.

На підприємствах, де використовується пара, необхідно захистити джерела його постачання, заглибити в ґрунт комунікації паропостачання і встановити запірні пристосування.

Запас резервних матеріалів необхідно розраховувати на такі строки роботи підприємства, за які можливе відновлення регулярного постачання.

Передбачити, на випадок перебоїв в постачанні підприємствамисуміжниками, створення місцевих матеріалів, сировини для виготовлення комплектуючих виробів і інструментів силами свого підприємства.

Для підвищення стійкості та забезпечення збереження (відновлення) будівель і споруд в умовах воєнного часу необхідно:

– провести оцінку можливих ступенів руйнування будівель і споруд господарства населеного пункту, визначити обсяг невідкладних ремонтних робіт, потреби в будівельних матеріалах.

– створити і підготувати спеціальні формування для ремонтновідновних, будівельних та інших робіт на об'єкті.

– розробити комплекс протипожежних заходів, які виключали б можливість виникнення масових пожеж.

Для забезпечення надійності системи управління і зв'язку потрібно організувати захищений пункт управління, забезпечити його засобами зв'язку, які б дали можливість швидко доводити сигнали ЦЗ до всіх виробничих підрозділів і населення у місцях проживання. При цьому необхідно здійснити планування збору даних про обстановку, передачу команд і розпоряджень в умовах впливу на об'єкт уражаючих факторів. Для підвищення стійкості системи управління і зв’язку в умовах воєнного часу необхідно організувати використання радіозасобів, засобів телефонного зв'язку, а також забезпечити зв’язок із колонами евакуйованого населення, що перебувають у дорозі, і

відповідальними особами, які супроводжують їх під час евакуації, забезпечити дублювання ліній і каналів зв'язку.

Отже, підвищення стійкості роботи об'єктів господарської діяльності у воєнний час можна забезпечити шляхом організації і проведення сукупності заходів, які включають організаційні, інженерно-технічні й технологічні заходи. Надзвичайно важливим є планування та організація захисту комплексів критичної інфраструктури, зокрема водопостачання, водовідведення, газопостачання та зв’язку.

ВИСНОВКИ

У межах цього дослідження проведено систематичний пошук, у результаті якого ідентифіковано декілька сотень програмних інструментів, присвячених темі «якість даних». За допомогою шести попередньо визначених критеріїв виключення відібрано 17 інструментів для глибшого дослідження, з яких оцінено 13 щодо списку з 43 вимог, розподілених за трьома категоріями: профілювання даних, вимірювання якості даних та безперервний моніторинг якості даних. Хоча ринок інструментів якості даних безперервно змінюється, проведене дослідження дає комплексне уявлення про сучасний стан інструментів якості даних та про те, як вимірювання якості даних наразі сприймається на практиці компаніями на противагу науковим дослідженням.

Основні результати дослідження можна узагальнити так. По-перше, попри припущення про те, що ринок інструментів якості даних усе ще перебуває на стадії розвитку, систематичний пошук виявив значну кількість інструментів, більшість з яких ніколи не включалися до жодного з наявних оглядів. Приблизно половина виявлених інструментів виявилися предметно-залежними.

По-друге, більшість оглянутих інструментів тією чи іншою мірою підтримували профілювання даних, однак, з огляду на стан досліджень, існує потенціал для функціонального вдосконалення профілювання даних, особливо щодо багатостовпцевого профілювання та виявлення залежностей. Виявлено потребу як у більшій автоматизації профілювання даних, так і в чіткому декларуванні та поясненні виконуваних обчислень і алгоритмів. У низці інструментів графіки генерувалися або викиди виявлялися без чіткого декларування використаного порогового значення чи функції відстані.

По-третє, не було знайдено інструмента, який реалізує ширший спектр метрик якості даних для найважливіших вимірів, як це пропонується в наукових працях. Виявлені реалізації метрик мають низку недоліків: деякі застосовні лише на рівні атрибута, деякі потребують еталона, який може не існувати, а деякі містять помилки реалізації. Дослідження дає підстави поставити під сумнів

поточне використання вимірів і метрик якості даних: наукові зусилля щодо безпосереднього вимірювання вимірів якості даних за допомогою єдиної загальнозастосовної метрики мають незначну практичну релевантність та майже не трапляються в інструментах. На практиці виміри якості даних використовуються для групування предметно-залежних правил якості даних на вищому рівні.

По-четверте, в інструментах якості даних загального призначення моніторинг якості даних вважається преміальною функцією, яка є платною та надається лише в професійних версіях. Винятком є спеціалізовані інструменти моніторингу з відкритим кодом, як-от Apache Griffin або MobyDQ, які підтримують автоматизацію правил, але не мають попередньо визначених функцій та можливостей профілювання даних.Отримані результати мають практичну цінність як для фахівців з якості даних, що допомагає їм обрати найбільш відповідний інструмент для конкретного сценарію використання, так і для подальших наукових досліджень, оскільки вони висвітлюють поточні можливості сучасних інструментів якості даних та виявляють розрив між теоретичними напрацюваннями й практичними реалізаціями.

Перспективними напрямами подальшої роботи є: розроблення практичної методології якості даних, що зосереджується на безпосередньо вимірюваних аспектах якості даних, а не на абстрактних вимірах; дослідження потенціалу автоматизованого профілювання даних «з коробки» з чітким декларуванням використаних параметрів; застосування аналітики часових рядів для дослідження результатів моніторингу якості даних з метою прогнозування трендів та раптових змін у якості даних; а також подальше дослідження виключених предметно-залежних інструментів якості даних щодо їх предметних областей та функціонального обсягу. З огляду на ажіотаж навколо штучного інтелекту й машинного навчання, порівняно невелика кількість клієнтів провідних постачальників інструментів якості даних свідчить про значний нереалізований потенціал застосування таких інструментів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Abedjan Z., Golab L., Naumann F. Profiling relational data: a survey. The VLDB Journal. 2015. Vol. 24, No. 4. P. 557–581.

2. Abedjan Z., Golab L., Naumann F., Papenbrock T. Data Profiling. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2019. 154 p.

3. Batini C., Scannapieco M. Data and Information Quality: Dimensions, Principles and Techniques. Cham : Springer International Publishing, 2016. 500 p.

4. Askham N., Cook D., Doyle M. et al. The Six Primary Dimensions for Data Quality Assessment. DAMA UK Working Group, 2013. 17 p.

5. Chrisman N. R. The role of quality information in the long-term functioning of a geographic information system. Cartographica. 1983. Vol. 21, No. 2–3. P. 79–88.

6. Chien M., Jain A. Magic Quadrant for Data Quality Tools. Technical Report. Stamford : Gartner, Inc., 2019.

7. Cichy C., Rass S. An overview of data quality frameworks. IEEE Access. 2019. Vol. 7. P. 24634–24648.

8. Dasu T., Johnson T. Exploratory Data Mining and Data Cleaning. New York : John Wiley & Sons, Inc., 2003. 224 p.

9. Ehrlinger L., Wöß W. Automated data quality monitoring. Proceedings of the 22nd MIT International Conference on Information Quality (ICIQ). Little Rock, 2017. P. 15.1–15.9.

10. Ge M., Helfert M. A review of information quality research. Proceedings of the 12th International Conference on Information Quality (ICIQ). Cambridge : MIT, 2007. P. 76–91.

11. Heinrich B., Hristova D., Klier M., Schiller A., Szubartowicz M. Requirements for data quality metrics. Journal of Data and Information Quality. 2018. Vol. 9, No. 2. P. 12:1–12:32.

12. Hildebrand K., Gebauer M., Hinrichs H., Mielke M. Daten- und Informationsqualität. 3rd ed. Wiesbaden : Springer Vieweg, 2015. 416 p.

13. IEEE Standard for a Software Quality Metrics Methodology. IEEE Std 1061-1998. Institute of Electrical and Electronics Engineers, 1998.

14. ISO/IEC 25012:2008. Systems and Software Engineering — Systems and Software Quality Requirements and Evaluation (SQuaRE) — Data Quality Model. Geneva : ISO, 2008.

15. ISO 8000-8:2015. Data Quality — Part 8: Information and Data Quality: Concepts and Measuring. Geneva : ISO, 2015.

16. Kitchenham B. Procedures for Performing Systematic Reviews. Technical Report TR/SE-0401. Keele University, 2004. 33 p.

17. Maydanchik A. Data Quality Assessment. Bradley Beach : Technics Publications, LLC, 2007. 336 p.

18. Moore S. How to Create a Business Case for Data Quality Improvement. Stamford : Gartner, Inc., 2018.

19. Naumann F. Data profiling revisited. ACM SIGMOD Record. 2014. Vol. 42, No. 4. P. 40–49.

20. Otto B., Österle H. Corporate Data Quality: Prerequisite for Successful Business Models. Berlin : Springer Gabler, 2016. 232 p.

21. Piro A. Datenqualität als kritischer Erfolgsfaktor: Vorgehensmodelle und Konzepte. Lohmar : Josef Eul Verlag, 2014.

22. Wang R. Y. A product perspective on total data quality management. Communications of the ACM. 1998. Vol. 41, No. 2. P. 58–65.

23. Sebastian-Coleman L. Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. Waltham : Morgan Kaufmann, 2013. 376 p.

24. Wang R. Y., Strong D. M. Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems. 1996. Vol. 12, No. 4. P. 5–33.

25. Selvage M. Y., Judah S., Jain A. Magic Quadrant for Data Quality Tools. Technical Report. Stamford : Gartner, Inc., 2017.

26. Голінько В. І. Охорона праці в галузі інформаційних технологій: навч. посіб. / В. І. Голінько, М. Ю. Іконніков, Я. Я. Лебедєв; М-во освіти і науки

України, Держ. вищий навч. закл. "Нац. гірн. ун-т". - Дніпропетровськ: НГУ, 2015. - 246 с.

27. Микитишин А. Г. Застосування методів машинного навчання для класифікації даних в комп'ютеризованих системах керування / Андрій Григорович Микитишин, І. С. Дідич, Р. І. Яцишин // Тези XIII МНПК „Актуальні задачі сучасних технологій", 11-12 грудня 2024 року. — Т. : ФОП Паляниця В. А., 2024. — С. 14–16. — (Нові матеріали, міцність і довговічність елементів конструкцій).

28. Карнаухов, О. К. (2024). Дослідження розробки електронного кабінету абітурієнта ТНТУ ім. І. Пулюя. Тези доповідей V міжнародної науково-практичної конференції учених та студентів "Цифрова економіка як фактор інновацій та сталого розвитку суспільства", 46-47.

29. Кучеренко О. А. Особливості передобробки даних для методів прогнозування / О. А. Кучеренко, О. О. Кучеренко // ІМСТТ, 13-14 грудня 2023 року. — Т. : ТНТУ, 2023. — С. 72. — (Інформаційні системи та технології, кібербезпека).

30. Яцишин В. Процеси забезпечення якості даних при проектуванні систем машинного навчання / В. Яцишин, Ю. Журихін // Матеріали V науково-технічної конференції „Інформаційні моделі, системи та технології", 1-2 лютого 2018 року. — Т. : ТНТУ, 2018. — С. 68. — (Секція 3. Комп'ютерні системи та мережі).

31. Чорновус, Р. М. Визначення якості тестування програмного забезпечення та аналіз отриманих даних. Матеріали конференції. Тернопіль: ТНТУ, 2017. URL: http://elartu.tntu.edu.ua/handle/123456789/18917.

32. Яцишин В. В. Оцінювання якості даних для систем машинного навчання / В. В. Яцишин, Ю. О. Журихін // Збірник тез доповідей VI Міжнародної науково-технічної конференції молодих учених та студентів „Актуальні задачі сучасних технологій", 16-17 листопада 2017 року. — Т. : ТНТУ, 2017. — Том 2. — С. 196. — (Комп'ютерно-інформаційні технології та системи зв'язку).

33. Гандзюк М.П. Основи охорони праці: Підручник. 4-е вид./Гандзюк М.П., Желібо Є.П., Халімовський М.О. - Київ: Каревела, 2008. – 384с.

34. Техноекологія та цивільна безпека. Частина «Цивільна безпека»: Навчальний посібник; укл.: Стручок В. С. Тернопіль: ФОП Паляниця В.А., 2022. 150 с.

35. Безпека в надзвичайних ситуаціях. Методичний посібник для здобувачів освітнього ступеня «магістр» всіх спеціальностей денної та заочної (дистанційної) форм навчання / укл.: Стручок В. С. Тернопіль: ФОП Паляниця В. А., 2022. 156 с.

36. Умови праці працівників, які використовують у роботі персональні комп'ютери. Zolochiv.Net. URL: https://zolochiv.net/umovy-pratsi-pratsivnykiv-iaki-vykorystovuiut-u-roboti-personal-ni-komp-iutery/ (дата звернення: 25.10.2024).

ДОДАТКИ

Додаток А

Тези доповіді

Міністерство освіти і науки України
Тернопільський національний технічний університет
імені Івана Пулюя
Маріборський університет (Словенія)
Технічний університет в Кошице (Словаччина)
Каунаський технологічний університет (Литва)
Львівський національний університет
імені Івана Франка
Гірничо-металургійна академія ім. Станіслава Сташиця (Польща)
Луцький національний технічний університет
Чернівецький національний університет
імені Юрія Федьковича
Вроцлавський економічний університет (Польща)
Університет технологій та економіки
імені Хелени Ходковської (Польща)
Донбаська державна машинобудівна академія

Студентське наукове товариство

IX МІЖНАРОДНА

студентська науково - технічна конференція

"ПРИРОДНИЧІ ТА ГУМАНІТАРНІ НАУКИ. АКТУАЛЬНІ ПИТАННЯ"

24-25 квітня 2026 р.

(збірник тез конференції)

Тернопіль 2026

УДК 004.6:005.6
Боб О., Мага С., Лотоцький Д., Боднарчук І.
Тернопільський національний технічний університет імені Івана Пулюя

ДО ПИТАННЯ ЯКОСТІ ДАНИХ В НАУКОВИХ ДОСЛІДЖЕННЯХ ТНТУ

O.O. Bob, S.Yu. Maha, D.V. Lototskyi, I.O. Bodnarchuk
Ternopil Ivan Puluj National Technical University

ON THE ISSUE OF DATA QUALITY IN SCIENTIFIC RESEARCH AT TNTU

Задачею поточного дослідження було систематизувати в загальних рисах наукові публікації з афіліацією Тернопільського національного технічного університету імені Івана Пулюя на тему визначення, забезпечення та покращення якості даних.

Найбільша увага прикута саме до того, як забезпечити якість даних для «мозку» сучасних технологій – систем машинного навчання та штучного інтелекту. У ранніх, але важливих роботах [7] та [10] автори піднімають важливу проблему управління якістю та ефективності ML-систем, без чого загальна якість інформаційних систем суттєво знижується. Дослідження [7] крім теоретичних викладок пропонує також власний метод оцінювання якості даних на основі міжнародного стандарту ISO/IEC 25012, та описує власне програмне рішення для задач контролю якості даних. Робота [8] детально і покроково описує весь процес забезпечення якості для машинного навчання: від профілювання та очищення до стандартизації й усунення дублікатів.

Багато уваги приділяється також процесам підготовки даних. У публікації [6] підготовку та попередню обробку даних подано, як один з найважливіших кроків для підвищення їхньої якості перед подальшою обробкою прогностичними моделями. І це логічно, оскільки якість прогнозу буде високою, коли базові дані матимуть належну якість. У дослідженні [3] автори продовжують цю тему, аналізуючи, як саме параметри моделей впливають на якість класифікації, та розглядаючи методи очищення, нормалізації та масштабування даних.

Окремий частина знайдених робіт присвячена структурній надійності даних та їхній перевірці. На сьогодні ці характеристики даних мають критичну важливість. Наприклад, робота [5] містить огляд можливостей забезпечення цілісності даних в технологіях блокчейн у розподілених системах зберігання. Дуже практичний підхід продемонстровано в [1], де йдеться про використання Django ORM для управління базою даних, що буквально гарантує надійну цілісність даних при автоматизованому зборі наукових публікацій. Також публікація [4] описує реальну проблему: при розробці електронного кабінету абітурієнта ТНТУ виявилося, що однією з основних і найбільших проблем була саме складність отримання та верифікації персональних даних.

Окремої уваги заслуговують менш очевидні, проте не менш значущі напрями досліджень. Зокрема, у роботі [2] розглядається проблема впорядкування метаданих наукових документів: обговорюються підходи до підвищення їхньої повноти та точності шляхом автоматизованого збагачення з відкритих джерел. Дослідження [9], попри зосередженість на тестуванні програмного забезпечення, також торкається аналізу отриманих даних і систем відстеження дефектів – аспектів, що є невід'ємною складовою загального циклу забезпечення якості даних.

Таким чином, можна констатувати, що ТНТУ не лише декларує увагу до проблематики якості даних, а й володіє ґрунтовною дослідницькою базою у цій сфері. Вона охоплює як теоретико-методологічний рівень – зокрема, застосування моделі ISO/IEC 25012 – так і широкий спектр прикладних напрямів: від методів машинного навчання та технологій блокчейну до верифікації даних вступників.

Література

1. Юрчишин Д. І. Розробка інформаційної системи надання наукових сервісів з використанням бібліотеки Scrapy та ORM для взаємодії з базою даних : робота на здобуття кваліфікаційного ступеня бакалавра : спец. 121 - інженерія програмного забезпечення / наук. кер. М. Р. Петрик. Тернопіль : Тернопільський національний технічний університет імені Івана Пулюя, 2025. 59 с.
2. Сучков С. С. Система автоматичного формування блоку метаданих наукових метаданих наукових документів з використанням відкритих баз даних : робота на здобуття кваліфікаційного ступеня магістра : спец. 121 - інженерія програмного забезпечення / наук. кер. І. В. Бойко. Тернопіль : Тернопільський національний технічний університет імені Івана Пулюя, 2024. 77 с.
3. Микитишин А. Г. Застосування методів машинного навчання для класифікації даних в комп'ютеризованих системах керування / Андрій Григорович Микитишин, І. С. Дідич, Р. І. Яцишин // Тези XIII МНПК „Актуальні задачі сучасних технологій", 11-12 грудня 2024 року. — Т. : ФОП Паляниця В. А., 2024. — С. 14–16. — (Нові матеріали, міцність і довговічність елементів конструкцій).
4. Карнаухов, О. К. (2024). Дослідження розробки електронного кабінету абітурієнта ТНТУ ім. І. Пулюя. Тези доповідей V міжнародної науково-практичної конференції учених та студентів "Цифрова економіка як фактор інновацій та сталого розвитку суспільства", 46-47.
5. Гладій В. В. Технології створення розподілених комп'ютерних систем зберігання даних на основі блокчейн : кваліфікаційна робота на здобуття освітнього ступеня магістр за спеціальністю "123 — комп'ютерна інженерія" / В. В. Гладій. — Тернопіль: ТНТУ, 2023. — 82 с.
6. Кучеренко О. А. Особливості передобробки даних для методів прогнозування / О. А. Кучеренко, О. О. Кучеренко // ІМСТТ, 13-14 грудня 2023 року. — Т. : ТНТУ, 2023. — С. 72. — (Інформаційні системи та технології, кібербезпека).
7. Журихін Ю. О. Методи забезпечення якості даних при проектуванні систем машинного навчання: автореферат дипломної роботи магістра за спеціальністю 123 «Комп'ютерна інженерія»/ Ю. О. Журихін – Тернопільський національний технічний університет імені Івана Пулюя – Тернопіль, ТНТУ, 2018. – 8 с.
8. Яцишин В. Процеси забезпечення якості даних при проектуванні систем машинного навчання / В. Яцишин, Ю. Журихін // Матеріали V науково-технічної конференції „Інформаційні моделі, системи та технології", 1-2 лютого 2018 року. — Т. : ТНТУ, 2018. — С. 68. — (Секція 3. Комп'ютерні системи та мережі).
9. Чорновус, Р. М. Визначення якості тестування програмного забезпечення та аналіз отриманих даних. Матеріали конференції. Тернопіль: ТНТУ, 2017. URL: http://elartu.tntu.edu.ua/handle/123456789/18917.
10. Яцишин В. В. Оцінювання якості даних для систем машинного навчання / В. В. Яцишин, Ю. О. Журихін // Збірник тез доповідей VI Міжнародної науково-технічної конференції молодих учених та студентів „Актуальні задачі сучасних технологій", 16-17 листопада 2017 року. — Т. : ТНТУ, 2017. — Том 2. — С. 196. — (Комп'ютерно-інформаційні технології та системи зв'язку).