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 багатовимірного опрацювання великих даних розумних міст з врахуванням процесів забезпечення конфіденційності процесів

Виконав: студент VI курсу, групи СНнм-61
спеціальності 122 Комп'ютерні науки
(шифр і назва спеціальності)

_____ Вітів І.В.
(підпис) (прізвище та ініціали)

Керівник _____ Никитюк В.В.
(підпис) (прізвище та ініціали)

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

Завідувач кафедри _____ Боднарчук І.О.
(підпис) (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 » квітня 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 Вітів Іван Володимирович
(прізвище, ім'я, по батькові)
1. Тема роботи Методи конфіденційного багатовимірного опрацювання великих даних розумних міст

Керівник роботи Никитюк Вячеслав Вячеславович, к.т.н., доцент кафедри КН
(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 Наукові публікації щодо багатовимірного опрацювання даних, великих даних, розумних міст та методів забезпечення конфіденційності даних

4. Зміст роботи (перелік питань, які потрібно розробити)
Вступ. 1 Актуальність, передумови та концептуальні засади багатовимірного аналізу даних у системах «розумних міст». 2 Конфіденційність процесів багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». 3 Методи, архітектурні рішення та експериментальні дослідження багатовимірного аналізу великих даних «розумних міст». 4 Охорона праці та безпека в надзвичайних ситуаціях. Висновки. Додатки

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
1 Титульна сторінка. 2 Тема, Мета, Об'єкт, Предмет дослідження. 3 Завдання дослідження. 4 Актуальність дослідження. 5 Структура процесу багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних у медичних закладах «розумного міста». 6 Хмарна відеоаналітика зі збереженням конфіденційності для розумних міст. 7 Інноваційні застосування конфіденційних великих багатовимірних даних «розумних міст». 8 Аномінізація великих за обсягом наборів та колекцій даних в системах аналізу великих даних «розумних міст». 9 Аналіз методів багатовимірної анонімізації в аналітиці даних «розумних міст». 10 Архітектура хмарної EHR-системи. 11 Порівняльна характеристика технік анонімізації даних «розумних міст». 12 Характеристика методів опрацювання великих за обсягом наборів та колекцій даних «розумних міст». 13 Класифікація аналітичних метрик великих за обсягом наборів та колекцій даних «розумних міст». 14 Висновки. 15 Завершальний слайд.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Сенчишин В.С., доцент кафедри 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	13.04.2026	
2.	Аналіз науково-технічних публікацій та збір даних по	13.04.2026-20.04.2026	
	темі кваліфікаційної роботи		
3.	Виконання дослідження згідно мети кваліфікаційної	21.04.2026-03.05.2026	
	роботи		
4.	Оформлення розділу «Актуальність, передумови та	04.05.2026-10.05.2026	
	концептуальні засади багатовимірного аналізу даних у		
	системах «розумних міст»»		
5.	Оформлення розділу «Конфіденційність процесів	04.05.2026-10.05.2026	
	багатовимірного аналітичного опрацювання великих		
	за обсягом наборів та колекцій даних «розумних міст»		
6.	Оформлення розділу «Методи, архітектурні рішення	04.05.2026-10.05.2026	
	та експериментальні дослідження багатовимірного		
	аналізу великих даних «розумних міст»		
7.	Виконання завдання до підрозділу «Охорона праці»	27.04.2026-10.05.2026	
8.	Виконання завдання до підрозділу «Безпека в	27.04.2026-10.05.2026	
	надзвичайних ситуаціях»		
9.	Оформлення кваліфікаційної роботи	11.05.2026-13.05.2026	
10.	Нормоконтроль	14.05.2026	
11.	Перевірка на плагіат	15.05.2026	
12.	Попередній захист кваліфікаційної роботи	18.05.2026	
13.	Захист кваліфікаційної роботи	26.05.2026	

Студент _____ Вітів І.В.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Никитюк В.В.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Методи багатовимірного опрацювання великих даних розумних міст з врахуванням процесів забезпечення конфіденційності процесів // Кваліфікаційна робота освітнього ступеня «Магістр» // Вітів Іван Володимирович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра комп'ютерних наук, група СНнм-61 // Тернопіль, 2026 // С. 86, рис. – 7, табл. – 8, кресл. – 15, додат. – 1, бібліогр. – 88.

Ключові слова: багатовимірна анонімізація, багатовимірне аналітичне опрацювання, великі дані, конфіденційність даних, маскування даних, розумне місто, хмарні архітектури.

Кваліфікаційна робота присвячена дослідженню методів багатовимірного опрацювання великих даних розумних міст з врахуванням процесів забезпечення конфіденційності процесів.

В першому розділі подано актуальність досліджень в галузі аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». Розглянуто передумови багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». Висвітлено основи моделей, методів та технік багатовимірного аналітичного опрацювання великих за обсягом даних, що зберігають конфіденційність. Проаналізовано хмарні архітектури та платформи для розширеного аналітичного опрацювання, що зберігає конфіденційність. В другому розділі досліджено конфіденційність високовимірних даних «розумних міст». Розглянуто конфіденційність OLAP «розумних міст». Проаналізовано конфіденційність великих багатовимірних даних «розумних міст» у нових сценаріях застосування. Досліджено архітектури та платформи для розширеного аналітичного опрацювання збереження конфіденційності в хмарах. В третьому розділі подано порівняльний опис інструментів та підходів аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст».

ANNOTATION

Methods for Multidimensional Processing of Smart City Big Data with Consideration of Privacy Preservation Processes // The educational level "Master" qualification work // Ivan Vitiv // Ternopil Ivan Pulyuy National Technical University, Faculty of Computer Information Systems and Software Engineering, Department of Computer Science, SNnm-61 group // Ternopil, 2025 // P. 86, fig. – 7, tables – 8, posters – 15, annexes – 1, ref. – 88.

Key words: multidimensional anonymization, multidimensional analytical processing, big data, data privacy, data masking, smart city, cloud architectures.

The qualification thesis is devoted to the research of methods for multidimensional analytical processing of smart city big data, taking into account the processes of ensuring process privacy.

The first chapter presents the relevance of research in the field of analytical processing of large-scale datasets and collections of "smart city" data. The prerequisites for multidimensional analytical processing of large-scale datasets and collections of "smart city" data are considered. The fundamentals of models, methods, and techniques for privacy-preserving multidimensional analytical processing of big data are highlighted. Cloud architectures and platforms for enhanced privacy-preserving analytical processing are analyzed.

The second chapter investigates the privacy of high-dimensional "smart city" data. The privacy of "smart city" OLAP is considered. The privacy of large multidimensional "smart city" data in new application scenarios is analyzed. Architectures and platforms for enhanced privacy-preserving analytical processing in the cloud are investigated.

The third chapter provides a comparative description of tools and approaches for the analytical processing of large-scale datasets and collections of "smart city" data.

ПЕРЕЛІК СКОРОЧЕНЬ І ТЕРМІНІВ

CCG (англ. Closest Common Generalization) – найближче спільне узагальнення.

EHR (англ. Electronic Health Record) – електронний медичний запис.

GADP (англ. General Additive Data Perturbation) – загальне адитивне збурення даних.

GDPR (англ. General Data Protection Regulation)– загальний регламент про захист даних.

HIPAA (англ. Health Insurance Portability and Accountability Act) – Закон США про підзвітність і наступність медичного страхування.

HPG (англ. Homogeneous Patient Groups) – однорідні групи пацієнтів.

OLAP (англ. OnLine Analytical Processing) – аналітична обробка у режимі реального часу.

SVD (англ. Singular-Value Decomposition) – рандомізований сингулярний розклад.

ЕКГ – електрокардіограма.

ЗМІСТ

ВСТУП

Актуальність теми. Тема аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» із збереженням конфіденційності зараз набуває обертів завдяки обширному переліку сучасних сценаріїв застосування, де її можна успішно впровадити. Вони охоплюють сфери від застосунків для «розумного міста» до електронного урядування, від соціальних мереж до інструментів графового аналізу, від логістичних транспортних систем до систем бізнес-аналітики тощо. Системи охорони здоров'я, наприклад, стають одним із основних прикладних середовищ, де інструменти аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» із збереженням конфіденційності застосовуються надзвичайно широко через специфічні вимоги, диктовані контекстом, тобто захист конфіденційності особи. Методи багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» є особливим випадком методів, де основна увага приділяється використанню багатовимірних метафор, які мали успішний досвід застосування в системах онлайн-аналітичної обробки та бізнес-аналітики «розумних міст». Зроблений крок уперед, а саме багатовимірне аналітичне опрацювання великих за обсягом наборів та колекцій даних «розумних міст» із збереженням конфіденційності, становить надзвичайний інтерес для дослідницької спільноти, оскільки вона поєднує внутрішню проблему збереження конфіденційності з новими методами багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст».

Мета і задачі дослідження. Метою даної кваліфікаційної роботи ступеня «Магістр» є підвищення ефективності захисту інформаційних ресурсів «розумних міст» та процесів їх аналітичного опрацювання. Для досягнення поставленої мети потрібно виконати ряд завдань, зокрема:

– Проаналізувати стан наукових досліджень в галузі багатовимірного аналізу та хмарних платформ «розумних міст».

– Виконати обґрунтування моделей та методів захисту та процесів багатовимірної анонімізації великих даних «розумних міст».

– Здійснити аналітичне порівняння технічних методів маскування, трансформації та конфіденційного OLAP-опрацювання інформації «розумних міст».

– Дослідити архітектури та налаштування інформаційно-технологічного середовища аналізу великих даних «розумних міст».

– Проведення оцінювання метрик цілісності та конфіденційності процесів.

Об’єкт дослідження – процеси багатовимірного аналітичного опрацювання та захисту великих за обсягом наборів і колекцій даних у хмарних інформаційно-технологічних архітектурах «розумних міст».

Предмет дослідження методи, моделі, технічні засоби маскування та багатовимірної анонімізації великих даних, що забезпечують конфіденційність і цілісність аналітичних процесів у системах «розумних міст».

Наукова новизна одержаних результатів полягає у вдосконаленні методів багатовимірної анонімізації, маскування та трансформації великих даних в хмарних архітектурах, що, на відміну від існуючих підходів, дає змогу здійснювати розширене аналітичне OLAP-опрацювання даних «розумних міст» із гарантованим збереженням балансу між їхньою конфіденційністю та цілісністю.

Практичне значення одержаних результатів полягає у створенні архітектурних рішень та спеціалізованих програмно-технічних засобів маскування даних, які дають можливість розгортати в хмарних середовищах захищені системи багатовимірного аналізу для безпечного опрацювання реальних потоків інформації у муніципальних сервісах «розумних міст».

Апробація результатів магістерської роботи. Основні результати проведених досліджень обговорювались на XIII науково-технічній конференції «Інформаційні моделі, системи та технології» Тернопільського національного технічного університету імені Івана Пулюя (м. Тернопіль, 2025 р.).

Публікації. Основні результати кваліфікаційної роботи опубліковано у двох працях конференції (Див. додатки А).

Структура й обсяг кваліфікаційної роботи. Кваліфікаційна робота складається зі вступу, чотирьох розділів, висновків, списку літератури з 88

найменувань та одного додатка. Загальний обсяг кваліфікаційної роботи складає 86 сторінок, з них 55 сторінок основного тексту, який містить 7 рисунків та 8 таблиць.

1 АКТУАЛЬНІСТЬ, ПЕРЕДУМОВИ ТА КОНЦЕПТУАЛЬНІ ЗАСАДИ БАГАТОВИМІРНОГО АНАЛІЗУ ДАНИХ У СИСТЕМАХ «РОЗУМНИХ МІСТ»

1.1 Актуальність досліджень в галузі аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст»

У контексті аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст», багатовимірне опрацювання відіграє провідну роль, головним чином завдяки широкому сімейству реальних застосувань, де вона може бути успішно впроваджена. Дійсно, існує довгий список прикладних сценаріїв, що варіюються від великих соціальних даних до великих сенсорних даних, від великих графових даних до великих енергетичних даних тощо, де внутрішня природа багатовимірних абстракцій [1] забезпечує явну перевагу над класичними, наприклад, базованими на SQL інтерфейсами та методологіями завдяки безсумнівній виражальній здатності, яка не має рівних у відомому класі методологій аналітичного опрацювання. Як результат, декілька доменів даних можуть бути легко змодельовані в термінах вимірів, ієрархій, рівнів та мір.

Багатовимірне аналітичне опрацювання великих за обсягом наборів та колекцій даних «розумних міст» може зробити більше, ніж класичні підходи. Це критичне питання, яке потребує розгляду, оскільки воно являє чітку революцію для актуальних досліджень великих за обсягом наборів та колекцій даних [1]. Прикладні сценарії створюють нетривіальний дослідницький виклик, а саме збереження конфіденційності цільових наборів даних. Аналітичне опрацювання великих за обсягом наборів та колекцій даних «розумних міст» із збереженням конфіденційності [2] є гострою темою в дослідженнях наступного покоління, яка привернула увагу широкої спільноти як академічних, так і промислових дослідників. Проблема сформована питанням збереження конфіденційності критично важливої інформації, що міститься в цільових наборах даних. Чіткий приклад такого сценарію сформований у сфері охорони здоров’я [3]. Дійсно, персональна інформація, як от ім’я, прізвище, вік, адреса тощо, не може бути

розкрита в отриманих результатах аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» через, наприклад, юридичні обмеження на кшталт GDPR – чинний у європейських країнах загальний регламент про захист даних [4]. З точки зору методології це трансформується в необхідність збереження конфіденційності ідентифікуючих атрибутів через процеси аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст».

Як наслідок, можна стверджувати, що проблема відповідно поширюється на глибший випадок вирішення питань багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». Цілком інтуїтивно зрозуміло, що через високу багатовимірність та складність цільових наборів даних [5], забезпечення конфіденційності всього процесу багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» становить виклик. У цьому контексті багатовимірне аналітичне опрацювання великих за обсягом наборів та колекцій даних «розумних міст» із збереженням конфіденційності [6] є спеціальним відгалуженням основної сфери аналітичного опрацювання, де увага зосереджена на поєднанні методологій багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» та парадигм збереження конфіденційності. Цей дослідницький контекст передбачає збереження конфіденційності чутливого діапазону даних при одночасній підтримці завдань багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». Це дуже амбітна мета, яка виявляється справжнім викликом для нинішніх і майбутніх дослідницьких зусиль. Варто зазначити, що цей рух підігрівається нещодавнім вражаючим вибухом сучасних хмарних платформ та платформ обробки великих за обсягом наборів та колекцій даних [7].

Декілька традиційних літературних підходів зосередили увагу на спорідненому контексті конфіденційності OLAP [1], де проблема полягає в тому, щоб на основі багатовимірного куба даних OLAP обчислити версію цього куба із збереженням конфіденційності відповідно до заданого критерію, наприклад,

незмінність відповідей на діапазонні запити, компроміс між конфіденційністю та точністю вихідних кубів даних із збереженням конфіденційності [8]. Передбачається, що після обчислення куба даних OLAP із збереженням конфіденційності процеси багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» також будуть забезпечувати конфіденційність. Це, дійсно, один із можливих підходів, але, згідно з іншим баченням, збереження конфіденційності також має стосуватися внутрішньої моделі процесу багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст», залучаючи таким чином саму модель процесу, окрім цільових наборів даних.

У межах ширшого бачення, вищезазначена парадигма реалізує підхід, орієнтований на архітектурний каркас, де проектування конфіденційності стосується як усього процесу багатовимірного аналітичного опрацювання, так і цільового набору великих за обсягом даних «розумних міст». Слід зазначити, що останнє є інноваційним баченням порівняно з поточними сучасними пропозиціями, які, на противагу цьому, розглядають ці два поєднані питання окремо [1].

Розглядаючи актуальну наукову літературу, визнаємо, що лише декілька пропозицій можуть бути конкретно розміщені в цій обстановці. Тому існує чітка потреба в систематичному становленні дослідницької галузі багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» із збереженням конфіденційності, а також у висвітленні визначень, властивостей, переваг та обмежень.

Грунтуючись на наведених вище мотивах, проведемо огляд моделей, методів і технік, що належать до багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» із збереженням конфіденційності, а також критичний огляд переваг та обмежень сучасних пропозицій.

Зважаючи на складність і заплутаність досліджуваної наукової галузі, це дослідження запроваджує ряд спеціалізованих напрямків, кожен з яких належить до основної сфери багатовимірного аналітичного опрацювання великих за

обсягом наборів та колекцій даних «розумних міст» із збереженням конфіденційності, але з особливим акцентом на певних аспектах основного наукового та методологічного спрямування. У дослідженні [1] виділено основні спеціалізовані напрямки:

– багатовимірне аналітичне опрацювання великих за обсягом наборів та колекцій даних «розумних міст»;

– конфіденційність високовимірних даних;

– конфіденційність OLAP;

– конфіденційність великих за обсягом багатовимірних наборів та колекцій даних у нових сценаріях застосування;

– конфіденційність через анонімізацію в системах аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст»;

– конфіденційність через багатовимірну анонімізацію в системах аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст»;

– архітектури та платформи для просунутого аналітичного опрацювання із збереженням конфіденційності у хмарах.

Ці теми є головним фокусом дослідження [1]. Слід зазначити, що розглянуті окремо, ці теми є власною дослідницькою областю, проте основною метою є побудова та оцінка «єдиної» дослідницької перспективи, де вони спільно роблять внесок у виникаючий дослідницький виклик, сформований парадигмою багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» із збереженням конфіденційності.

1.2 Передумови багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст»

Багатовимірне аналітичне опрацювання великих за обсягом наборів та колекцій даних «розумних міст» [9] відноситься до нового напряму досліджень, де акцент робиться на багатовимірності аналітичної моделі, яка може

поєднуватися або не поєднуватися з багатовимірною природою цільових великих за обсягом наборів та колекцій даних.

Термін був запропонований у 2011 році [5] як новий виклик для досліджень великих за обсягом наборів та колекцій даних «розумних міст» наступного покоління, чому передували деякі роботи, де стандартне подання даних та завдання інтелектуального аналізу були вперше поєднані з концепціями багатовимірного моделювання [1]. З часом інтерес з боку дослідницької спільноти зріс, що з одного боку було зумовлено вражаючим розвитком платформ соціальних мереж та хмарних обчислень, а з іншого – застосунками класу «розумне місто» та біоінформатики.

Багатовимірне аналітичне опрацювання великих за обсягом наборів та колекцій даних «розумних міст» розширює можливості опрацювання шляхом впровадження в аналітичні моделі нових концепцій, здебільшого успадкованих від класичного досвіду OLAP та BI, завдяки аксіомам моделі, визнаним під термінами виміри та міри [1]. Для заданого домену великих за обсягом наборів та колекцій даних «розумних міст» $\mathbf{D}$ спочатку ідентифікуються виміри. Виміри – це атрибути моделі, відповідно до яких розробляється основний процес аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». Потім ідентифікуються міри. Міри – це атрибути моделі, відповідно до яких виконується основний процес аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст».

Щоб навести перший простий приклад, розглянемо добре зрозумілу сферу охорони здоров’я, наприклад, пов’язану зі спалахом COVID-19 [10]. Тут аналітики даних можуть бути зацікавлені в аналізі кількості інфікованих людей, тобто міри цільового процесу багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст», залежно від зони, часу, профілю користувача [11], історії хвороби, коморбідних факторів тощо, тобто вимірів цільового процесу багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». Усі ці концепції допомагають визначити чітку багатовимірну модель, на основі якої можна виконувати не лише класичні операції OLAP [12], як от півотинг «pivoting»,

деталізація «drill-down», згортання «roll-up» тощо, але й, що найважливіше, застосовувати алгоритми інтелектуального аналізу даних, як от пошук частих шаблонів, та методи машинного навчання, як от регресія, збільшуючи таким чином виражальну потужність базового процесу [1].

На рисунку 1.1 подано еталонну архітектуру для тематичного дослідження, пов’язаного з актуальною ситуацією в охороні здоров’я.

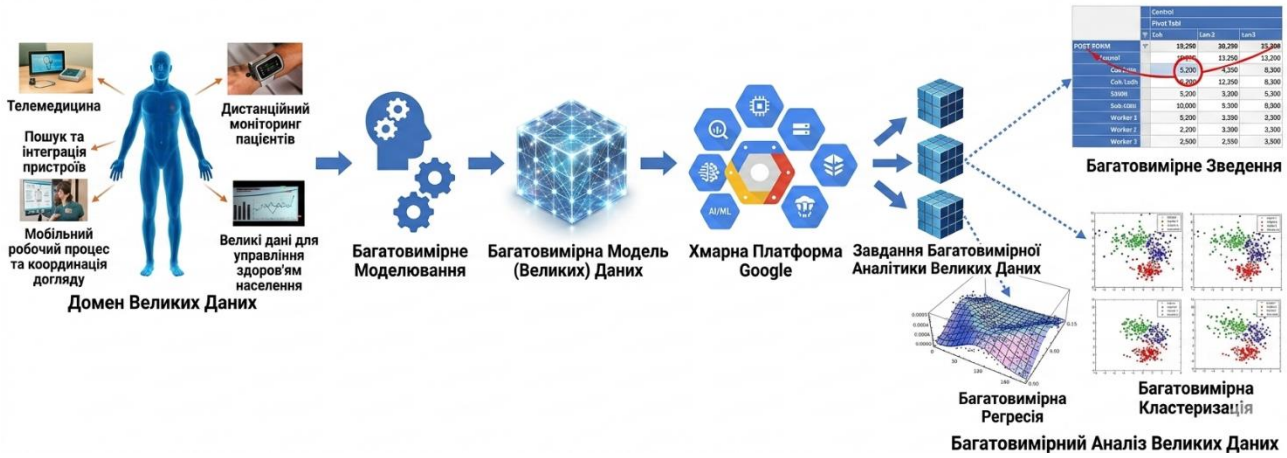

Рисунок 1.1 – Структура процесу багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних у медичних закладах [1]

Як показано на рис. 1.1, у еталонній архітектурі можна виділити декілька рівнів:

1. Домен великих даних.
2. Багатовимірне моделювання.
3. Багатовимірна модель великих за обсягом наборів та колекцій даних «розумних міст».
4. Google Cloud Platform.
5. Завдання багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст».
6. Багатовимірний аналіз великих за обсягом наборів та колекцій даних «розумних міст».

1.3 Основи моделей, методів та технік багатовимірного аналітичного опрацювання великих за обсягом даних, що зберігають конфіденційність

Багатовимірне аналітичне опрацювання великих за обсягом наборів та колекцій даних «розумних міст» – це знання, що забезпечують прийняття рішень і надаються шляхом багатовимірного аналізу великих за обсягом наборів та колекцій даних «розумних міст» над багатовимірними моделями даних, визначеними в багатовимірних системах зберігання великих за обсягом наборів та колекцій даних «розумних міст», зокрема, ті, що використовують можливості сховищ даних, наприклад, OLAP-сервери [1]. Багатовимірний аналіз великих за обсягом наборів та колекцій даних «розумних міст» – це завдання, що передбачає індивідуального та орієнтованого вивчення всього масиву даних, збережених у відповідній багатовимірній моделі, як от OLAP. Зазвичай багатовимірний аналіз великих за обсягом наборів та колекцій даних «розумних міст» стає можливим завдяки націлюванню на конкретні виміри даних та їх перевірці, наприклад, з використанням операції агрегації, в точці їхнього перетину для отримання корисних знань для конкретного застосування. Приклади систем, що забезпечують багатовимірне аналітичне опрацювання великих за обсягом наборів та колекцій даних «розумних міст», подані в [13].

Конфіденційність високовимірних даних є відкритим викликом. Дійсно, багато методів конфіденційності даних виявляються недостатніми при роботі з великою кількістю атрибутів під час захисту «високовимірних» даних. Було встановлено, що більшість існуючих методів конфіденційності не можуть належним чином та ефективно обробляти високовимірні дані [14]. З іншого боку, застосування існуючих методів конфіденційності до високовимірних даних пов’язане з погіршенням якості отриманих даних, що використовуються для аналізу даних [15]. У [16] було зазначено, що такі методи, як k-анонімність, залежать від просторової локальності, вимірюваної за допомогою метрики відстані, точок даних для виконання анонімізації даних, проте просторова локальність спотворюється у випадку високовимірних просторів [17]. Рішення на основі вибору ознак не є дуже корисним, оскільки воно призводить до

відкидання низки ознак – процесу, який може негативно вплинути на корисність даних. Проте вартим уваги перспективним напрямком є той, що описаний у [18], де пропонується використовувати підхід вертикальної фрагментації для поділу даних відповідно до їхніх атрибутів та незалежної анонімізації кожної частини.

OLAP – це метод підтримки прийняття рішень, який дає змогу видобувати цінні знання шляхом дослідження багатовимірної структури даних. Проте в багатьох випадках таке дослідження має виконуватися в конфіденційний спосіб. Дійсно, наприклад, чутлива інформація може бути розкрита для обслуговування зловмисних намірів, і тому потреба в конфіденційності є очевидною. Конфіденційність OLAP може бути реалізована декількома способами: наприклад, автори [19] описують підхід контролю доступу до кубів OLAP, який обмежує доступ для ненадійних користувачів шляхом приховування зрізів даних перед наданням доступу до куба. Приховування здійснюється двома шляхами: шляхом представлень або шляхом правил. Шлях представлень обмежує відображення певних чутливих вимірів, тоді як шлях правил обмежує доступ до чутливих даних для певних користувачів, а не для інших.

Інший приклад [16] описує метод пертурбації клієнтських даних, що запитуються через OLAP за допомогою схем заміни при зберіганні, які обирають цільові дані для рандомізації, перед їхньою фактичною інтеграцією на сервер. Потім дані реконструюються за допомогою техніки інверсії матриці або ітераційної байєсівської техніки перед застосуванням агрегацій. Ще один приклад [20] пропонує підхід на основі методу вибірки для захисту OLAP кубів.

Нові технології диктують суворіші, ніж будь-коли, умови для процесів, методів та середовищ багатовимірних великих за обсягом наборів та колекцій даних «розумних міст». Зокрема, методи забезпечення конфіденційності для великих багатовимірних даних мають еволюціонувати, стаючи дієвішими, ефективнішими, гнучкішими та масштабованішими [1]. Дійсно, з огляду на нові способи зберігання, обробки та аналізу даних, необхідно визначити комплексніші методи конфіденційності, щоб упоратися з технологічними проривами, які кидають виклик приватності.

У випадку «розумної» медицини, сьогодні електронні медичні карти (EHR) вважаються новим еталонним джерелом даних, що дає можливість, наприклад, значно покращити процеси в галузі охорони здоров'я серед багатьох інших переваг [21]. У [22] обмін електронними медичними картами захищений за допомогою інноваційної системи, що використовує динамічний контроль доступу на основі блокчейну в поєднанні з локальною диференційною конфіденційністю. Інший приклад пояснює, як можна розробити структуру даних мобільності із вбудованою конфіденційністю «за проектом» [23].

Ще один приклад описує метод анонімізації даних «розумних» лічильників [24]. Підхід полягає в розрізненні високочастотних даних, що використовуються для оцінки регулярного енергоспоживання користувача, та низькочастотних «розумних» даних, що використовуються для щомісячного виставлення рахунків користувачам, та виконанні анонімізації виключно ідентифікатора користувача «user ID», пов'язаного з їхніми генерованими високочастотними даними вимірювань, щоб забезпечити безпечний обмін даними «розумного» обліку між компонентами установок «розумного» обліку. Анонімізація такого ідентифікатора користувача базується на сервісі ескроу – депонування [1].

Аналітичне опрацювання великих за обсягом наборів та колекцій даних «розумних міст» є корисним для прийняття рішень [1]. Оскільки це стає обов'язковою вимогою, посилення відповідальності систем аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» за конфіденційність тягне за собою впровадження модуля приватності, який належним чином дбає про приховування чутливих даних під час застосування аналітичних процедур. У цьому сенсі автор [25] описує метод анонімізації в контексті системи натільних сенсорних мереж «Body Area Network», де носні сенсорні пристрої відповідають за збір даних ЕКГ.

Метод полягає в додаванні шуму до характеристик набору даних ЕКГ на основі їхнього порядку важливості, визначеного на етапі попередньої обробки даних, для досягнення чітко визначеного порогу втручання диференційної конфіденційності. Поріг втручання оновлюється в режимі реального часу в міру вимірювання нових даних сенсорними пристроями та враховує важливість ознак.

Поріг втручання відповідає за гарантування певного рівня конфіденційності даних. Інший приклад такої системи описаний у [26], де запропонована система використовує гомоморфну криптосистему Пейє «Paillier Homomorphic Cryptosystem» для запобігання витоку інформації на рівні мережевих каналів у контексті системи федеративного навчання.

Багатовимірне обмеження даних є вирішальним аспектом, який слід враховувати в контексті системи аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» із збереженням конфіденційності. Аспект багатовимірності даних додає ще одну перешкоду для збереження приватності, незалежну від конкретного застосунка, у сфері аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». У спробі уникнути згаданих проблем, один із прикладів, запропонований у [27], використовує варіант алгоритму шифрування BGN для забезпечення багатовимірної агрегації зашифрованих даних, необхідної для збору даних із джерел периферійної мережі «edge network» до центру управління, на основі якого аналітичне опрацювання може бути виконано конфіденційно.

1.4 Хмарні архітектури та платформи для розширеного аналітичного опрацювання, що зберігає конфіденційність

Хмарні сервіси потрібні більше, ніж будь-коли, для обширного спектру інтенсивних процесів, пов’язаних із даними, як от аналітика великих за обсягом наборів та колекцій даних «розумних міст». Надзвичайно великий обсяг даних, що генерується за допомогою нових технологій, диктує потребу в більшій кількості ресурсів для виконання завдань, пов’язаних із даними. Зокрема, аналітичне опрацювання даних із збереженням конфіденційності вимагає значних обчислювальних ресурсів і ресурсів зберігання для адекватної роботи в умовах жорсткіших обмежень, пов’язаних з обсягом даних. Приклад рішення вищезгаданих проблем описаний у [28]. Дане рішення являє собою хмарну відеоаналітику як сервіс, що захищає відеопотоки від атак по сторонніх каналах, а також від прямого витоку приватного відеоконтенту (див. рисунок 1.2).

Рисунок 1.2 – Хмарна відеоаналітика зі збереженням конфіденційності для розумних міст

Інший приклад [29] описує рішення, яке використовує методи диференційної конфіденційності для забезпечення функцій спільного використання даних та аналітичне опрацювання хмарних сервісів. Автори [30] стверджують, що найкращими для використання методами анонімізації в хмарному контексті є підходи до деідентифікації:

– псевдонімізація;
– узагальнення;
– придушення;
– маскування даних.

Вони виключають придатність інших підходів, що забезпечують анонімізацію, як от гомоморфне шифрування або також часткове шифрування в поєднанні з довіреними клієнтськими чи гібридними хмарами в хмарному контексті з кількох зазначених причин [1].

1.5 Висновок до першого розділу

В першому розділі кваліфікаційної роботи освітнього рівня «Магістр» подано актуальність досліджень в галузі аналітичного опрацювання великих за

обсягом наборів та колекцій даних «розумних міст». Розглянуто передумови багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». Висвітлено основи моделей, методів та технік багатовимірного аналітичного опрацювання великих за обсягом даних, що зберігають конфіденційність. Проаналізовано хмарні архітектури та платформи для розширеного аналітичного опрацювання, що зберігає конфіденційність.

2 КОНФІДЕНЦІЙНІСТЬ ПРОЦЕСІВ БАГАТОВИМІРНОГО АНАЛІТИЧНОГО ОПРАЦЮВАННЯ ВЕЛИКИХ ЗА ОБСЯГОМ НАБОРІВ ТА КОЛЕКЦІЙ ДАНИХ «РОЗУМНИХ МІСТ»

У цьому розділі проаналізуємо множину найбільш релевантних наукових публікацій щодо конфіденційності процесів багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст».

Автор [31] пропонує фреймворк для аналізу багатовимірних даних шляхом застосування алгоритмів кластеризації та класифікації до багатовимірного подання вхідних клінічних даних. Підхід зводиться до побудови кривих росту шляхом аналізу даних за допомогою OLAP з використанням багатовимірних моделей фактів – «Dimensional Fact Models», які слугували б еталонами для порівняння росту плодів і, таким чином, виявлення аномалій. Основне припущення полягає в тому, що здорові плоди мають схожі базові характеристики росту, якщо вони мають однаковий гестаційний вік, схожі генетичні властивості, наприклад, етнічна приналежність, і зростали в подібних умовах навколишнього середовища [1]. Базові криві росту плодів полегшують визначення груп плодів, що називаються однорідними групами пацієнтів «HPG». Приналежність плодів до однієї з груп «HPG» надаватиме інформацію про їхнє здоров’я, благополуччя або патології. Багатовимірні подання створюються на основі цільових даних після застосування алгоритму класифікації до даних для пошуку найближчої групи «HPG» для конкретного пацієнта.

У [32] автори пропонують інструмент для аналізу багатовимірних медичних даних, пов’язаних із біологічним лікуванням, та моніторингу їхнього введення пацієнтам. Вони пропонують інструмент на основі мікросервісів, де різні компоненти розгортаються за допомогою контейнерної віртуалізації, наприклад, Docker, для аналізу даних на базі ROLAP, що зберігаються в базі даних PostgreSQL. Вони пропонують багатовимірну модель фактів для вітрини даних та аналізу введення біологічних препаратів пацієнтам.

2.1 Конфіденційність високовимірних даних «розумних міст»

У цьому параграфі розглянемо деякі з найбільш релевантних суміжних робіт у контексті конфіденційності високовимірних даних.

Автори [33] розглядають техніку для фіксації кореляції даних під час їхньої анонімізації, враховуючи максимізацію корисності даних, що ілюструє, як генерується кореляційна матриця, наступним чином: виходячи з «транзакційних даних», сформованих деякими клієнтами, будується загальна матриця шляхом проставлення «1» – якщо товар є у транзакції, та «0» – якщо ні. Чутливі атрибути навмисно пов’язуються з чутливими елементами, щоб полегшити обробку. Спочатку вони розпочинають із перетворення загальної матриці даних у стрічкову матрицю за допомогою алгоритму [34].

Далі вони використовують її для створення формувань чутливих транзакцій за допомогою запропонованого ними жадібного алгоритму «CAHD». «CAHD» працює таким чином: ітерація над чутливою транзакцією полягає в групуванні p сусідніх, вхідний ступінь конфіденційності p, чутливих транзакцій разом на основі поняття неконфліктних атрибутів. Дві конфліктні чутливі транзакції мають спільне чутливе значення та шляхом вибору разом лише тих транзакцій, які мають найближчу кількість спільних квазі-ідентифікаторів «Quasi-ID» [1]. Під час процесу формування груп ведеться гістограма для гарантування того, що обмеження конфіденційності дотримано. Алгоритм зупиняється, коли більше не залишається незгрупованих чутливих транзакцій або коли неможливо сформувати додаткові групи.

У [35] автори описують розширення l-різноманітності «l-diversity» – MSA-різноманітність «MSA-diversity», що є алгоритмом, здатним захистити від того, що вони називають атаками з використанням знань про неприналежність: ймовірність вирахувати запис жертви обмежена величиною $1/(l-i)$ за наявності i біт знань про неприналежність. Знання про неприналежність: зловмисник знає i значень чутливих атрибутів про клас еквівалентності жертви, що є інформацією, яка звужує кількість можливих записів, пов’язаних із жертвою, шляхом скорочення кількості потенційних записів жертви. Алгоритм прямує до

зазначеної мети шляхом, по-перше, лінеаризації значень багатовимірних атрибутів квазі-ідентифікаторів (кортежів) у впорядкований одновимірний список, а потім намагається побудувати SA-відмінні, через визначену MSA-різноманітність, групи з цих кортежів шляхом створення пов’язаних матриць, де жодні два кортежі не мають спільного рядка чи стовпця. Потім групи узагальнюються або анатомізуються. Конфіденційність високовимірних даних «розумних міст» подано на рисунку 2.1.

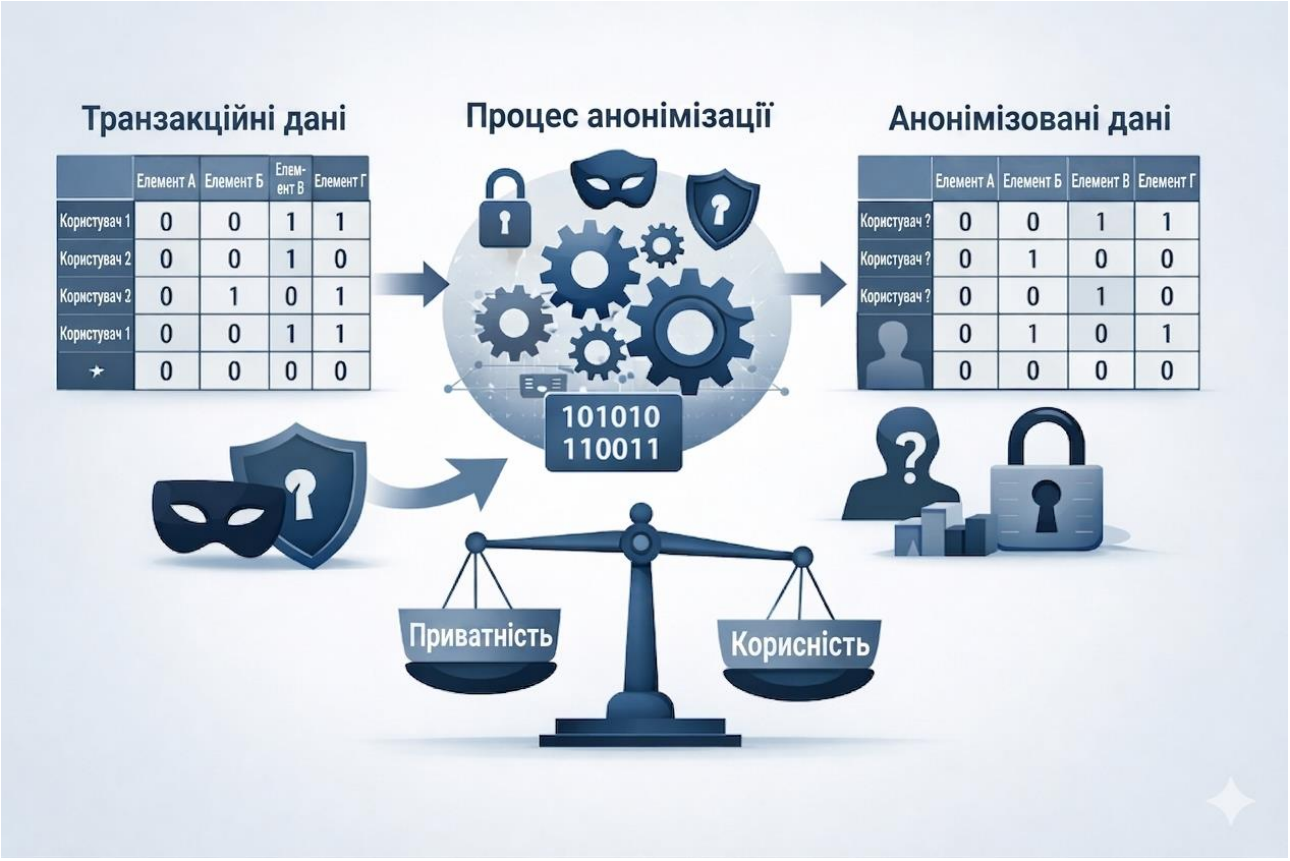

Рисунок 2.1 – Конфіденційність високовимірних даних«розумних міст»

Автор [36] досліджує техніку багатовимірної анонімізації, де трансформації узагальнення застосовуються до цільового набору даних ітеративно. Трансформації встановлюються у вершинах решітки, і алгоритм проходить крізь решітку, враховуючи максимізацію корисності даних та певний визначений поріг ризику для даних. Трансформація підтверджується та використовується лише в тому випадку, якщо вона підвищує корисність даних і якщо дотримано порогів ризику для даних. Для досягнення своєї мети алгоритм перемикається між підходом жадібного обходу та пошуком за критерієм

«перший ліпший» «Best-first search». Якщо досягнуто межі решітки, алгоритм виконує повернення назад «backtrack», щоб мати змогу продовжити обхід. Якщо було зроблено n кроків, алгоритм також перейде до використання жадібного алгоритму, але тут уже без виконання будь-якого повернення назад [1]. Під час обходу решітки алгоритм підтримує чергу, де трансформації впорядковані у порядку спадання відповідно до корисності даних. Інтуїтивно зрозуміло, що чим більше трансформація зберігає корисність даних, тим меншим є пов’язане з нею узагальнення даних.

2.2 Конфіденційність OLAP «розумних міст»

У цьому параграфі дослідимо концепцію конфіденційного OLAP «розумних міст», зосереджуючись на дослідженні множини релевантних сучасних наукових робіт (див. рисунок 2.2).

Рисунок 2.2 – Конфіденційність OLAP «розумних міст»

Автори [37] пропонують триетапний метод вибірки «sampling» для анонімізованого аналізу кубів даних, а саме: розподіл простору зберігання для зразків, витягнутих із робочого навантаження запитів «QWL», для отримання

приблизної відповіді куба даних; власне етап вибірки робочого навантаження запитів даних; та фаза уточнення, включена для покращення результатів з точки зору точності відповіді та конфіденційності даних, забезпечуючи таким чином збалансований компроміс між двома цільовими метриками. Автори також визначають методи для вирішення численних проблемних аспектів, що перешкоджають точній та конфіденційній відповіді на запити до куба даних [1]. Наприклад, вони пропонують розглядати запит над набором даних як набір міні-запитів, визначених над набором «регіонів», які є обмежувальними просторами згаданих запитів. Виходячи з цього, вони розглядають запит як набір регіонів, які повинні відповідати декільком умовам для забезпечення детального та ефективного аналізу, зберігаючи при цьому запитувані дані в безпеці від зловмисних цілей.

У [38] автори пропонують нову техніку анонімізації кубів даних у розподіленому середовищі. Вони використовують матричну декомпозицію «CUR» – метод розкладання матриць, що використовується для обчислення наближених представлень великих матриць, – для анонімізації 2D-подання розподілених кубів. У розподіленому сценарії вони пропонують протокол безпечної розподіленої агрегації «OLAP» – «SDO» для анонімізації всієї мережі кубів даних. Щоб підкреслити процес анонімізації, досягнутий завдяки наближенню CUR, вони проводять поглиблену оцінку результатів реконструкції, відтворюючи метод, вперше запропонований в алгоритмі пертурбації із заміною при зберіганні «Retention Replacement Perturbation». Ідея полягає в тому, щоб достатньо точно кількісно оцінити значення, шляхом встановлення «розумних» і суворих меж, отриманих відповідей на запити діапазонного підсумовування «range-SUM» на реконструйованих даних. Також проводиться порівняння результатів згаданих запитів як над сконструйованою версією, так і над початковою, щоб підтвердити дієвість та ефективність використаного методу декомпозиції CUR [1]. Вони також демонструють деякі корисні математичні властивості, що включають те, що вони називають повною диференційною конфіденційністю – кількісна оцінка різниці між початковим і трансформованим поданням, та маргінальною диференційною конфіденційністю – кількісна оцінка

часткової різниці між початковим і трансформованим поданням у домені запитуваного атрибута.

Автори [39] розкривають вплив на корисність даних та втрату інформації в анонімізованих кубах даних трьох типів підходів до анонімізації, а саме:

– глобального узагальнення;

– локального узагальнення;

– бакетизації «bucketization».

Вони вказують на те, як згадані підходи до анонімізації можуть змінити ефективність аналізу куба даних стосовно трьох врахованих метрик:

– розміру куба даних;

– перекриття комірок;

– втрати інформації.

Вони стверджують, що на ці метрики неминуче впливає зміна даних, що є результатом узагальнення або бакетизації. Крім того, вони висвітлюють, як розглянуті методи анонімізації формують кінцевий анонімізований куб даних.

Наприклад, бакетизація має тенденцію до збільшення кількості комірок, оскільки бакети (чутливі значення) приписуються всім пов’язаним записам квазі-ідентифікаторів в анонімізованому кубі даних [1]. Подібним чином, локальне узагальнення може призвести до збільшення кількості комірок у кінцевому анонімізованому кубі даних, тоді як глобальне узагальнення призводить до меншої кількості комірок. Крім того, вони наводять приклад того, як локальне узагальнення призводить до більшої кількості перекриттів комірок, і обґрунтовують це тим фактом, що метод дає змогу використовувати декілька правил запису.

Щодо метрики втрати інформації, вони стверджують, що всі розглянуті методи анонімізації призводять до більш-менш значної втрати інформації. Крім того, вони пропонують четверту метрику, що називається медіанна відносна похибка. Підхід, який вони використовують для акцентування на згаданій метриці, полягає, по-перше, у запиті кубоїдів «cuboids» на основі вибраних значень атрибутів або діапазонів значень атрибутів, і, по-друге, в обчисленні відносної похибки шляхом виконання кожного типу запиту (точкового та

діапазонного) на різних кубоїдах та розрахунку показника, що складається з різниці між результатом запиту до анонімізованого кубоїда та оригінального набору даних [1], поділеної на результат, отриманий з оригінального куба даних.

2.3 Конфіденційність великих багатовимірних даних «розумних міст» у нових сценаріях застосування

У цьому параграфі розглянемо деякі суміжні наукові роботи щодо конфіденційності великих багатовимірних даних «розумних міст» у нових сценаріях застосування. Розглянемо різні дослідження та методології, які вирішують питання приватності в контекстахк IoT, «розумні міста», медичні дані [40]тощо.

У [41] цілісність даних та конфіденційність користувачів є основними аспектами збереження приватності. Для досягнення своєї мети автори пропонують техніку, де дані спочатку упаковуються за допомогою китайської теореми про залишки, а по-друге, шифруються за допомогою методу шифрування, який не покладається на встановлений механізм шифрування з відкритим ключем, наприклад на основі методу Пейє, а скоріше на підхід, заснований на повідомленні користувача з використанням техніки переговорів про ключі між користувачами. Згадана техніка покладається на додавання міток часу до даних кожного користувача для досягнення своєї мети. За словами авторів, техніка шифрування гарантує стійкість проти мережевої атаки повторного відтворення. Фази запропонованої схеми:

– звітування про дані «SM»;
– анонімізована агрегація даних «GW»;
– фаза парсингу даних «CC».

Розроблена система спрямована на збір звітів про споживання електроенергії з житлових районів, що включають багатьох мешканців «SM», для їх шифрування та подальшої передачі через шлюзи «GW» в агрегованому вигляді [1]. Потім їх аналіз проводиться через центри управління «CC». З огляду на таку структуру, само собою зрозуміло, що передача даних має бути безпечною

заради блага користувачів та інстанції центру управління. Сценарій використання цього підходу пов’язаний із «розумними» мережами «smart grids», які мають наповнюватися результатами аналізу даних для вдосконалення стратегій постачання електроенергії мешканцям.

Автори [42] пропонують інноваційний алгоритм гомоморфного кодування на основі китайської теореми про залишки та шифрування Пейє. Шифрування виконується після етапу збору даних IoT пристроями і перед їх агрегацією на периферійних вузлах «edge nodes» та відправленням до центру управління для безпечного та надійного аналізу. Модель системи, яку вони пропонують, спрямована на надання інтелектуальних транспортних послуг шляхом аналізу даних пристроїв. Детальніше: периферійні вузли отримуватимуть шифротекст від IoT пристроїв, виконуватимуть над ними агрегації (гомоморфні операції) та надсилатимуть агрегований результат для аналізу до центру управління після етапу дешифрування. Китайська теорема про залишки була використана для перетворення вхідних векторів повідомлень у цілочисельні вектори, а метод Пейє використовувався безпосередньо для шифрування цих векторів у шифротекст. Агрегації поверх зашифрованих даних можливі, оскільки гомоморфне кодування даних це дозволяє.

Дослідники [43] розглядають збереження конфіденційності вихідних даних, що зберігаються в медичному центрі, а потім переносяться в хмару для полегшення обробки, та запиту, наприклад, запиту на суму в діапазоні – «range SUM query», надісланого в хмару користувачами-лікарями. Вони використовують R-дерево для індексації набору даних та полегшення виконання запитів до даних через діапазонні SUM-запити, визначивши для цієї мети спеціальний алгоритм. Приклади інноваційних застосувань конфіденційних великих багатовимірних даних «розумних міст» подано в таблиці 2.1.

Таблиця 2.1 – Приклади інноваційних застосувань конфіденційних великих багатовимірних даних «розумних міст»

Сфера застосування	**Методологічний підхід**	**Технічні рішення**	**Очікуваний ефект**
Енергетичні мережі «smart grids»	Триетапна модель збору та шифрування даних	Китайська теорема про залишки, часові мітки, безключове шифрування	Безпечна передача даних споживання енергії між мешканцями та центром
Інтелектуальний транспорт і IoT-інфраструктура	Гомоморфне шифрування для периферійних вузлів	Метод Пейє, CUR-декомпозиція, агрегація шифрованих даних	Конфіденційний аналіз трафіку та мобільності в місті
Медичні сервіси міських лікарень	Анонімізований пошук у хмарних сховищах	-дерево, гомоморфне порівняння, матричні обчислення	Захист персональних медичних даних у міських хмарах
Туманні обчислення для міських сенсорних систем	Трирівнева архітектура SS–FN–DT	Збір, змішування, агрегація, звітування	Безпечна аналітика даних сенсорів у міських середовищах

Основна ідея зводиться до пошуку даних, що перетинаються, між цільовою частиною даних, встановленою у вузлі, що обходиться, та інтервалами запиту. Обхід R-дерева здійснюється методом пошуку вглиб. Зокрема, пошук у дереві зверху вниз триває лише в тому випадку, якщо запит перетинається з даними поточного вузла, і зрештою листки додаються до відповіді на запит, якщо вони задовольняють зазначену умову. Автори уподібнюють межі даних кожного вузла та запиту до цілих чисел і оперують їхньою версією у вигляді бінарних векторів для їх порівняння, надаючи таким чином анонімізовану версію згаданого алгоритму через порівняння меж даних та запиту. Порівняння надаватиме інформацію про перетин запиту з даними і, таким чином, про набір відповідей на запит. Порівняння цілих чисел трансформується в тест на рівність на основі визначеної ними функції, яка варіюється відповідно до векторних змінних, встановлених з векторної версії розглянутих цілих чисел через спряження та

транспонування вихідних цілочисельних векторів [1]. Це порівняння має на меті приховати результати проміжних операцій (оцінку нерівностей), необхідних для отримання результуючого перетину між запитом і даними. Зокрема, вони застосовують техніку порівняння даних на основі гомоморфного кодування, включно з шифруванням векторів даних і запиту та матричні обчислення над закодованими версіями векторів, яка видає результат тесту на рівність, що інформує про перетин між вихідними векторами.

У [44] пропонується рішення для агрегації даних із збереженням конфіденційності, реалізоване в межах трирівневої туманної системи «fog-based system», що має на меті аналіз даних за допомогою визначених користувачем запитів при забезпеченні безпеки та захищеності даних. Автори пропонують архітектуру, сформовану із:

– сервера послуг «Service Server» (SS);
– туманних вузлів «Fog Nodes» (FN);
– терміналів даних «Data Terminals» (DT).

Вузли FN підключаються до SS, тоді як DT підключаються до вузлів FN, кожен DT збирає дані, наприклад, із давачів за допомогою Bluetooth. Пропонуються декілька кроків для виконання запитів до даних, а саме:

– ініціалізація системи;
– збір даних;
– змішування даних;
– агрегація даних;
– звітування про результати.

2.4 Конфіденційність через анонімізацію в системах аналізу великих даних «розумних міст»

У цьому параграфі розглянемо релевантні суміжні роботи щодо конфіденційності за допомогою анонімізації в системах аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст», зосереджуючись на різних методах анонімізації, як от k-анонімність, l-

різноманітність та t-близькість, а також на їх застосуванні для збереження чутливої інформації у великих наборах даних.

У [45] автори виходять із припущення щодо обмеження порушення конфіденційності, яке полягає в тому, що більшість актів викрадення записів даних стають можливими завдяки знанню значень атрибутів, що відповідають лише частині всього набору «QID» – повного набору атрибутів квазі-ідентифікаторів набору даних. На основі цього вони пропонують орієнтований на корисність метод конфіденційності під назвою LKC-privacy. А саме, цей метод передбачає, що для кожного кортежу атрибутів QID розміром менше L, включеного в повний набір «QID», принаймні K записів повинні мати однакові значення атрибутів, гарантуючи K-анонімність.

Крім того, ймовірність виведення чутливого значення в межах класів еквівалентності не повинна перевищувати C%, гарантуючи C-десоціацію. Вони пропонують два типи сценаріїв, де виконується анонімізація даних про операції та переливання крові: централізований підхід, де Центральне державне агентство охорони здоров'я «CGHA» анонімізує дані у спосіб LKC-privacy для передачі одержувачу, і розподілений підхід, де дані анонімізуються окремо в кожній лікарні, що здійснює переливання крові пацієнтам, уникаючи таким чином потреби проходити через «CGHA».

Автори [46] пропонують підхід на основі k-анонімізації для медичних показників IoT, зібраних із вбудованих пристроїв пацієнтів, зокрема, медичних сенсорів. Розглянута архітектура системи передбачає, що дані зібрані з пристроїв пацієнтів мають надсилатися через WAN, інтернет або мобільну мережу для аналізу медичними службами та, зрештою, слугуючи базою для системи, що дає можливість рекомендувати лікування пацієнтам. Автори уподібнюють архітектуру системи до клієнт-серверної, між якими передаються дані. Завдання анонімізації відбувається всередині мережі, де зловмисник-перехоплювач може втрутитися в передачу даних на ненадійний сервер. Автори пропонують анонімізувати дані за допомогою підходу на основі кластеризації з використанням техніки $alpha$-десоціації на вузлах агрегації даних у мережі [1].

Спочатку кожен кортеж розглядається як кластер, після чого обирається репрезентативний вектор для порівняння відстаней між кластерами та створення матриці відмінностей. Кластери групуються відповідно до подібності між кортежами. Зокрема, відмінності кластерів використовуються для оцінки відстані між парами кластерів.

Автори [47] визначають алгоритм анонімізації шляхом кластеризації на основі MapReduce. Алгоритм використовує алгоритм кластеризації k-means для отримання декількох кластерів із вихідного набору даних. Кількість є вхідним параметром, пов’язаним з анонімізацією у паралельний спосіб MapReduce. Алгоритм виконується у два раунди операцій MapReduce. У першому раунді MapReduce мапер «mapper» виконує роль сортування кластерів на основі кількості значень їхніх атрибутів для формування кортежу «ключ-значення», де ключем є значення атрибута, а значенням – кількість значень атрибута, пов’язана з чутливим атрибутом. Редюсер «reducer» після цього відповідає за виведення кількості значень чутливих атрибутів для кожного кластера.

Аномінізація великих за обсягом наборів та колекцій даних в системах аналізу великих даних «розумних міст» зображена на рисунку 2.3.

Рисунок 2.3 – Аномінізація великих за обсягом наборів та колекцій даних в системах аналізу великих даних «розумних міст»

У [48] автори пропонують анонімізувати як числові, так і категоріальні атрибути за допомогою перетворень на основі атрибутів. Перетворення для числових атрибутів передбачає обчислення для кожного квазі-ідентифікатора (числового атрибута) ширини інтервалу «IW», що складається із середнього значення найбільшого інтервалу значень розглянутого атрибута. Після чого алгоритм будує інтервали класів еквівалентності, які інкрементуються відповідно до «IW» від найменших до найбільших значень атрибутів. Після побудови діапазонів для класів еквівалентності «EQ», початкові значення кортежів враховуються для розрахунку середнього значення, що призводить до анонімізуючих значень, які замінять фактичний діапазон значень кожного EQ. У випадку, коли значення атрибутів у межах одного побудованого EQ всі рівні, алгоритм переходить до зміни значення, щоб приховати вихідне значення атрибута. У випадку категоріальних атрибутів вони пропонують підхід анонімізації шляхом бакетизації на основі підметоду кластеризації, який використовує підхід k-medoid для створення кластерів із набору даних для анонімізації та розділення кластерів, уподібнених до бакетів, для генерації l-різноманітних чутливих бакетів [1]. Перевірка значень чутливих атрибутів на l-різноманітність, генерація чутливих бакетів, базується на розрахунку розподілу ймовірностей чутливих значень у межах кожного нарізаного бакета, який повинен відповідати порогу l-пропорційності.

2.5 Конфіденційність через багатовимірну анонімізацію в системах аналітичного опрацювання великих даних

У цьому параграфі дослідимо концепцію забезпечення конфіденційності за допомогою багатовимірної анонімізації в системах аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст», зосереджуючись на розгляді деяких релевантних сучасних наукових робіт.

Автори [49] пропонують жадібні алгоритми анонімізації, які максимізують корисність даних і мінімізують запропоновану метрику під назвою зважений нормалізований штраф за визначеність «Normalized Certainty Penalty», «NCP».

Пропонуються два підходи: перший – це висхідний підхід «bottom-up», який починається з призначення кожного з кортежів набору даних певній групі, а потім намагається об'єднати кожну з них, коли умова k-анонімності не виконується, з іншою групою та/або кортежами, шляхом проходу сканування набору даних, що мінімізувало б метрику «NCP». Якщо отримана об'єднана група перевищує певний поріг, її розділяють на менші групи, які задовольняють обмеження анонімізації щодо необхідного розміру групи k. Для обчислення метрики «NCP» між групами та/або групами кортежів використовувалися відстані для числових значень атрибутів і ієрархії для категоріальних значень атрибутів.

Другий підхід – це низхідний підхід «top-down», де вихідний набір даних розбивається на послідовні підмножини з урахуванням мінімізації метрики «NCP». Дійсно, розділи повинні мати можливість бути далі розбитими на менші групи кортежів, які мінімізують метрику «NCP» [1]. На кожній ітерації розділи повинні підтверджувати обмеження k-анонімності. Зрештою, кожна група має задовольняти k-анонімність, інакше виконується коригування.

Автори [50] пропонують підхід на основі ієрархічної таксономії атрибутів для анонімізації значень атрибутів вхідного набору даних шляхом локального перекодування. При цьому пропонується нова метрика, а саме зважена ієрархічна відстань «Weighted Hierarchical Distance» («WHD»), для оцінки результатів узагальнення початкового набору даних. Метрика вимірює відношення суми ваг кожного з рівнів домену між початковим значенням і узагальненим значенням до суми ваг усіх рівнів ієрархічного домену [1]. Спотворення між двома кортежами, початковим та узагальненим, потім виводиться шляхом підсумовування всіх відстаней WHD між рівнями ієрархічного домену кожного зі значень атрибутів початкового та узагальненого кортежів. Спотворення між таблицями тоді є сумою спотворень кожного кортежу таблиці. Також визначено відстані між двома кортежами, а також між класами еквівалентності. В таблиці 2.2 подано аналіз методів багатовимірної анонімізації в аналітиці даних «розумних міст».

Таблиця 2.2 – Аналіз методів багатовимірної анонімізації в аналітиці даних «розумних міст»

Алгоритм / Підхід	**Метрики**	**Застосування в розумних містах**	**Оптимізація великих даних**
Жадібна k-анонімізація *(bottom-up / top-down)*	NCP (штраф за визначеність)	Моніторинг ЖКГ та енергоспоживання. Зберігає точність числових масивів (обсяги води/світла), мінімізуючи похибку групування.	Ітераційне розбиття великих міських масивів на підмножини з контролем порогу анонімності.
Кластеризація в ієрархіях	WHD (ієрархічна відстань)	Медичні та демографічні реєстри (e-Health). Використовує міські таксономії (*Вулиця* → *Район* → *Місто*) для безпечного кодування зв'язків.	Локальне перекодування кластерів; розділення порцій даних на «основи» (*stubs*) та «стовбури» (*trunks*).
Анонімізація за чутливістю	Вираз чутливості (рівень доступу + k)	Транспорт та IoT-давачі. Економить ресурси, анонімізуючи лише атрибути з малим діапазоном доменів (наприклад, часові інтервали).	Паралельні обчислення великих міських стеків даних через Apache Pig на базі Hadoop MapReduce.
MRMondrian *(масштабований)*	Поріг потужності вузла пам'яті	Геолокація, GPS-треки та камери. Швидко ділить географічні зони міста на розділи для безпечного аналізу мобільності містян.	Розподілений обхід через дерева індексації (PID-tree); обробка в пам'яті (In-Memory MapRe

Перша обчислюється шляхом, по-перше, ідентифікації найближчого спільного узагальнення «CCG» між кортежами, і, по-друге, врахування суми спотворень першого кортежу з «CCG» та суми другого з «CCG». Зрештою, відстань між класами еквівалентності визначається як зважена за розміром класу еквівалентності сума попередньої відстані, оскільки кортежі в межах класу еквівалентності є ідентичними.

Потім визначається основний алгоритм під назвою: K-анонімізація шляхом кластеризації в ієрархіях атрибутів «K-Anonymization by Clustering in Attribute hierarchies», який проходить циклом по всіх класах еквівалентності – кластерах вхідного набору даних та ітерує ті з них, що мають розмір менше k, знаходить найближчий до згаданого клас еквівалентності – кластер і узагальнює шляхом локального перекодування обидва класи еквівалентності – той, що ітерується в даний момент, і його найближчий [1]. Алгоритм зупиняється, коли не залишається жодного класу еквівалентності розміром менше k. Щоб допомогти анонімізації локального перекодування шляхом кластеризації, автори вводять поняття основи «stub» і стовбура «trunk» разом із технікою розділення для розмежування порцій даних, які мають бути узагальнені або оброблені.

Автори [51] описують підхід до багатовимірної анонімізації, який передбачає розділення набору даних на основі атрибутів «Quasi-ID» перед виконанням k-анонімності, що знижує значне навантаження, пов'язане з анонімізацією наборів даних із великою кількістю атрибутів. Ідея полягає в тому, щоб розглядати для анонімізації виключно домени атрибутів, діапазон яких є малим, що веде до покращення отримання інформації та ефективності. Для реалізації цієї ідеї вони вводять математичний вираз на основі чутливості, який використовує поняття ймовірностей атрибутів «Quasi-ID». Значення чутливості диктуватиме ступінь анонімізації, що застосовується до даних – мінімальні значення відповідних ймовірностей «Quasi-ID» або також діапазони інтервалів, які слід враховувати для анонімізації атрибута «Quasi-ID». Ймовірність, що стосується числового атрибута «Quasi-ID», є нічим іншим, як ймовірністю вибору числа в межах домену атрибута. Для категоріального випадку – це вибір певного значення з набору значень розглянутого атрибута Quasi-ID. Лінійний коефіцієнт виразу на основі чутливості базується на цих ймовірностях і параметрі k – k-анонімність [1]. Крім того, вираз чутливості лінійно пов'язаний із параметром рівня володіння даними користувача, який обернено впливає на значення чутливості: чим вищий цей параметр, тим нижче значення чутливості і тим вищий рівень анонімності, що застосовується до даних. Зокрема, у описаному випадку діапазон доменів, що розглядаються для анонімізації, буде

більшим, що сильніше впливає на сирі дані. Детальніше, запропонований алгоритм анонімізації даних працює ітеративно, розглядаючи на кожній ітерації підмножину атрибутів «Quasi-ID» для застосування k-анонімності. На кожній ітерації для анонімізації розглядаються лише ті атрибути «Quasi-ID», які мають найнижчі показники ймовірності – великі діапазони доменів атрибутів, тоді як записи групуються за рештою атрибутів «Quasi-ID». Реалізація цього алгоритму була досягнута завдяки «Apache Pig» [52], що дало змогу визначити скрипти, необхідні для трансформації записів даних.

Автори [7] пропонують ітераційну версію алгоритму «Mondrian» на основі підходу MapReduce. Кешоване дерево індексації, що містить основну інформацію про розділи даних «PID-tree» підтримується та спільно використовується вузлами системи. Щоб впоратися з надмірними обчислювальними витратами, які можуть спричинити великі набори даних «розумних міст», вони пропонують застосовувати послідовний MapReduce до кожного нового рівня розділів, які можуть бути отримані за інструкціями драйвера після завершення нової ітерації, замість того, щоб робити це рекурсивно для кожного нового розділу. Вони пропонують алгоритм, який враховує два типи атрибутів: числові та категоріальні. Ітерації драйвера з «розділення даних» тривають доти, доки всі розділи не задовольнятимуть визначений користувачем поріг потужності вузла – це максимальна кількість записів, які вузол може вмістити в пам’яті. Впродовж ітерації алгоритм прагне отримати набір розділів, оптимально розділених для анонімізації. При цьому анонімізація виконується в пам’яті з використанням парадигми MapReduce.

2.6 Архітектури та платформи для розширеного аналітичного опрацювання збереження конфіденційності в хмарах

У цьому параграфі розглянемо наукові роботи щодо архітектур і платформ, які забезпечують розширене аналітичне опрацювання даних «розумних міст» зі збереженням конфіденційності у хмарах. Зокрема, заглибимося у фреймворки та

системи, розроблені для підтримки приватності даних при забезпеченні складного аналізу в межах хмарного середовища.

Автори [53] пропонують хмарну систему електронної охорони здоров’я, що забезпечує доступ до електронних медичних карток пацієнтів (EHR) за запитом, приділяючи особливу увагу аспекту конфіденційності даних. Запропонована архітектура системи передбачає використання сторонніх центрів обробки даних для зберігання інформації з метою її подальшого використання за потреби. Для того, щоб дійти до кінцевих користувачів, дані передаються через мережу, де проблеми перевантаження та пропускної здатності вирішуються шляхом реплікації даних. Автори також пропонують модель збору даних, яка дає змогу проводити швидкий і надійний аналіз даних з урахуванням часу обробки та пріоритетності (див. рисунок 2.4).

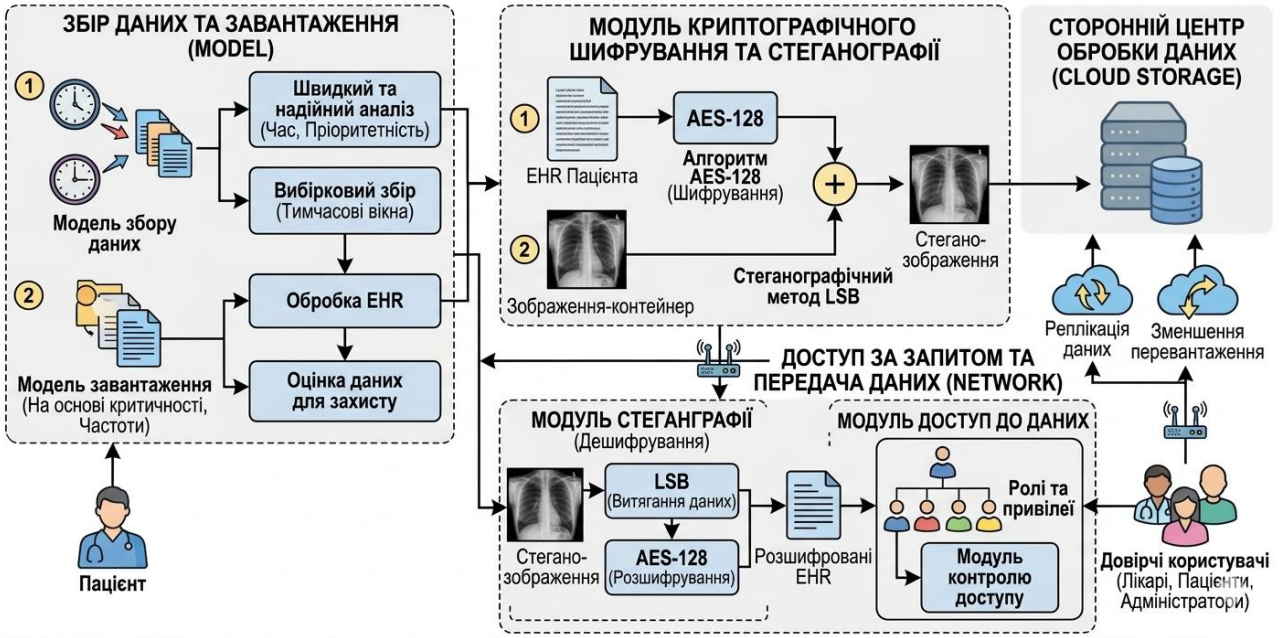

Рисунок 2.4 – Архітектура хмарної EHR-системи

Крім того, процедура завантаження даних враховує частоту консультацій пацієнта та критичність його стану. Дійсно, пропонується використовувати тимчасову інформацію на основі вікон для збору даних, щоб вибірково розмежувати найбільш інформативні дані для захисту.

Запропонована архітектура застосовує:

– модуль контролю доступу, який надає користувачам доступ до даних на основі їхніх ролей і привілеїв;

– модуль криптографічного шифрування, який відповідає за приховування інформації EHR у зображенні-контейнері за допомогою алгоритму AES-128 та стеганографічного методу LSB, що зберігається у сторонньому центрі обробки даних;

– модуль стеганографії, який гарантує, що розшифровані дані будуть доступні для модуля доступу до даних.

За запитом користувача модуль контролю доступу перевіряє дані через модуль анонімізації шляхом приховування чутливих значень та анонімізації записів квазі-ідентифікаторів «Quasi-ID» перед тим, як дані зможуть бути розкриті та доставлені користувачеві, який зробив запит.

У [54] автори пропонують систему для збору даних з IoT давачів для медичних цілей, яка є «дружньою» до споживання енергії. Дані надходять на медичний сервер через сполучну мережу, що забезпечується WI-FI, Інтернетом або стільниковими мережами. При цьому цілісність даних захищена за допомогою гомоморфного MAC (H-MAC), щоб жоден зловмисник не міг змінити вміст даних, що передаються мережею. У цьому випадку сторона, що приймає (медичний сервер), не потребуватиме розшифровки надісланого повідомлення для виконання обчислень над ним. Автори вважають, що це забезпечить додатковий захист даних і зменшить ризикований вплив мережі на дані. Процес автентифікації між сенсорними IoT пристроями та координатором, що генерує дані, а також останнього з медичним сервером захищені за допомогою традиційних хеш-функцій та механізмів секретних і відкритих ключів [55]. Крім того, конфіденційність даних гарантується за допомогою гомоморфного шифрування відомого тим, що дає змогу проводити обчислення над зашифрованими даними. Фактично, система спрямована на зниження енергоспоживання шляхом зменшення швидкості передачі даних у мережі, на яку припадає 70% загального споживання енергії, решта переважно приписується споживанню енергії, спричиненому агрегацією даних. Щоб досягти цього, вони використовують «DPM» [56] – підхід, призначений для

навчання моделей прогнозування як на сенсорних вузлах, так і на стороні медичного центру, щоб інформувати про те, які дані є найбільш релевантними для відправлення на медичний сервер. Основна ідея, яку вони використовують, полягає в тому, що якщо дані можуть бути достатньо точно, в межах визначеного порогу, спрогнозовані на сенсорному вузлі, то вони також можуть бути спрогнозовані на медичному сервері, оскільки історичні дані, що використовуються для навчання моделей, однакові для обох сторін. Крім того, виконується одночасне оновлення пам'яті для зберігання даних, необхідної для навчання моделі, на обох сторонах з використанням прогнозованих значень даних, а також нових отриманих значень. Завдяки такому підходу можна врешті-решт відрізнити корисні дані від надлишкових перед відправленням їх до медичного центру. Зрештою, вдасться уникнути безглуздих передач даних, що спричинить зменшення мережевого трафіку і, як наслідок, зниження загального енергоспоживання системи.

Дослідники [57] розглядають хмарну систему, що дає можливість організаціям, як от лікарні або медичні центри, накопичувати анонімізовані дані пацієнтів. Їхня система зводиться до постачальника хмарних систем, який здійснює внутрішню анонімізацію для подальшого розповсюдження прихованих даних. Підхід до анонімізації, який вони застосовують, полягає в накладанні декількох кроків, що забезпечують анонімізацію, які, за їхніми словами, утворюють ефективну та результативну техніку анонімізації клінічних даних. Основний робочий процес анонімізації зводиться до уточнення та стандартизації даних за допомогою використання нормального розподілу як еталона для виявлення аномальних записів «outliers» та їх відкидання, оскільки вони становлять найбільш вразливі записи для атак. По-друге, вони використовують k-means++ [1], щоб кластеризувати записи способом, який, як вони стверджують, є корисним для придатності даних. Дійсно, вони пояснюють, що за такого підходу анонімізація даних, через k-анонімність, стає простішою та ефективнішою, оскільки вимагає меншого узагальнення, тому що елементи даних адекватно кластеризовані на основі значень «QID». Нарешті, вони застосовують метод k-анонімності до кожного кластера окремо, отримуючи

таким чином анонімізовані фрагменти даних, які об’єднуються у фінальний анонімізований набір даних, готовий до доставки в організації охорони здоров’я.

У [58] висвітлюються обмеження сучасних методів анонімізації; дійсно, автори стверджують, що при використанні стандартних методів анонімізації чутлива інформація все ще перебуває під ризиком розкриття. Вони додають, що якщо чутлива інформація класу «QID» підпадає під спільний домен семантичної близькості, виведення такої чутливої інформації стає приблизно можливим через те, що вони називають «порушенням категоріальної близькості» (categorical proximity breach), коли чутлива інформація приблизно визначається через її приналежність до потенційного домену семантичної близькості. Для реалізації своєї ідеї вони пропонують метрику відстані для чутливих атрибутів. Математично метрика визначається як відстань між двома чутливими значеннями відносно максимально можливої відстані між двома чутливими значеннями, що визначені деревом таксономії даного атрибута [1]. Відстань між двома чутливими значеннями певного атрибута обчислюється через їхній найближчий спільний предок. Відповідно до їхнього фреймворку, анонімність чутливих значень забезпечується шляхом встановлення порогового рівня несхожості. На додаток до цього, вони визначають індекс близькості для двох чутливих значень на основі їхньої відстані, як згадувалося раніше, і поширюють це визначення на групи «QID». Крім того, визначено метрику спотворення для фіксації спотворення даних, що виникає в результаті процесу узагальнення. Через згадані визначення вони описують підхід до кластеризації з урахуванням близькості, який має на меті мінімізувати як близькість кластерів, так і спотворення, що виникає в результаті процесу узагальнення.

Для подальшого вдосконалення свого підходу вони побудували масштабовану версію раніше згаданого алгоритму. Дійсно, вони визначають двоетапну кластеризацію з урахуванням близькості за допомогою MapReduce. Перший крок процесу генерує кластери за допомогою методу t-предків «t-ancestors method», де кластери формуються шляхом групування записів разом з їхнім найближчим предком [1]. Потім предок кластера переобчислюється з урахуванням нового формування: обчислення запису-предка кластера

виконується шляхом розрахунку медіани значень атрибутів «QID» кластера (числові значення) або найближчого спільного предка тих самих значень атрибутів QID записів кластера (категоріальні значення). Другий крок полягає в застосуванні методу агломеративної кластеризації до кожного зі згенерованих кластерів для отримання набору підкластерів із кожного початкового кластера. Потім виконується узагальнення для кожного з кластерів, отриманих на другому етапі. Ідея алгоритму агломеративної кластеризації полягає в об’єднанні найближчих кластерів разом, причому відстань між кластерами вважається функцією двох найвіддаленіших записів, що належать до кожного кластера, наприклад, максимальне значення відстаней близькості між записами розглянутих кластерів [1].

2.7 Висновок до другого розділу

В другому розділі кваліфікаційної роботи досліджено конфіденційність високовимірних даних «розумних міст». Розглянуто конфіденційність OLAP «розумних міст». Проаналізовано конфіденційність великих багатовимірних даних «розумних міст» у нових сценаріях застосування. Описано конфіденційність через багатовимірну анонімізацію в системах аналітичного опрацювання великих даних. Досліджено архітектури та платформи для розширеного аналітичного опрацювання збереження конфіденційності в хмарах.

3 МЕТОДИ, АРХІТЕКТУРНІ РІШЕННЯ ТА ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ БАГАТОВИМІРНОГО АНАЛІЗУ ВЕЛИКИХ ДАНИХ «РОЗУМНИХ МІСТ»

3.1 Загальні інструменти та підходи аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст»

Корисність (даних) (Utility) стосується придатності даних для аналізу, зазвичай після анонімізації. Щоб зафіксувати корисність даних, вирази метрик можуть включати обчислення як над кінцевими анонімізованими даними, так і над вихідними даними. Зокрема, міра корисності даних «розумних міст» повинна фіксувати різницю між розподілами даних до та після анонімізації [1]. Двома важливими аспектами при оцінці якості даних за допомогою метрики корисності є втрата інформації внаслідок анонімізації та важливість атрибутів. Добре визначена метрика корисності має бути здатною фіксувати внутрішні аспекти, що впливають на якість даних для розглянутого застосування. Задля ясності, оскільки часто відбувається змішування понять корисності даних та якості даних, слід наголосити, що ці два поняття є різними. У той час як корисність даних доцільно оцінювати перед застосуванням методів інтелектуального аналізу даних, якість даних застосовується у ширшому контексті. Дійсно, якість даних стосується характеристик самих даних, на відміну від їхньої корисності для іншого процесу [1], зокрема:

– точність даних «розумних міст», наприклад, після процесу очищення даних вона оцінює, наскільки значення атрибута відрізняється від його початкового сирого значення;

– повнота даних «розумних міст» – оцінює ступінь пропущених даних після очищення;

– узгодженість даних «розумних міст» – ступінь кореляції атрибутів, яка зберігається після очищення.

Характеристики якості даних «розумних міст» подано в таблиці 3.1.

Таблиця 3.1 – Характеристики якості даних «розумних міст»

Характеристика якості даних	**Визначення**	**Що саме оцінюється**	**Приклади в «розумному місті»**	**Наслідки низької якості**
Точність даних	Наскільки виміряні або очищені дані відповідають реальному стану міської інфраструктури та середовища	Відхилення між даними сенсорів і реальними фізичними значеннями	Показники якості повітря, трафіку, температури, рівня шуму після фільтрації або очищення	Хибні прогнози заторів, неточні екологічні оцінки, помилкові управлінські рішення
Повнота даних	Ступінь наявності всіх необхідних даних у міських потоках після збору та обробки	Частка пропущених записів або відсутніх значень у часових рядах сенсорів	Відсутні вимірювання з окремих IoT-давачів (паркування, енергоспоживання, транспорт)	«Сліпі зони» в моніторингу міста, зниження якості аналітики та моделей прогнозування
Узгодженість даних	Ступінь логічної та статистичної узгодженості між різними джерелами міських даних	Збереження кореляцій і відсутність суперечностей між пов’язаними атрибутами	Узгодженість між потоками: трафік ↔ витрати пального ↔ рівень CO_2; енергоспоживання ↔ температура	Конфліктні дані між системами, некоректні інсайти для управління містом

Реконструкція даних стосується набору алгоритмів, які намагаються відновити початкові дані «розумних міст», що пройшли через процес із втратами. Як випливає з їхньої назви, такі процеси спричиняють втрату даних, тому головним викликом методів реконструкції даних є забезпечення відновлення початкових даних у повному обсязі – реконструкція без втрат. Прикладів методів реконструкції даних у літературі доволі багато [1]. Стиснені дані про трафік

реконструюються в контексті масштабного Інтернету транспортних засобів «Internet of Vehicles» за допомогою підходу на основі нейронних мереж.

Деякі дослідження зосереджені на конкретному застосуванні реконструкції даних [59], як, наприклад, для зображень дистанційного зондування або для відновлення даних у багатовимірних часових рядах, тоді як інші є загальнішими та застосовуються до будь-якого типу даних, де пропонується метод, заснований на апроксимації найменших квадратів у поєднанні з технікою рандомізованого сингулярного розкладу SVD.

Крім того, деякі інші наукові праці розглядають реконструкцію даних у сфері збереження конфіденційності. Наприклад, [60] описують метод відновлення розподілу для анонімізованих даних, який намагається відновити оптимальну корисність для чутливих даних користувачів. Іншою вартою уваги науковою роботою є [61], яка надає рішення для реконструкції даних після застосування методу анонімізації з використанням диференційної конфіденційності для захисту геолокації в контексті периферійної мережі «edge network» «розумного міста».

3.2 Методи аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» за впливом на цілісність

Техніка втрати даних стосується методу обробки даних, який призводить до часткової втрати даних після виконання завдання із «втратами», зокрема, процедури анонімізації. Наприклад, в [1] згадано модель анонімізації, розроблена для роботи з потоками транзакційних даних. Модель використовує нормалізацію та пригнічення «suppression» даних для анонімізації чутливих значень. На жаль, відомо, що ці методи спричиняють втрату даних «розумних міст». Інша варта уваги наукова робота описана в [1], де визначено розширену версію (α, k)-анонімності. Запропонований метод передбачає застосування (α, k)-анонімності та генерацію двох таблиць для конфіденційної публікації. Ці дві таблиці мають спільний стовпець ID групи («group-ID»), який присвоює кожному кортежу ідентифікатор класу еквівалентності. Водночас

демонструється, як цей метод може зменшити спотворення даних порівняно з традиційною публікацією даних, однією узагальненою таблицею [1]. Крім того, автори стверджують, що вихідні дані неможливо відновити, якщо не буде виконано з'єднання із втратами «lossy join» згенерованих таблиць.

Порівняльна характеристика технік анонімізації даних «розумних міст» подана в таблиці 3.2.

Таблиця 3.2 – Порівняльна характеристика технік анонімізації даних «розумних міст»

Техніка	**Тип підходу**	**Використані методи**	**Переваги**	**Недоліки / обмеження**
Анонімізація з втратами даних (lossy anonymization)	З втратами даних	Нормалізація даних, suppression (пригнічення даних)	Забезпечує конфіденційність чутливої інформації	Призводить до втрати частини даних та зниження точності аналітики
Розширена k-анонімність із двома таблицями	З втратами даних	k-анонімність, поділ таблиці, lossy join	Менше спотворення даних порівняно з однією узагальненою таблицею; підвищений захист конфіденційності	Неможливість повного відновлення вихідних даних без з'єднання із втратами
Анонімізація без втрат для асоціативних правил	Без втрат даних	Алгоритм оптимізації світлячка (Firefly Optimization Algorithm)	Збереження цілісності та повноти даних	Висока обчислювальна складність оптимізаційних алгоритмів
Очищення та відновлення даних без втрат	Без втрат даних	Оптимізація роєм частинок (PSO) + алгоритм оптимізації кита (WOA)	Дані можуть бути повністю відновлені після обробки	Складність реалізації та потреба в оптимальному підборі параметрів
Анонімізація бази відбитків пальців	Без втрат даних	Pixel Embeddability Criterion, логічні операції NOT та AND	Повне відновлення початкового зображення; захист біометричних даних	Орієнтованість переважно на графічні/біометричні дані

Метод обробки даних без втрат «lossless» стосується техніки, яка зберігає дані недоторканими впродовж усього процесу обробки. У літературі є численні наукові праці, що розв’язують це завдання, серед яких доцільно згадати авторів [62], де детально описується метод анонімізації без втрат для асоціативних правил. Запропонована техніка спрямована на захист чутливих правил, доступ до яких здійснюється з бази даних асоціативних правил, за допомогою алгоритму оптимізації світлячка.

Інше дослідження, що стосується методів без втрат, описане в [63], де запропонований алгоритм генерує оптимальні ключі для очищення даних, а також для їх відновлення за допомогою комбінації методів, а саме: оптимізації роєм частинок та алгоритму оптимізації кита. Ще одна цікава наукова робота [65] пропонує анонімізацію бази даних відбитків пальців за допомогою нової техніки, що визначає «критерій вбудовування пікселів» (pixel embeddability criterion), який вирішує, чи може певний піксель у зображенні відбитка пальця бути виділений як елемент приховування зображення. Використовуючи вираз, пов’язаний із взаємозв’язком пікселів, заснований на логічних «NOT» та «AND», початковий відбиток пальця потім відновлюється через процес реконструкції шаблону.

3.3 Захист та приватність процесів аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст»

Анонімізація – це процес приховування всіх ключових ідентифікаційних ознак, які можуть розкрити конфіденційну інформацію про осіб, коли власники даних (лікарні, страхові компанії тощо) передають свої дані третій стороні (клінікам, дослідницьким центрам тощо) [1]. Її кінцевою метою є захист чутливої інформації, наприклад, про пацієнтів у сфері охорони здоров’я, від розкриття ненадійній стороні шляхом деперсоналізації записів даних. Дійсно, анонімізація записів даних покликана зберегти приватність та інформацію про осіб. Зазвичай набір даних розділяється за ознаками на значення атрибутів «QID» «квазі-

ідентифікаторів» та чутливі значення. Процес анонімізації значень атрибутів «QID» полягає в:

– пригніченні «suppressing»;
– узагальненні «generalizing»;
– збуренні «perturbing»;
– перестановці «permuting»;
– анатомізації «anatomizing/bucketing».

Пригнічення відбувається, коли значення атрибутів замінюються зірочкою. Узагальнення досягається шляхом заміни фактичних значень узагальненим значенням з їхньої ієрархічної лінії – узагальнення базується на заздалегідь визначеній ієрархії значень атрибутів. Збурення здійснюється шляхом додавання шумних даних до записів, що анонімізуються, розрахованим способом, щоб зберегти певні статистичні характеристики записів незмінними [1]. Перестановка досягається шляхом розділення записів на групи, де чутливі значення перемішуються всередині кожної групи. Анатомізація (або анатомія, або бакетизація) реалізується шляхом відокремлення квазі-ідентифікаторів та чутливих атрибутів у дві окремі таблиці зі збереженням зв'язкової інформації, щоб зробити чутливі значення такими, що можна віднести до конкретного кортежу квазі-ідентифікаторів [1]. Анатомізація часто застосовується в поєднанні з k-анонімністю або l-різноманітністю.

Потреби у збереженні конфіденційності випливають із вимог до конфіденційності даних, які мають на меті захистити приватність пацієнтів. Ці вимоги зазвичай накладаються національними та міжнародними законами про захист даних, такими як Закон США про підзвітність і наступність медичного страхування «HIPAA» [64] та Загальний регламент про захист даних (GDPR). Збереження конфіденційності встановлюється за допомогою набору методів і підходів, спрямованих на перетворення чутливих даних у безповоротний формат.

Порівняльна характеристика методів анонімізації даних «розумних міст» подана в таблиці 3.3.

Таблиця 3.3 – Порівняльна характеристика методів анонімізації даних «розумних міст»

Метод анонімізації	**Суть методу**	**Як змінюються дані**	**Переваги**	**Недоліки**
Пригнічення (Suppressing)	Видалення або приховування окремих значень атрибутів QID	Частина значень замінюється символами типу `*`, `NULL` або повністю вилучається	Простота реалізації; високий рівень захисту конфіденційності	Значна втрата інформації та зниження аналітичної цінності даних
Узагальнення (Generalizing)	Заміна точних значень більш загальними категоріями	Наприклад, точний вік замінюється діапазоном, адреса — регіоном	Зберігає структуру даних; підтримує статистичний аналіз	Зменшує точність та деталізацію даних
Збурення (Perturbing)	Додавання випадкових змін або шуму до даних	Значення модифікуються випадковим чином без повного видалення	Зберігає загальні статистичні властивості набору даних	Може спотворювати результати аналізу та прогнозування
Перестановка (Permuting)	Перемішування значень між записами	Значення атрибутів переставляються між різними кортежами	Зменшує ризик ідентифікації особи; зберігає розподіл даних	Може порушувати логічні зв’язки між атрибутами
Анатомізація / Bucketing	Поділ даних на групи з окремим збереженням QID та чутливих атрибутів	Дані розбиваються на «кошики» або таблиці з непрямим зв’язком	Знижує ризик розкриття чутливих даних; краще зберігає корисність інформації	Можливість повторної ідентифікації при додаткових зовнішніх даних

Найчастіше збереження конфіденційності пов’язане з поняттям обміну даними. Дійсно, обмін даними – це практика надання третій стороні можливості отримувати дані для отримання нових знань і, зрештою, впровадження високоточної персоналізованої медицини. Важливим застосуванням у сфері «розумної» охорони здоров’я є анонімізація, метою якої є максимально можливе приховування особистості осіб/пацієнтів при збереженні зв’язків із їхньою

чутливою інформацією. Слід зазначити, що чутлива інформація має піддаватися мінімальному пригніченню «suppression» для підтримки корисності даних. Приклади методів збереження конфіденційності включають: k-анонімність [66] та l-різноманітність [68].

Шифрування даних «розумних міст» описує набір методів, що використовуються для кодування даних, зазвичай з використанням алгоритму шифрування з відкритим/приватним ключем, перед маніпулюванням ними, наприклад, надсиланням їх через мережу. За допомогою процесу шифрування звичайні дані перетворюються на зашифровані. І навпаки, процес дешифрування перетворює зашифровані повідомлення на звичайний текст. Основна ідея полягає у використанні ітерацій (так званих раундів) підстановки тексту та перестановки в повідомленні для шифрування, де біти відповідно змінюються. Приклад таких методів описано в [67].

3.4 Технічні методи маскування та трансформації великих за обсягом наборів та колекцій даних «розумних міст»

Маскування даних охоплює набір методів, відповідальних за приховування даних з метою збереження конфіденційності. На відміну від інших методів анонімізації, маскування даних спрямоване на приховування даних при збереженні початкової структури значень атрибутів незмінною за допомогою методів:

– перемішування «Shuffling» [69], що змінює порядок значень атрибутів у стовпці набору даних; анулювання (Nulling) (використання загального символу замість фактичних літер/цифр);

– підстановка «Substitution», яка використовує інше синтетичне (фейкове) значення атрибута для маскування цільового значення даних;

– рандомізація «Randomization» [70], що полягає в заміні значень атрибутів випадковими даними;

– перекіс «Skewing», який впливає на значення атрибутів через дисперсію розподілу домену, щоб змінити їхні початкові справжні значення.

– розмиття «Blurring» також є методом маскування даних, який має на меті змінити кожне окреме значення атрибута відповідним чином, щоб отримати оновлені значення, що підпадають під заздалегідь визначений діапазон.

Менш поширеним методом є усереднення «Averaging», де зберігається підсумковий агрегований зв’язок із реальними даними, поки певні значення атрибутів проходять оновлення [1].

Збурення «Perturbation» – це техніка, що використовується для зміни значення атрибута з метою його анонімізації. Зазвичай ця техніка використовує відомий розподіл, наприклад, розподіл даних, що підлягають анонімізації, для зміни вихідних даних, подаючи їх таким чином у формі, що захищає конфіденційність – збурення на основі розподілу ймовірностей. З іншого боку, збурення може бути виконане «випадковим» чином з використанням синтетично згенерованих даних для остаточної заміни початкових значень новими – фіксоване збурення даних [1]. В обох типах збурень шумові дані додаються до оригінальних даних ідеальним чином, щоб зберегти певні статистичні знання, закладені в даних, як от середнє значення домену, дисперсія або кореляція. На жаль, багато описаних у літературі методів збурення при застосуванні до даних «розумних міст» вносять упередженість «bias», що визначається як різниця у відповідях на запити при використанні збурених та оригінальних даних, що призводить до стирання певної статистичної інформації – підсумкових показників даних, як от середнє значення, а також до знищення кореляцій між значеннями атрибутів. Слід зазначити, що однією з технік збурення, яка витримує всі чотири типи упередженості [711], є загальне адитивне збурення даних «GADP» [71].

Бакетизація (Bucketization) – це процес анонімізації, який відокремлює класи еквівалентності, зокрема значення атрибутів «QID» у групах записів, від чутливих атрибутів, проте зберігає зв’язок між ними для цілей аналізу даних. Шляхом послаблення зв’язків між «QID» та чутливими атрибутами, чутливі атрибути всередині бакета можуть бути однаково приписані до певного запису, оскільки вони належать до бакета, а не до впорядкованого списку.

Характеристика методів маскування великих за обсягом наборів та колекцій даних «розумних міст» подана в таблиці 3.4.

Таблиця 3.4 – Характеристика методів маскування великих за обсягом наборів та колекцій даних «розумних міст»

Метод маскування	**Опис методу**	**Застосування в даних розумних міст**	**Переваги**	**Недоліки**
Перемішування (Shuffling)	Зміна порядку значень у межах одного атрибута; значення залишаються тими самими, але їх прив’язка до конкретних записів порушується	Дані про поїздки транспорту або споживання енергії перемішуються між користувачами/часовими записами	Зберігає розподіл даних; проста реалізація	Руйнує зв’язок «користувач–значення», ускладнює поведінковий аналіз
Анулювання (Nulling)	Заміна значень на загальний маркер відсутності (NULL або символи типу *)	Приховування точних координат або персональних даних у міських сервісах	Високий рівень конфіденційності	Втрата інформації та зниження аналітичної цінності
Підстановка (Substitution)	Реальні значення замінюються синтетичними (фейковими), але правдоподібними даними зі збереженням структури	Підміна ідентифікаторів користувачів або адрес у smart city платформах	Зберігає структуру даних; зменшує ризик ідентифікації	Може вводити в оману при некоректному генеруванні синтетичних значень
Рандомізація (Randomization)	Значення замінюються або модифікуються випадковими даними (додавання шуму або повна заміна)	Дані сенсорів IoT (трафік, енергія, шум)	Захищає конфіденційність; ускладнює реідентифікацію	Спотворює аналітику та прогнози
Перекіс (Skewing)	Системна зміна значень, що змінює статистичний розподіл даних і домен атрибутів	Дані навантаження на міські мережі або транспортні потоки	Ускладнює статистичну ідентифікацію	Викривляє аналітичні моделі та розподіли
Розмиття (Blurring)	Точні значення замінюються наближеними значеннями в межах заданого діапазону	Геолокація користувачів або швидкість транспорту	Зберігає загальні тенденції даних	Втрата точності та деталізації

Технічно атрибути «QID» деасоціюються від чутливих атрибутів, що призводить до створення двох таблиць, з’єднаних через стовпець приналежності до групи «Group-ID» [1]. Характеристика методів трансформації великих за обсягом наборів та колекцій даних «розумних міст» подана в таблиці 3.5.

Таблиця 3.5 – Характеристика методів трансформації великих за обсягом наборів та колекцій даних «розумних міст»

Метод трансформації	**Опис методу**	**Застосування в даних розумних міст**	**Переваги**	**Недоліки**
Усереднення (Averaging)	Частина значень атрибутів оновлюється шляхом агрегування, при цьому зберігається лише узагальнений зв’язок із реальними даними	Середній рівень трафіку, енергоспоживання або забруднення по району	Зменшує обсяг даних; зберігає загальні тренди	Втрачається деталізація та індивідуальні спостереження
Збурення (Perturbation)	Значення атрибутів змінюються шляхом додавання шуму або заміни на синтетичні дані; може бути розподільчим або випадковим; включає методи на основі ймовірностей або фіксовані	Сенсорні дані IoT (шум, температура, трафік) з доданими випадковими змінами	Захищає конфіденційність; зберігає загальні статистики	Вносить bias, спотворює середні значення та кореляції
Бакетизація (Bucketization)	Розділення даних на класи еквівалентності (бакети), де QID відокремлюються від чутливих атрибутів, але зв’язок між ними частково зберігається	Дані пацієнтів, транспортні або соціальні дані з групуванням користувачів	Зменшує ризик ідентифікації; зберігає корисність для аналізу	Можлива неоднозначність відповідності між QID і чутливими даними

Стовпець «Group-ID» відповідає за «приблизне» відстеження походження чутливих значень і того, до якої групи записів вони належать. Також підтримується підрахунок кількості чутливих значень у кожному бакеті. Зауважте, що кількість різних ідентифікаторів у стовпці «Group-ID» подано загальну кількість згенерованих бакетів. Це може призвести до посилення конфіденційності та зниження ризиків атак на дані.

Перекодування даних означає застосування до них трансформації, що зберігає конфіденційність, наприклад, узагальнення. Локальне перекодування даних полягає в тому, щоб дало змогу одному значенню атрибута в класі еквівалентності «QID» мати два або більше трансформованих, наприклад, узагальнених, значень. Іншими словами, локальне перекодування домену значень означає можливість гнучкого призначення різних захисних, наприклад, узагальнюючих, значень для цього домену [1]. Відображення домену значень на його узагальнений аналог виконується на основі кортежів. Дійсно, значення атрибутів трансформуються, наприклад, узагальнюються, «за кортежами» (а не за доменами). Зокрема, для локального перекодування домену значень «сусідня» порція входжень буде розглянута для певного узагальнення, тоді як інші «близькі» групи входжень будуть розглянуті для інших узагальнюючих значень. Перевагою такої гнучкості в роботі з доменними значеннями є потенційне забезпечення більшої корисності даних шляхом мінімізації втрат інформації під час процесу анонімізації.

Глобальне перекодування значень домену під час анонімізації полягає в узагальненні всіх входжень атрибутного домену до одного узагальнюючого значення, щоб досягти цільового порогу анонімізації. Зауважте, що такі методи можуть надмірно узагальнювати значення домену, що призводить до збільшення втрати даних [1]. Втрата інформації в результаті глобального перекодування набагато важливіша, ніж внаслідок локального перекодування. Крім того, глобальне перекодування можна вважати окремим випадком локального перекодування.

3.5 Спеціалізовані методи опрацювання великих за обсягом наборів та колекцій даних «розумних міст»

Методи, орієнтовані на атрибути, мають на меті анонімізувати значення атрибутів великих за обсягом наборів та колекцій даних «розумних міст» шляхом виконання операцій, пов'язаних із цими атрибутами. Зазвичай цей тип техніки розрізняє процедуру анонімізації великих за обсягом наборів та колекцій даних «розумних міст», яку слід застосувати, залежно від типу категоріального чи числового атрибута. В обох випадках техніка, орієнтована на атрибути, дасть можливість анонімізувати значення атрибутів за допомогою двох різних підходів. один для числових значень атрибутів, а інший – для категоріальних [1]. Приклади включають ідентифікацію ширини інтервалу для групування числових значень та використання середнього значення для заміни цільових значень. Для категоріальних атрибутів приклади включають техніку зіставлення на основі ідентифікаторів або кластеризацію кортежів за допомогою алгоритму PAM k-medoid з подальшим розділенням кортежів, де згруповані в бакети кортежі проходять перевірку на анонімізацію, наприклад, l-різноманітність.

На відміну від методів, орієнтованих на атрибути, методи, орієнтовані на діапазони даних, зосереджують увагу на забезпеченні анонімності в цільових наборах даних шляхом застосування специфічних процедур обробки даних над спеціальними «ad-hoc» колекціями елементів даних, що зберігаються в цих наборах даних. Основна ідея тут полягає в розробці моделей і методів, які інтелектуально використовують специфічні властивості даних для досягнення бажаної анонімізації. Наприклад, розподіли, діапазони, максимальне значення, мінімальне значення тощо. Це стосується анонімізації на основі агрегації [72] або анонімізації на основі обфускації [73]. Деякі підходи також поєднують декілька методів для досягнення цього ефекту [74].

Характеристика методів опрацювання великих за обсягом наборів та колекцій даних «розумних міст» подана в таблиці 3.6.

Таблиця 3.6 – Характеристика методів опрацювання великих за обсягом наборів та колекцій даних «розумних міст»

Метод опрацю-вання	**Підхід**	**Суть**	**Приклади в «розумних містах»**	**Плюси**	**Мінуси**
Атрибутно-орієнтовані	Поатрибутна анонімізація	Обробка кожного атрибута окремо залежно від типу (числовий/категоріальний)	Сенсорні та демографічні дані	Гнучкість	Складність і втрата точності
Орієнтовані на діапазони	Статистичні властивості	Використання мін/макс, розподілів, обфускації	Трафік, енергоспожи-вання	Масштабова-ність	Втрата деталей
Агрегаційні	Об’єднання даних	Обчислення сум/середніх, у т.ч. над зашифрованими даними	Міські KPI, енергетика	Захист приватності	Менша деталіза-ція
Гібридні	Комбінація методів	Поєднання декількох технік анонімізації	Аналітичні платформи smart city	Баланс якості й безпеки	Висока склад-ність

Метод агрегації використовує дані, зібрані з декількох джерел, для виконання певних операцій агрегування над ними [1]. У сфері збереження конфіденційності великих за обсягом наборів та колекцій даних «розумних міст» агрегація може виконуватися над зашифрованими даними, щоб забезпечити безпеку даних під час виконання завдання агрегування. Техніка, що забезпечує такі приватні агрегації, називається гомоморфним шифруванням. Агрегація використовується для приховування оригінальних даних, водночас даючи можливість здійснювати подальші маніпуляції з ними компонентами системи анонімізації. Наприкінці процесу агреговані дані будуть доставлені кінцевому користувачу, щоб дати йому можливість отримати знання зі статистичних властивостей, що містяться в цих агрегаціях.

3.6 Аналітичні метрики великих за обсягом наборів та колекцій даних «розумних міст»

Метрики на основі помилок застосовують певні критерії, що визначаються на основі заданих метрик помилок, які, у свою чергу, можуть бути базовими або похідними від декількох (інших) метрик [1]. Наприклад, у випадку OLAP-кубів великих за обсягом наборів та колекцій даних «розумних міст» популярна метрика на основі помилок полягає в мінімізації помилки відповіді на запит за умови сукупності «типових» запитів, що застосовуються до цільового куба даних.

Інші ініціативи використовують методи шифрування великих за обсягом наборів та колекцій даних «розумних міст», засновані на максимізації помилок розшифрування, щоб досягти безпеки конфіденційності цільових записів [75], зокрема в медичних установах. Крім того, інші підходи розглядають помилку, яку може зробити гіпотетичний зловмисник при виявленні властивостей набору даних, наприклад, розподіл конкретних атрибутів у цільовому наборі даних [76].

Класифікацію аналітичних метрик великих за обсягом наборів та колекцій даних «розумних міст» подано на рисунку 3.1.

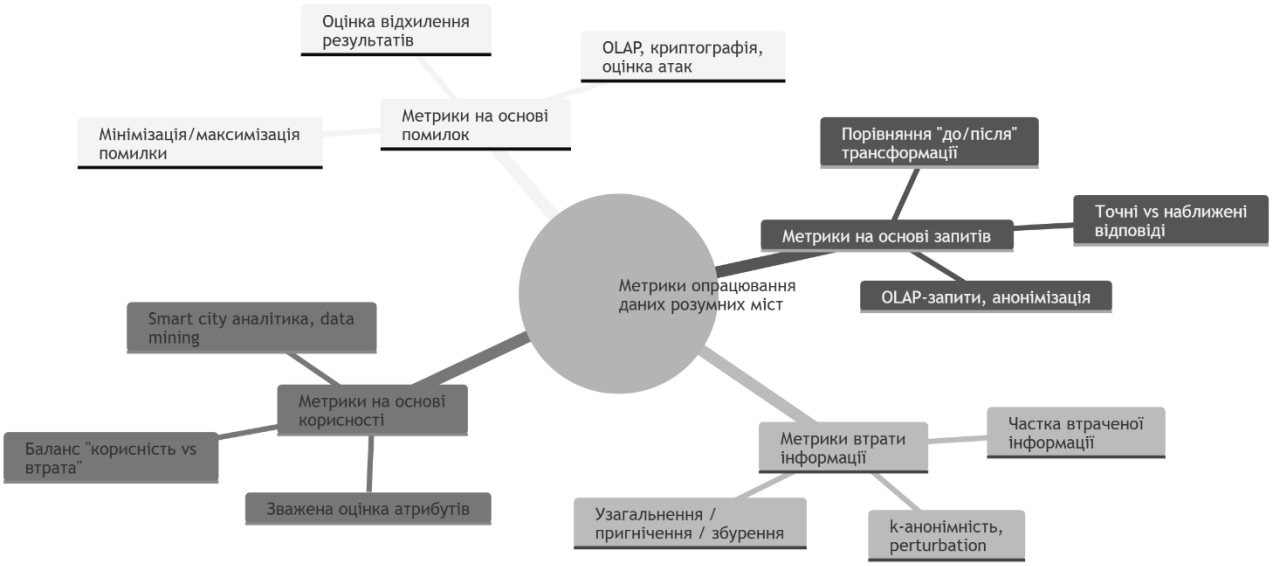

Рисунок 3.1 – Класифікація аналітичних метрик великих за обсягом наборів та колекцій даних «розумних міст»

Метрики на основі запитів – це тип метрик, що базуються на порівнянні відповідей до та після застосування техніки трансформації до цільових великих за обсягом наборів та колекцій даних «розумних міст», наприклад, анонімізації, стиснення тощо [1]. Це важливо для того, щоб інформувати користувачів про якість трансформованих даних і про те, чи відповідають вони певним вимогам якості. Ця метрика також може бути корисною для кількісної оцінки змін у даних, спричинених застосованою трансформацією. Наприклад, у сфері OLAP метрики на основі записів є оцінкою впливу методу трансформації даних (анонімізація, стиснення тощо) на баланс між корисністю даних та втратою даних або втратою інформації). Крім того, метрики на основі OLAP-запитів можуть оцінювати вплив на дані шляхом порівняння точної відповіді та приблизної відповіді – тих, що були отримані на основі оригінальних даних та трансформованих даних відповідно.

Метрики втрати інформації використовуються для оцінки ступеня втраченої інформації під час процесу анонімізації великих за обсягом наборів та колекцій даних «розумних міст». Зокрема, метрика оцінює частку втраченої інформації внаслідок анонімізації певного домену. Наприклад, у випадку узагальнення вона може бути виражена як різниця між верхньою та нижньою межами узагальненого домену відносно такої ж різниці меж того самого домену.

Тоді як у випадку пригнічення «suppression», метрика втрати інформації може бути виражена через відношення кількості пригнічених записів до загальної кількості записів у домені. Крім того, у випадку збурення «perturbation» втрата інформації вимірюється через несхожість між початковими та очищеними даними.

Метрики на основі корисності вимірюють вплив анонімізації на корисність великих за обсягом наборів та колекцій даних «розумних міст», враховуючи як втрату інформації для кожного атрибута, так і їхню важливість для аналізу даних [1]. На відміну від метрики втрати інформації, метрика на основі корисності використовує зважування атрибутів залежно від їхньої придатності для аналізу даних.

3.7 Робота з великими даними та середовищами «розумних міст»

3.7.1 Інформаційно-технологічна архітектура та налаштування

Ці налаштування стосуються централізованого середовища, де алгоритми або системи використовують централізований підхід для зберігання та обробки великих за обсягом наборів та колекцій даних «розумних міст». Усі необхідні операції виконуються в одному місці. Алгоритми анонімізації, що використовують такі налаштування, виконують завдання на одній машині в одному місці над усім набором даних [67].

Розподілені налаштування стосуються розподіленого середовища «розумних міст», де алгоритми або системи використовують розподілені можливості для зберігання та обчислень. Зазвичай це передбачає використання мережі для транспортування даних або обчислених значень між двома географічно віддаленими локаціями, наприклад, налаштування клієнт-сервер.

У випадку анонімізації даних розподілені компоненти можуть потребувати паралельної роботи незалежним чином, коли завдання анонімізації розділяється між «вузлами»-учасниками, яких часто називають «воркерами», таким чином, щоб це сприяло продуктивності, наприклад, часу обробки, зберігаючи при цьому хорошу якість отриманого анонімізованого набору даних, наприклад, з точки зору корисності [1]. Багато наукових праць розглядали розподілені налаштування для анонімізації великих наборів даних, наприклад, у сфері «розумної» охорони здоров’я [45] або ширше для великих за обсягом наборів та колекцій даних «розумних міст» [77].

Масштабованість стосується здатності системи «розумних міст» справлятися з більшим навантаженням шляхом динамічного, адекватного та автоматичного самостійного розподілу достатньої кількості необхідних ресурсів для зберігання, обчислення та належного функціонування [1]. Загальна евристика, яку можуть приймати алгоритми, що забезпечують масштабованість, полягає в зниженні обсягу даних для передачі мережею шляхом розгляду спрощеного подання даних. Варто зауважити, що декілька літературних праць

забезпечують масштабованість за допомогою розподілених архітектур, що використовують парадигму програмування MapReduce у хмарі.

Інший приклад масштабованих систем описаний у [77], де був визначений метод анонімізації на основі «Mondrian» шляхом розширення алгоритму «Mondrian» [78] через інноваційний підхід до розбиття на основі квантилів, що використовує ранжування значень атрибутів набору даних «розумних міст». У їхньому підході вузлам-виконавцям «worker nodes» призначаються ексклюзивні частини набору даних, що оптимізує швидкість обміну даними між вузлами, які потім можуть незалежно анонімізувати свої відповідні локальні фрагменти даних. Порції даних визначаються за допомогою координатора, який відповідно встановлює умови розбиття.

3.7.2 Аналіз та моделювання великих за обсягом наборів та колекцій даних «розумних міст»

Багатовимірні дані – це великі за обсягом набори та колекції даних «розумних міст», які характеризуються великою кількістю ознак або атрибутів, що здебільшого притаманно сфері охорони здоров’я [1]. У ширшому сенсі, для певного зібраного спостереження кількість спостережуваних змінних буде дуже високою. Наприклад, у клінічних даних, якщо записи пацієнтів є розгляданими спостереженнями, то кількість характеристик запису – спостережуваних змінних є особливо великою. Такий тип даних викликає багато занепокоєнь щодо того, чи можна їх ефективно та результативно обробляти і, що найважливіше, анонімізувати, оскільки це залишається відкритим викликом, нині загальновідомим як «прокляття розмірності». Наприклад, було висловлено припущення [79], що для захисту конфіденційності багатовимірних даних необхідно пригнітити велику кількість атрибутів, що негативно впливає на корисність даних і перешкоджає продуктивному аналізу даних.

OLAP-запити – це запити, що подаються до OLAP-кубів даних «розумних міст» і передбачають агрегацію над визначеними цільовими комірками з метою аналізу цих OLAP-кубів [1]. Найчастіше ці запити спрямовуються на спеціальні

сервери, відомі як сховища даних «Data Warehouses», де розташовані OLAP-куби даних. Виділимо важливі типи OLAP-запитів, що називаються діапазонними запитами «range queries», де операція агрегування застосовується до цільових комірок, що містять цільові показники, які охоплені в межах заданих розмежувальних числових діапазонів. Наприклад, сімейство запитів «range-SUM» обчислює суму показників, пов'язаних із комірками в межах зазначених меж. Іншою часто використовуваною операцією є сімейство запитів «range-COUNT», яке підраховує кількість комірок, охоплених зазначеним діапазоном.

3.7.3 Багатовимірна модель аналізу великих даних «розумних міст»

Багатовимірна модель аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» – це схема, що дає змогу проводити багатовимірний аналіз даних, враховуючи структуру базових даних [1]. Значна кількість наукових праць визначила моделі багатовимірного аналітичного опрацювання. Добре визначена аналітична модель розробляється на основі типу запланованого аналізу та сфери застосування. Загальновідомим інструментом, що забезпечує багатовимірне аналітичне опрацювання даних, є OLAP-куби. Дійсно, OLAP-куб дає можливість здійснювати багатовимірне подання даних на етапі попереднього аналізу. Зазвичай OLAP використовується для побудови придатної моделі для цільового аналітичного застосування.

Наприклад, [80] чітко визначають модель багатовимірного ROLAP-куба даних і вказують схеми «зірка» та «сніжинка» для даних про забруднення повітря, щоб проаналізувати їх і в кінцевому підсумку забезпечити прийняття рішень та виявлення знань. Інший приклад описаний у [81], де пропонується модель сховища даних про тероризм. Зокрема, визначена схема «галактика» для полегшення розслідування терористичних актів фахівцями та забезпечення прийняття рішень. Система також дає змогу виконувати MDX-запити до даних про терористичні атаки, щоб отримувати результати залежно від інтуїції слідчого. Крім того, модель аналітичного опрацювання даних, що стосується збереження конфіденційності, описана в [2], яка застосовується до випадку

часових відкритих великих за обсягом наборів та колекцій даних «розумних міст».

3.8 Проблеми та майбутні напрямки досліджень

Шляхом аналізу наукової літератури [1] виникає декілька викликів та майбутніх напрямів досліджень у контексті багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» зі збереженням конфіденційності. Ось найбільш значущі:

1. Масштабні великі багатовимірні дані. Одним із критичних викликів досліджуваної області є те, як зробити методи та алгоритми багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» зі збереженням конфіденційності масштабованими, тоді як обсяг базових сховищ багатовимірних великих за обсягом наборів та колекцій даних «розумних міст» зростає. Це дуже критичний виклик, який заслуговуватиме на велику увагу в найближчі роки.

2. Підтримка конфіденційності великих багатовимірних даних. Коли конфіденційність базових сховищ багатовимірних великих за обсягом наборів та колекцій даних «розумних міст» досягнута, як підтримувати цю конфіденційність в умовах даних, що швидко змінюються? Як широко відомо, типова характеристика багатовимірних великих наборів даних засвідчена тим фактом, що вони за своєю суттю є даними, що еволюціонують, і це може легко вплинути на перевірку фактичних обмежень конфіденційності в наборах даних.

3. Походження «Provenance» багатовимірних великих за обсягом наборів та колекцій даних «розумних міст» стосується виявлення та аналізу джерела та історії «lineage» даних. При застосуванні до багатовимірних великих за обсягом наборів та колекцій даних «розумних міст» ця проблема стає критичною, оскільки ці дані є складними та дуже гетерогенними за своєю природою [1]. Таким чином, походження багатовимірних великих за обсягом наборів та колекцій даних «розумних міст» уособлює складніший випадок для

дослідження, і воно відіграватиме провідну роль у майбутніх дослідженнях великих за обсягом наборів та колекцій даних «розумних міст».

4. Невизначені багатовимірні великі дані. Невизначеність є типовою особливістю багатовимірних великих за обсягом наборів та колекцій даних «розумних міст» через те, що такі дані піддаються помилкам різного характеру, включаючи помилки кодування, помилки передачі, людські помилки тощо.

5. Як методи та алгоритми багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» зі збереженням конфіденційності справляються з такою невизначеністю в даних [82]?

6. Інтеграція з сучасними платформами обробки великих за обсягом наборів та колекцій даних «розумних міст». Стандартизація методів та алгоритмів багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» зі збереженням конфіденційності також залежатиме від їхньої здатності бути інтегрованими в сучасні платформи обробки великих за обсягом наборів та колекцій даних «розумних міст», як от Hadoop і Spark [1].

3.9 Висновок до третього розділу

В третьому розділі кваліфікаційної роботи подано порівняльний опис інструментів та підходів аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». Розглянуто методи аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» за впливом на цілісність. Описано захист та приватність процесів аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст». Досліджено технічні методи маскування та трансформації великих за обсягом наборів та колекцій даних «розумних міст». Проаналізовано спеціалізовані методи опрацювання великих за обсягом наборів та колекцій даних «розумних міст». Розглянуто аналітичні метрики великих за обсягом наборів та колекцій даних «розумних міст». Висвітлена робота з великими даними та середовищами «розумних міст». Описано проблеми та майбутні напрямки досліджень.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1 Вимоги щодо охорони праці при роботі з комп’ютерами. Інструкція для програміста

Дослідження в межах магістерської кваліфікаційної роботи спрямоване на методи багатовимірного опрацювання великих даних розумних міст з врахуванням процесів забезпечення конфіденційності процесів. Оскільки розробка, впровадження та використання програмних продуктів такого масштабу неминуче пов'язані з багатогодинним перебуванням перед комп'ютерним обладнанням, критично важливим стає комплексний аналіз правил охорони праці, які орієнтовані на формування ергономічного простору, мінімізацію фізичного виснаження, запобігання перевтомі зорових аналізаторів та збереження загального тонусу фахівця під час виконання його щоденних службових обов'язків.

Ефективне виконання завдань перед монітором вимагає від розробника перебування в індивідуально налаштованому та конструктивно збалансованому робочому секторі. Дисплей необхідно зафіксувати на такій відстані, щоб вона дорівнювала приблизно півметру або сімдесяти сантиметрам від обличчя, тоді як верхню межу екрана слід орієнтувати точно на рівні лінії погляду чи з легким нахилом донизу, що гарантує анатомічно правильне положення шийного відділу та суттєво знижує напругу очей. Саму площину монітора корисно розвернути під ледь помітним кутом, аби нейтралізувати небажані світлові бліки і тим самим уберегти органи зору від передчасного стомлення. Поверхня робочого столу за своїми габаритами повинна дозволяти вільно та безперешкодно оперувати клавіатурою, мишкою та іншим допоміжним обладнанням без створення захаращеності, а робоче крісло має бути оснащене механізмами зміни висоти та нахилу спинки з обов'язковим анатомічним акцентом на поперeку, завдяки чому підтримується природна постава тіла протягом тривалих сесій. Стопи інженера мають упевнено спиратися на поверхню підлоги або на спеціалізовану похилу

опору, що дозволяє запобігти шкідливому тиску на судини та суглоби нижніх кінцівок [83].

Світлове оформлення робочого приміщення має відповідати критеріям фізіологічного комфорту та повної безпеки, оскільки інтенсивність світлового потоку повинна гарантувати безпроблемне читання текстових символів як на екрані, так і на паперових носіях, не створюючи при цьому засліплюючих спалахів або глибоких тіней, які провокують зоровий дискомфорт. Найбільш вигідним вважається раціональне комбінування природної інсоляції та штучних освітлювальних приладів із прив'язкою до геометрії віконних прорізів та орієнтації монітора, що виключає падіння прямих сонячних променів на робочу зону дисплея. Разом із цим особливу увагу слід приділяти стабільності внутрішнього мікроклімату, де температурні показники, рівень вологості та якість повітрообміну мають суворо вкладатися у встановлені санітарні межі, що уможливлюють тривале перебування за терміналом без проявів фізичного нездужання.

Інтенсивна інтелектуальна діяльність за комп'ютером супроводжується серйозним нервово-психічним та соматичним тиском, який зазвичай трансформується у хронічну втому, падіння швидкості роботи, перенапруження очей або локальні м'язові болі. Через це програмісту необхідно вибудувати свій щоденний графік таким чином, щоб забезпечити організму регулярну можливість для релаксації та повноцінної регенерації сил. Періодична корекція положення тіла, систематичні короткочасні паузи для відпочинку очей та виконання комплексу простих гімнастичних рухів допомагають знімати негативні наслідки тривалої статичної пози. Шляхом гнучкого перемикання між різними видами професійної діяльності впродовж робочої зміни, наприклад, поєднуючи безпосереднє написання коду з аналітичним плануванням чи формуванням супровідної документації, можна нівелювати загрозу монотонного виснаження та зберігати високу продуктивність праці на всіх етапах [84].

Невіддільним елементом збереження професійного довголіття та здоров'я є неухильне дотримання правил особистої та виробничої гігієни. Регулярне очищення робочих поверхонь, дисплея і засобів введення, продумане

розміщення супутнього інструментарію, а також використання прогресивних моніторів із функціями придушення мерехтіння чи фільтрації синього спектра випромінювання допомагають суттєво зменшити навантаження на очі та стабілізувати загальний стан організму. До того ж надважливо грамотно розподіляти робочий тиск, оскільки навіть за умов критичного накопичення завдань чіткий розпорядок дня, своєчасні технологічні перерви, виважена координація робочих процесів та раціональне тайм-менеджмент-планування виступають надійним фундаментом для збереження здоров'я та високої працездатності.

Запропонований звід рекомендацій виконує роль базового нормативного орієнтира, яким інженер-програміст має керуватися під час побудови свого робочого простору. Його головна мета полягає у формуванні таких умов життєдіяльності, за яких багаторічна експлуатація комп'ютерної техніки не завдаватиме шкоди організму, а навпаки, буде максимально ефективною та безпечною для здоров'я. Практична реалізація цих порад є свідченням зрілого та відповідального ставлення до власної професійної місії, що значно підвищує потенціал для успішної, результативної та стабільної діяльності в царині проектування сучасних інформаційних комплексів [85].

Окрему увагу в контексті розробки інтелектуальних міських сервісів слід приділити психологічному комфорту та ментальній гігієні інженера-дослідника. Створення алгоритмів багатовимірної анонімізації в реальному часі для екстремальних обсягів Big Data вимагає безперервної концентрації уваги, що за умов дефіциту часу чи заплутаності кодової бази може призвести до виникнення синдрому професійного вигорання та когнітивного перевантаження. З огляду на це, невіддільною частиною безпеки праці стає створення сприятливого психоемоційного клімату, мінімізація стресових чинників за допомогою чіткого розмежування етапів обробки великих даних, а також використання технік тайм-боксінгу для запобігання надмірній понаднормовій праці. Інтеграція таких превентивних ментальних практик у загальну систему організації праці дозволяє не лише нівелювати загрозу психосоматичних розладів, а й безпосередньо впливає на архітектурну якість створюваного програмного продукту.

4.2 Планування та порядок проведення евакуації населення з районів наслідків впливу НС техногенного та природного характеру

У межах магістерської кваліфікаційної роботи, присвяченої розробці методів багатовимірного опрацювання великих даних «розумних міст», особливе місце посідає моделювання екстремальних сценаріїв, таких як масове переміщення людей. Проектування інформаційних систем для координації логістичних та евакуаційних потоків у мегаполісі вимагає не лише аналізу колосальних масивів інформації з IoT-давачів і відеокамер моніторингу в режимі реального часу, а й суворого забезпечення конфіденційності персональних даних містян, їхніх геолокаційних треків та медичних карт. Створення захищених алгоритмів багатовимірного аналізу дозволяє органам управління «розумних міст» ефективно прогнозувати навантаження на інфраструктуру під час надзвичайних ситуацій, спираючись на чинне законодавство у сфері цивільного захисту.

Організаційна та правова основа для виведення громадян із зон надзвичайних ситуацій природного чи техногенного походження в Україні регламентується Постановою Кабінету Міністрів № 841 від 30.10.2013. Цей нормативно-правовий документ визначає системний порядок евакуаційних заходів, регулює питання тимчасового розміщення та забезпечення життєдіяльності евакуйованих, а також окреслює механізми збереження матеріальних засобів і пам'яток культури за умов виникнення критичних ризиків.

У системі цивільної безпеки виділяють кілька ключових форматів проведення евакуації, вибір яких залежить від специфіки та просторового поширення небезпечного чинника. Загальна евакуація оголошується під час катастроф глобального характеру, зокрема у випадку аварій із викидом радіації чи небезпечних хімічних сполук, а також у ситуаціях раптового катастрофічного затоплення, коли розрахунковий час підходу хвилі внаслідок пошкодження гідротехнічних споруд не перевищує чотирьох годин. Часткова евакуація ініціюється місцевими органами влади та передбачає виведення лише певних груп жителів або звільнення конкретно визначених локацій. Залежно від часових

рамок та можливості повернення, такі заходи класифікують як тимчасове відселення або безповоротне переміщення.

Транспортне забезпечення логістичних процесів під час евакуації передбачає мобілізацію всього наявного рухомого складу, включаючи комунальний транспорт територіальної громади, техніку комерційних структур та автомобілі приватних осіб, якщо виникає безпосередня небезпека для людських життів. Відповідно до правових норм, усім суб'єктам господарювання та громадянам, чий транспорт був використаний для рятувальних робіт, передбачено компенсацію понесених збитків із державних фінансових фондів [86].

Вертикаль управління та запуск евакуаційних процесів розподілені між різними ланками державного апарату, починаючи від найвищого рівня — Кабінету Міністрів, і закінчуючи обласними чи районними адміністраціями. Водночас у форс-мажорних обставинах, коли кожна хвилина має вирішальне значення для порятунку людей, право на негайне виведення населення з епіцентру небезпеки покладається на командира аварійно-рятувального підрозділу, який очолює ліквідацію наслідків інциденту на місці.

Заходи щодо евакуації матеріальних активів та об'єктів культурного фонду розробляються на основі аналітичного прогнозування та моніторингу динаміки розвитку загрози. За наявності достатнього часового резерву вивезення та захист цінностей здійснюються відповідно до галузевих інструкцій, затверджених Міністерством оборони України [87].

4.3 Організація оповіщення пожежної безпеки на підприємствах «розумних міст»

Інтеграція протипожежних комплексів у загальну екосистему «розумних міст» має колосальне значення, адже використання мережі смарт-сенсорів та технологій штучного інтелекту дозволяє локалізувати джерела займання на етапі їх виникнення, суттєво знижуючи рівень загрози для мешканців великих міст. Цифрова автоматизація захисних інженерних систем не лише оптимізує час

виклику та прибуття рятувальних підрозділів, а й виступає гарантом збереження високої працездатності розгалуженої технологічної інфраструктури мегаполісу.

Забезпечення пожежної безпеки є невіддільним юридичним обов'язком будь-якої організації чи підприємства. Відповідні нормативні положення та регламенти мають бути відображені у внутрішній установчій документації, включаючи статути компаній, а також чітко прописані у трудових договорах під час найму персоналу [88].

Безпосередній контроль та персональну відповідальність за дотримання протипожежних вимог покладено на керівний склад організації. Начальник підприємства зобов’язаний розподілити та затвердити посадові обов'язки підлеглих у сфері пожежного захисту, видати офіційні розпорядження про закріплення відповідальних фахівців за конкретними будівлями, територіями чи виробничими цехами, а також постійно контролювати експлуатаційну придатність та працездатність усіх наявних засобів пожежогасіння й автоматики.

Для кожної локації розробляється та вводиться в дію індивідуальний протипожежний режим. Він детально визначає правила облаштування спеціальних зон для куріння, заборону чи суворе обмеження використання відкритого вогню, специфіку організації зварювальних та інших небезпечних робіт, принципи безпечного складування сировини й готової продукції. Крім того, цей режим регулює періодичність очищення вентиляційних каналів від пожежонебезпечних пилових відкладень та визначає чітку послідовність дій колективу при виявленні ознак горіння.

Навчання персоналу правилам безпечної поведінки організоване у формі регулярних навчальних занять та інструктажів. Вступне ознайомлення проходять усі без винятку нові працівники під час оформлення на роботу. Первинний інструктаж організовують безпосередньо на конкретному робочому місці до початку виконання професійних завдань, а повторний проводять систематично за затвердженим графіком для актуалізації та перевірки отриманих раніше знань. Працівники, чия діяльність пов'язана з виконанням робіт високого рівня небезпеки, проходять обов'язкове щорічне тестування та переатестацію.

Для локацій із високою концентрацією відвідувачів, таких як освітні простори, готельні комплекси та медичні центри, обов'язково створюються деталізовані графічні схеми евакуації. Ці плани мають перебувати у зоні вільної видимості, оперативно коригуватися у випадку будь-якої архітектурної перебудови споруди та в обов'язковому порядку перевірятися під час практичних тренувальних евакуацій колективу щонайменше раз на пів року.

Керівництво зобов'язане забезпечити об'єкт функціонуючою мережею гучномовного та централізованого сповіщення, інформаційними плакатами з номерами телефонів рятувальних служб, а також уніфікованими покажчиками напрямку руху та знаками безпеки.

З метою оптимізації захисних заходів на масштабних підприємствах можуть додатково формуватися координаційні та виконавчі органи, зокрема пожежно-технічні комісії, а також залучатися добровільні пожежні команди чи дружини з-поміж штатних співробітників.

Керівник здійснює постійний нагляд за виконанням правил безпеки. У разі виявлення будь-яких відхилень від нормативних стандартів він має негайно організувати роботу з їх ліквідації, а до осіб, які допустили халатність, застосувати заходи дисциплінарного впливу або матеріальні стягнення.

4.4 Висновок до четвертого розділу

В четвертому розділі кваліфікаційної роботи описано вимоги щодо охорони праці при роботі з комп’ютерами. Інструкція для програміста. Розглянуто планування та порядок проведення евакуації населення з районів наслідків впливу НС техногенного та природного характеру. Подано опис процесу організації оповіщення пожежної безпеки на підприємствах «розумних міст».

ВИСНОВКИ

У кваліфікаційній роботі освітнього рівня «Магістр» розв'язано актуальне науково-практичне завдання щодо захисту інформаційних ресурсів мегаполіса шляхом обґрунтування теоретичних засад багатовимірної анонімізації та дослідження методів конфіденційного аналізу міських великих даних в хмарних середовищах. На основі проведеного обчислювального аналізу доведено ефективність запропонованих архітектурних рішень, метрик, а також технічних засобів маскування й трансформації даних, які забезпечують оптимальний баланс між приватністю містян та цілісністю аналітичних процесів «розумних міст».

В першому розділі кваліфікаційної роботи освітнього рівня «Магістр»:

– Подано актуальність досліджень в галузі аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст».

– Розглянуто передумови багатовимірного аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст».

– Висвітлено основи моделей, методів та технік багатовимірного аналітичного опрацювання великих за обсягом даних, що зберігають конфіденційність.

– Проаналізовано хмарні архітектури та платформи для розширеного аналітичного опрацювання, що зберігає конфіденційність.

В другому розділі кваліфікаційної роботи:

– Досліджено конфіденційність високовимірних даних «розумних міст».

– Розглянуто конфіденційність OLAP «розумних міст».

– Проаналізовано конфіденційність великих багатовимірних даних «розумних міст» у нових сценаріях застосування.

– Описано конфіденційність через багатовимірну анонімізацію в системах аналітичного опрацювання великих даних.

– Досліджено архітектури та платформи для розширеного аналітичного опрацювання збереження конфіденційності в хмарах.

В третьому розділі кваліфікаційної роботи:

– Подано порівняльний опис інструментів та підходів аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст»

– Розглянуто методи аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст» за впливом на цілісність.

– Описано захист та приватність процесів аналітичного опрацювання великих за обсягом наборів та колекцій даних «розумних міст».

– Досліджено технічні методи маскування та трансформації великих за обсягом наборів та колекцій даних «розумних міст».

– Проаналізовано спеціалізовані методи опрацювання великих за обсягом наборів та колекцій даних «розумних міст».

– Розглянуто аналітичні метрики великих за обсягом наборів та колекцій даних «розумних міст».

– Висвітлена робота з великими даними та середовищами «розумних міст».

– Описано проблеми та майбутні напрямки досліджень.

У розділі «Охорона праці та безпека в надзвичайних ситуаціях» описано вимоги щодо охорони праці при роботі з комп’ютерами. Інструкція для програміста. Розглянуто планування та порядок проведення евакуації населення з районів наслідків впливу надзвичайних ситуацій техногенного та природного характеру. Подано опис процесу організації оповіщення пожежної безпеки на підприємствах «розумних міст».

ПЕРЕЛІК ДЖЕРЕЛ

1 Cuzzocrea, Alfredo, and Selim Soufargi. "Privacy-preserving multidimensional big data analytics models, methods and techniques: A comprehensive survey." Expert Systems with Applications 270 (2025): 126387.

2 Oehmichen, A., Jain, S., Gadotti, A., & de Montjoye, Y. (2019). OPAL: high performance platform for large-scale privacy-preserving location data analytics. In 2019 IEEE international conference on big data (IEEE bigData), Los Angeles, CA, USA, December 9-12, 2019 (pp. 1332–1342). IEEE.

3 Coronato, A., & Cuzzocrea, A. (2022). An innovative risk assessment methodology for medical information systems. IEEE Transactions on Knowledge and Data Engineering, 34(7), 3095–3110.

4 Commission, E. (2023). Data protection in the EU. https://commission.europa.eu/law/ law-topic/data-protection/data-protection-eu_en.

5 Cuzzocrea, A., Song, I., & Davis, K. C. (2011). Analytics over large-scale multidimensional data: the big data revolution!. In DOLAP 2011, ACM 14th international workshop on data warehousing and OLAP, Glasgow, United kingdom, October 28, 2011, proceedings (pp. 101–104). ACM.

6 Cuzzocrea, A. (2021). Innovative paradigms for supporting privacy-preserving multidimensional big healthcare data management and analytics: The case of the EU H2020 QUALITOP research project. In CEUR workshop proceedings: vol. 3055, Proceedings of the fourth international workshop on semantic web meets health data management, SWH 2021 co-located with the 20th international semantic web conference, ISWC 2021, virtual conference, October 24, 2021 (pp. 1–7). CEUR-WS.org.

7 Zhang, X., Qi, L., Dou, W., He, Q., Leckie, C., Kotagiri, R., et al. (2022). MRMondrian: Scalable multidimensional anonymisation for big data privacy preservation. IEEE Transactions on Big Data, 8(1), 125–139.

8 Dediv, L., Dozorska, O., Kukuruza, V., Nykytyuk, V., Kovalyk, S. Computer Simulation Modeling of Voice Signals in the Matlab Environment for the Task of Computerized Diagnostic Systems Testing. The 1st International Workshop on

“Computer information technologies in Industry 4.0” (CITI-2023) will be held in Ternopil, Ukraine, from June 14 to 16, 2023. The Workshop is organized by the Faculty of Applied Information Technologies and Electrical Engineering of Ternopil Ivan Puluj National Technical University. 2023, 3468, pp. 257–262. Vol-3468 urn:nbn:de:0074-3468-8, ISSN 1613-0073.

9 Cuzzocrea, A., & Bringas, P. G. (2022). CORE-BCD-mAI: A composite framework for representing, querying, and analyzing big clinical data by means of multidimensional AI tools. In Lecture notes in computer science: vol. 13469, Hybrid artificial intelligent systems – 17th international conference, HAIS 2022, Salamanca, Spain, September 5-7, 2022, proceedings (pp. 175–185). Springer.

10 Oleksii Duda, Nataliia Kunanets, Serhii Martsenko, Vyacheslav Nykytyuk, Volodymyr Pasichnyk. Information technology platform for the selection and analytical processing of information on COVID-19. 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT). Volume 2, Lviv, Ukraine 22-25 Sept. 2021. P. 231-328. Electronic ISBN:978-1-6654-4257-2, Print on Demand(PoD) ISBN:978-1-6654-4258-9, Electronic ISSN: 2766-3639, Print on Demand(PoD) ISSN: 2766-3655. DOI: 10.1109/CSIT52700.2021.9648839.

11 Vyacheslav Nykytyuk, Vasil Dozorskyi, Oksana Dozorska, Andrii Karnaukhov and Liubomyr Matiichuk. The Method of User Identification by Speech Signal. The 2nd International Workshop on Information Technologies: Theoretical and Applied Problems (ITTAP-2022) Ternopil, Ukraine, November 22-24, 2022. Vol-3309 urn:nbn:de:0074-3309-1. P.225-232. ISSN 1613-0073 DOI: 10.1425/jsdtl.

12 Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., et al. (1997). Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub totals. Data Mining and Knowledge Discovery, 1(1), 29–53.

13 Guerrero-Prado, J. S., Alfonso-Morales, W., & Bravo, E. F. C. (2021). A data analytics/big data framework for advanced metering infrastructure data. Sensors, 21(16), 5650.

14 Aggarwal, C. C. (2007). On randomization, public information and the curse of dimensionality. In Proceedings of the 23rd international conference on data

engineering, ICDE 2007, the Marmara Hotel, Istanbul, Turkey, April 15-20, 2007 (pp. 136–145). IEEE Computer Society.

15 Fung, B. C., Trojer, T., Hung, P. C., Xiong, L., Al-Hussaeni, K., & Dssouli, R. (2012). Service-oriented architecture for high-dimensional private data mashup. IEEE Transactions on Services Computing, 5(3), 373–386.

16 Agrawal, R., Srikant, R., & Thomas, D. (2005). Privacy preserving OLAP. In Proceedings of the ACM SIGMOD international conference on management of data, Baltimore, Maryland, USA, June 14-16, 2005 (pp. 251–262). ACM.

17 Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional spaces. In Lecture notes in computer science: vol. 1973, Database theory – ICDT 2001, 8th international conference, London, UK, January 4-6, 2001, proceedings (pp. 420–434). Springer.

18 Zakerzadeh, H., Aggarwal, C. C., & Barker, K. (2014). Towards breaking the curse of dimensionality for high-dimensional privacy: An extended version. CoRR abs/1401. 1174.

19 Priebe, T., & Pernul, G. (2000). Towards OLAP security design – survey and research issues. In 3rd ACM international workshop on data warehousing and OLAP (DOLAP 2000), Washington, DC, USA, November 10, 2000 (pp. 33–40). ACM.

20 Cuzzocrea, A., Russo, V., & Saccà, D. (2008). A robust sampling-based framework for privacy preserving OLAP. In Lecture notes in computer science: vol. 5182, Data warehousing and knowledge discovery, 10th international conference, DaWaK 2008, Turin, Italy, September 2-5, 2008, proceedings (pp. 97–114). Springer.

21 Oleksii Duda, Nataliia Kunanets, Serhii Martsenko, Vyacheslav Nykytyuk, Volodymyr Pasichnyk. COVID-19 data collections and analytical processing. 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT). Volume 2, Lviv, Ukraine 22-25 Sept. 2021. P. 252-257. Electronic ISBN:978-1-6654-4257-2, Print on Demand (PoD) ISBN:978-1-6654-4258-9, Electronic ISSN: 2766-3639, Print on Demand (PoD) ISSN: 2766-3655. DOI: 10.1109/CSIT52700.2021.9648839.

22 Wu, G., Wang, S., Ning, Z., & Zhu, B. (2022). Privacy-preserved electronic medical record exchanging and sharing: A blockchain-based smart healthcare system. IEEE Journal of Biomedical and Health Informatics, 26(5), 1917–1927.

23 Monreale, A., Rinzivillo, S., Pratesi, F., Giannotti, F., & Pedreschi, D. (2014). Privacyby-design in big data analytics and social mining. EPJ Data Science, 3(1), 10.

24 Efthymiou, C., & Kalogridis, G. (2010). Smart grid privacy via anonymization of smart metering data. In 2010 IEEE international conference on smart grid communications (pp. 238–243). IEEE.

25 Lin, C., Song, Z., Song, H., Zhou, Y., Wang, Y., & Wu, G. (2016). Differential privacy preserving in big data analytics for connected health. Journal of Medical Systems, 40(4), 97:1–97:9.

26 Zhang, J., Chen, B., Yu, S., & Deng, H. (2019). PEFL: a privacy-enhanced federated learning scheme for big data analytics. In 2019 IEEE global communications conference, GLOBECOM 2019, Waikoloa, HI, USA, December 9-13, 2019 (pp. 1–6). IEEE.

27 Zhang, X., Huang, C., Gu, D., Zhang, J., Xue, J., & Wang, H. (2022). Privacy-preserving statistical analysis over multi-dimensional aggregated data in edge computing-based smart grid systems. Journal of Systems Architecture, 127, Article 102508.

28 Poddar, R., Ananthanarayanan, G., Setty, S. T. V., Volos, S., & Popa, R. A. (2020). Visor: Privacy-preserving video analytics as a cloud service. In 29th USENIX security symposium, USENIX security 2020, August 12-14, 2020 (pp. 1039–1056). USENIX Association.

29 Xu, M., Wang, T., Ding, B., Zhou, J., Hong, C., & Huang, Z. (2019). DPSAaS: Multidimensional data sharing and analytics as services under local differential privacy. Proceedings of the VLDB Endowment, 12(12), 1862–1865.

30 Bondel, G., Garrido, G. M., Baumer, K., & Matthes, F. (2020). The use of deidentification methods for secure and privacy-enhancing big data analytics in cloud environments. In Proceedings of the 22nd international conference on enterprise

information systems, ICEIS 2020, Prague, Czech Republic, May 5-7, 2020, volume 2 (pp. 338–344). SCITEPRESS.

31 Bochicchio, M. A., Cuzzocrea, A., & Vaira, L. (2016). A big data analytics framework for supporting multidimensional mining over big healthcare data. In 15th IEEE international conference on machine learning and applications, ICMLA 2016, Anaheim, CA, USA, December 18-20, 2016 (pp. 508–513). IEEE Computer Society.

32 Tanaglia, M., Ientile, V., L'Abbate, L., Combi, C., Scondotto, S., & Trifirò, G. (2021). Multidimensional design and analysis of a data mart related to healthcare treatments with biologic drugs. In IEEE symposium on computers and communications, ISCC 2021, Athens, Greece, September 5-8, 2021 (pp. 1–7). IEEE.

33 Ghinita, G., Kalnis, P., & Tao, Y. (2011). Anonymous publication of sensitive transactional data. IEEE Transactions on Knowledge and Data Engineering, 23(2), 161–174.

34 Cuthill, E. H., & McKee, J. (1969). Reducing the bandwidth of sparse symmetric matrices. In Proceedings of the 24th national conference, ACM 1969, USA, 1969 (pp. 157–172). ACM.

35 Abdalaal, A., Nergiz, M. E., & Saygin, Y. (2013). Privacy-preserving publishing of opinion polls. Computers & Security, 37, 143–154.

36 Prasser, F., Bild, R., Eicher, J., Spengler, H., Kohlmayer, F., & Kuhn, K. A. (2016). Lightning: Utility-driven anonymization of high-dimensional data. Transactions on Data Privacy, 9(2), 161–185.

37 Cuzzocrea, A., & Saccà, D. (2012). A theoretically-sound accuracy/privacy-constrained framework for computing privacy preserving data cubes in OLAP environments. In Lecture notes in computer science: vol. 7566, On the move to meaningful internet systems: OTM 2012, confederated international conferences: coopIS, DOA-sVI, and ODBASE 2012, Rome, Italy, September 10-14, 2012. proceedings, part II (pp. 527–548). Springer.

38 Cuzzocrea, A., & Bertino, E. (2014). A comprehensive theoretical framework for privacy preserving distributed OLAP. In Lecture notes in computer science: vol. 8842, On the move to meaningful internet systems: OTM 2014 workshops – confederated international workshops: OTM academy, OTM industry case studies

program, C&TC, EI2N, INBAST, ISDE, META4ES, MSC and OnToContent 2014, Amantea, Italy, October 27-31, 2014. proceedings (pp. 117–136). Springer.

39 Kim, S., Lee, H., & Chung, Y. D. (2017). Privacy-preserving data cube for electronic medical records: An experimental evaluation. International Journal of Medical Informatics, 97, 33–42.

40 Dozorskyi, V., Dediv, I., Sverstiuk, S., Nykytyuk, V., Karnaukhov, A. The Method of Commands Identification to Voice Control of the Electric Wheelchair. The Workshop is organized by the Faculty of Applied Information Technologies and Electrical Engineering of Ternopil Ivan Puluj National Technical University. The 1st International Workshop on “Computer information technologies in Industry 4.0” (CITI-2023) will be held in Ternopil, Ukraine, from June 14 to 16, 2023. The Workshop is organized by the Faculty of Applied Information Technologies and Electrical Engineering of Ternopil Ivan Puluj National Technical University. 2023, 3468, pp. 233–240. Vol-3468 urn:nbn:de:0074-3468-8, ISSN 1613-0073.

41 Liu, Z., Cao, Z., Dong, X., Zhao, X., Liu, T., Bao, H., et al. (2022). EPMDA-FED: efficient and privacy-preserving multidimensional data aggregation scheme with fast error detection in smart grid. IEEE Internet Things Journal, 9(9), 6922–6933.

42 Peng, C., Luo, M., Wang, H., Khan, M. K., & He, D. (2022). An efficient privacypreserving aggregation scheme for multidimensional data in IoT. IEEE Internet Things Journal, 9(1), 589–600.

43 Zheng, Y., Lu, R., Zhang, S., Guan, Y., Shao, J., Wang, F., et al. (2022). PMRQ: achieving efficient and privacy-preserving multidimensional range query in ehealthcare. IEEE Internet Things Journal, 9(18), 17468–17479.

44 Liu, H., Gu, T., Shojafar, M., Alazab, M., & Liu, Y. (2023). OPERA: optional dimensional privacy-preserving data aggregation for smart healthcare systems. IEEE Transactions on Industrial Informatics, 19(1), 857–866.

45 Mohammed, N., Fung, B. C. M., Hung, P. C. K., & Lee, C. (2010). Centralized and distributed anonymization for high-dimensional healthcare data. ACM Transactions on Knowledge Discovery from Data, 4(4), 18:1–18:33.

46 Onesimu, J. A., Karthikeyan, J., & Sei, Y. (2021). An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services. Peer-to-Peer Networking and Applications, 14(3), 1629–1649.

47 Lawrance, J. U., & Jesudhasan, J. V. N. (2021). Privacy preserving parallel clustering based anonymization for big data using MapReduce framework. Applied Artificial Intelligence, 35(15), 1587–1620.

48 Onesimu, J. A., Karthikeyan, J., Eunice, J., Pomplun, M., & Dang, H. (2022). Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. IEEE Access, 10, 86979–86997.

49 Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu, A. W. (2006a). Utilitybased anonymization for privacy preservation with less information loss. SIGKDD Explorations, 8(2), 21–30.

50 Li, J., Wong, R. C., Fu, A. W., & Pei, J. (2008). Anonymization by local recoding in data with attribute hierarchical taxonomies. IEEE Transactions on Knowledge and Data Engineering, 20(9), 1181–1194.

51 Al-Zobbi, M., Shahrestani, S. A., & Ruan, C. (2017). Improving MapReduce privacy by implementing multi-dimensional sensitivity-based anonymization. Journal of Big Data, 4, 45.

52 Swarna, C., & Ansari, Z. (2017). Apache pig-a data flow framework based on hadoop map reduce. International Journal of Engineering Trends and Technology (IJETT), 50(5), 271–275.

53 Zala, K., Thakkar, H. K., Jadeja, R., Singh, P., Kotecha, K., & Shukla, M. (2022). PRMS: design and development of patients' E-healthcare records management system for privacy preservation in third party cloud platforms. IEEE Access, 10, 85777–85791.

54 Othman, S. B., Almalki, F. A., Chakraborty, C., & Sakli, H. (2022). Privacypreserving aware data aggregation for IoT-based healthcare with green computing technologies. Computers & Electrical Engineering, 101, Article 108025.

55 Kryazhych O., Itskovych V., Iushchenko K., Hrytsyshyna V., Bruvier D., Nykytyuk V., Bodnarchuk I. (2023) The use of abstract moore automaton to control

the sensors of a service-oriented alarm and emergency notification network. Scientific Journal of TNTU (Tern.), vol 109, no 1, pp. 111–120. ISSN 2522-4433.

56 Zhang, Y., Deng, R. H., Han, G., & Zheng, D. (2018). Secure smart health with privacyaware aggregate authentication and access control in internet of things. Journal of Network and Computer Applications, 123, 89–100.

57 Abbasi, A., & Mohammadi, B. (2022). A clustering-based anonymization approach for privacy-preserving in the healthcare cloud. Concurrency Computations: Practice and Experience, 34(1).

58 Zhang, X., Dou, W., Pei, J., Nepal, S., Yang, C., Liu, C., et al. (2015). Proximityaware local-recoding anonymization with MapReduce for scalable big data privacy preservation in cloud. IEEE Transactions on Computers, 64(8), 2293–2307.

59 Vyacheslav Nykytyuk, Vasyl Dozorskyy, Nataliia Kunanets, Volodymyr Pasichnyk, Oleksandr Matsiuk, Ihor Bodnarchuk: Electrical Probe-Signal Processing and Criterion for the Determination of Time Parameters of the Teeth Filling Material Polymerization Process in Dentistry. 4th IDDM 2021: Valencia, Spain. P. 54-63.

60 ElSalamouny, E., & Palamidessi, C. (2022). Reconstruction of the distribution of sensitive data under free-will privacy. CoRR abs/2208.11268.

61 Jing, W., Miao, Q., Song, H., & Chen, X. (2019). Data loss and reconstruction of location differential privacy protection based on edge computing. IEEE Access, 7, 75890–75900.

62 Navale, G. S., & Mali, S. N. (2019). Lossless and robust privacy preservation of association rules in data sanitization. Cluster Computing, 22(Suppl 1), 1415–1428.

63 Shailaja, G. K., & Rao, C. V. G. (2022). Robust and lossless data privacy preservation: Optimal key based data sanitization. Evolutionary Intelligence, 15(2), 1123–1134.

64 Li, S., & Kot, A. C. (2010). Privacy protection of fingerprint database using lossless data hiding. In Proceedings of the 2010 IEEE international conference on multimedia and expo, ICME 2010, 19-23 July 2010, Singapore (pp. 1293–1298). IEEE Computer Society.

65 Centers for Medicare & Medicaid Services (1996). The health insurance portability and accountability act of 1996 (HIPAA). Online at

https://www.cms.gov/aboutcms/information-systems/privacy/health-insurance-portability-and-accountabilityact-1996.

66 Sweeney, L. (2002). K-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557–570.

67 Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). Ldiversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, 1(1), 3.

68 Balilo, B. B., Gerardo, B. D., & Byun, Y. (2019). CipherBit192: Encryption technique for securing data. In Applied computing and information technology (pp. 137–148). Springer.

69 Muralidhar, K., & Sarathy, R. (2006). Data shuffling – a new masking approach for numerical data. Management Science, 52(5), 658–670.

70 Cadar, C., Akritidis, P., Costa, M., Martin, J.-P., & Castro, M. (2008). Data randomization: Technical report, technical report TR-2008-120, Microsoft Research, 2008.

71 Muralidhar, K., Parsa, R., & Sarathy, R. (1999). A general additive data perturbation method for database security. Management Science, 45(10), 1399–1415.

72 Benkaouz, Y., Erradi, M., & Freisleben, B. (2015). Distributed privacy-preserving data aggregation via anonymization. In Lecture notes in computer science: vol. 9466, Networked systems – third international conference, NETYS 2015, Agadir, Morocco, May 13-15, 2015, revised selected papers (pp. 94–108). Springer.

73 Ardagna, C. A., Cremonini, M., Damiani, E., di Vimercati, S. D. C., & Samarati, P. (2007). Location privacy protection through obfuscation-based techniques. In Lecture notes in computer science: vol. 4602, Data and applications security XXI, 21st annual IFIP WG 11.3 working conference on data and applications security, Redondo Beach, CA, USA, July 8-11, 2007, proceedings (pp. 47–60). Springer.

74 Ren, W., Tong, X., Du, J., Wang, N., Li, S., Min, G., et al. (2021). Privacy enhancing techniques in the internet of things using data anonymisation. Information Systems Frontiers.

75 Abdulkadir, U., Waziri, V. O., Alhassan, J. K., & Ismaila, I. (2022). Ring learning with error-based encryption scheme for the privacy of electronic health records management. In 2022 5th information technology for education and development (pp. 1–5).

76 Hoh, B., & Gruteser, M. (2005). Protecting location privacy through path confusion. In First international conference on security and privacy for emerging areas in communications networks, SecureComm 2005, Athens, Greece, 5-9 September, 2005 (pp. 194–205). IEEE.

77 di Vimercati, S. D. C., Facchinetti, D., Foresti, S., Livraga, G., Oldani, G., Paraboschi, S., et al. (2023). Scalable distributed data anonymization for large datasets. IEEE Transactions on Big Data, 9(3), 818–831.

78 LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006). Mondrian multidimensional K-anonymity. In Proceedings of the 22nd international conference on data engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA (p. 25). IEEE Computer Society.

79 Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In Proceedings of the 31st international conference on very large data bases, Trondheim, Norway, August 30 – September 2, 2005 (pp. 901–909). ACM.

80 Doreswamy, & Harishkumar, K. S. (2018). Multidimensional data model for air pollution data analysis. In 2018 international conference on advances in computing, communications and informatics, ICACCI 2018, Bangalore, India, September 19-22, 2018 (pp. 1684–1689). IEEE.

81 Saidi, F., Trabelsi, Z., & Ghézala, H. B. (2018). Towards a multidimensional model for terrorist attacks analysis and mining. In 28th international conference on computer theory and applications, ICCTA 2018, Alexandria, Egypt, October 30 – November 1, 2018 (pp. 55–59). IEEE.

82 Huang, C., Wang, D., & Chawla, N. V. (2020). Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems. IEEE Transactions on Big Data, 6(4), 702–713.

83 Казюра, А. В., and І. В. Віштак. Ергономічні аспекти організації робочих місць як фактор забезпечення охорони праці та збереження здоров'я

працівників. Diss. Львівський державний університет безпеки життєдіяльності, 2025.

84 Герасимчук, О. В., and О. В. Кобилянський. Вплив тривалої роботи за комп’ютером на здоров’я студентів: шляхи мінімізації ризиків. Diss. ВНТУ, 2025.

85 Гурик, Олег Ярославович, et al. "Навчально-методичний посібник до практичних заняття з дисципліни «Безпека життєдіяльності, основи охорони праці» для студентів освітнього ступеня, бакалавр" усіх спеціальностей та форм навчання." (2025).

86 Методика планування заходів з евакуації: затверджена наказом Міністерства внутрішніх справ України від 10 липня 2017 року № 579. – Київ: Міністерство внутрішніх справ України, 2017. – 50 с.

87 Постанова Кабінету Міністрів України від 30.10.2013 № 841.

88 Баранов, Максим Дмитрович. "Розроблення автоматизованої системи пожежної безпеки на промисловому підприємстві." (2022).

ДОДАТКИ

Додаток А

Тези конференцій

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ

М А Т Е Р І А Л И

XIII НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

«ІНФОРМАЦІЙНІ МОДЕЛІ, СИСТЕМИ ТА ТЕХНОЛОГІЇ»

17-18 грудня 2025 року

ТЕРНОПІЛЬ
2025

ПРОГРАМНИЙ КОМІТЕТ

Голова: Микола Приймак– професор кафедри комп'ютерних систем та мереж, д.т.н., професор;

Співголови: Павло Марущак– докт. техн. наук, професор, проректор з наукової роботи.
Ігор Баран– канд. техн. наук, доцент, декан факультету ФІС.

Науковий секретар: Галина Семенишин– старший викладач.

Члени: докт. фіз.-мат. наук, професор Василь Кривень; докт. техн. наук, професор Ярослав Литвиненко; докт. техн. наук, професор Микола Карпінський; докт. фіз.-мат. наук, професор Михайло Петрик; канд. техн. наук, доцент Галина Осухівська; канд. пед. наук, доцент Жанна Баб'як; канд. техн. наук, доцент Наталія Загородна.

ОРГАНІЗАЦІЙНИЙ КОМІТЕТ

Голова: Юрій Скоренький– канд. фіз.-мат. наук, доцент, завідувач кафедри фізики

Члени: канд. техн. наук, доцент Вячеслав Никитюк; канд. техн. наук, доцент Дмитро Михалик; канд. техн. наук, доцент Марія Стадник; канд. техн. наук, доцент Євгенія Тиш; ст. викладач Ліліана Джиджора.

М34 Матеріали XIII науково-технічної конференції «Інформаційні моделі, системи та технології» Тернопільського національного технічного університету імені Івана Пулюя, (Тернопіль, 17–18 грудня 2025 р.). – Тернопіль : Тернопільський національний технічний університет імені Івана Пулюя, 2025. –162 с.

Адреса оргкомітету: ТНТУ ім. І. Пулюя, м. Тернопіль, вул. Руська, 56, 46001,
тел. (0352) 52-41-33, факс (0352) 254983.
E-mail: conffis2025@gmail.com

Редагування, оформлення та верстка: Галина Семенишин

СЕКЦІЇ КОНФЕРЕНЦІЇ, ЯКІ ПРЕДСТВЛЕНІ В ЗБІРНИКУ

- Математичне моделювання;
- Інформаційні системи та технології;
- Комп'ютерні системи та мережі;
- Програмна інженерія та моделювання складних розподілених систем;
- Новітні фізико-технічні та освітні технології.

В збірнику надруковано тези доповідей XIII науково-технічної конференції «Інформаційні моделі, системи та технології» (Тернопіль, 17–18 грудня 2025 р.) за такими науковими напрямками: математичне моделювання; інформаційні системи та технології; комп'ютерні системи та мережі; програмна інженерія та моделювання складних розподілених систем; новітні фізико-технічні та освітні технології.

Розрахований на науковців, викладачів та студентів вузів.

За зміст тез та дотримання норм академічної доброчесності відповідальність несе автор.

УДК 004.03
І. Вітів; А. Кривецький; М. Боднар
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

БАГАТОВИМІРНЕ АНАЛІТИЧНЕ ОПРАЦЮВАННЯ ВЕЛИКИХ ДАНИХ

UDC 004.03
I. Vitiv; A. Kryvetskyi; M. Bodnar

BIG DATA MULTIDIMENSIONAL ANALYTICAL PROCESSING

Актуальність використання аналітичного опрацювання великих за обсягом наборів та колекцій даних для новітніх фізико-технічних технологій полягає у необхідності швидкої та точної обробки експоненційно зростаючих обсягів даних, що генеруються високочутливими сенсорами, прискорювачами частинок та складними фізичними моделями, для оптимізації процесів та інновацій. Це дає змогу в режимі реального часу виявляти закономірності, прогнозувати поведінку систем, керувати міськими системами тощо. У контексті аналітичного опрацювання великих за обсягом наборів та колекцій даних багатовимірна аналітика великих даних відіграє провідну роль, здебільшого завдяки широкому переліку реальних застосувань, де її можна успішно застосувати. На даний момент часу, дослідники запропонували обширний перелік прикладних застосувань, зокрема [1]:

- великі за обсягом набори та колекції соціальних даних;
- великі за обсягом набори та колекції сенсорних даних;
- великі за обсягом набори та колекції графових даних;
- великі за обсягом набори та колекції енергетичних даних

Внутрішня природа багатовимірних абстракцій надає явну перевагу над класичними, наприклад, SQL-орієнтованими, підходами завдяки використанню розлогого класу аналітичних методологій. В результаті, декілька доменів даних можуть бути легко змодельовані в термінах вимірів, ієрархій, рівнів та мір, і на їхній основі можуть бути побудовані потужні аналітичні інструменти.

Традиційні системи оперативної аналітичної обробки (OLAP) та бізнес-аналітики (BI) помітно розширюють можливості SQL-орієнтованих інструментів аналітичного опрацювання даних. Оскільки добре відомо, що, як правило, набори даних у реальному житті проявляють багатовимірну та багаторівневу природу. Тому багатовимірні аналітичні інструменти більш придатні для видобування корисної інформації з великих за обсягом наборів та колекцій даних, ніж традиційні, безвимірні підходи.

Багатовимірна аналітика великих за обсягом наборів та колекцій даних, що зберігає приватність, є актуальним напрямком досліджень в галузі аналітики великих даних, що зберігає приватність, де фокус зосереджується навколо комбінації методологій багатовимірної аналітики великих даних та парадигм збереження приватності. Цей дослідницький контекст передбачає збереження приватності чутливого діапазону даних при одночасній підтримці завдань багатовимірної аналітики великих за обсягом наборів та колекцій даних. Цей напрямок досліджень активно стимулюється нещодавнім вибуховим розвитком сучасних хмарних інформаційно-технологічних платформ.

Література

1. Ohana, Ruben, et al. "The well: a large-scale collection of diverse physics simulations for machine learning." Advances in Neural Information Processing Systems 37 (2024): 44989-45037.

УДК 004.03
О. Котлінський; І. Вітів; С. Довгалюк
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

РОЗУМНІ МІСТА – КОНЦЕПТИ ТА НАПРЯМКИ ДОСЛІДЖЕНЬ

UDC 004.03
O. Kotlinskyi; I. Vitiv; S. Dovhaliuk

SMART CITIES – CONCEPTS AND RESEARCH DIRECTIONS

Концепція «розумного міста» вперше виникла зосереджується на використанні інформаційних та комунікаційних технологій для покращення інфраструктури та модернізації ресурсних мереж. Широке впровадження інформаційних технологій дало містам можливість покращити безпеку, процеси управлінням та надання послуг. Не існує загальноприйнятого визначення «розумного міста» [1]. Натомість в науковій літературі існує обширний перелік визначень. Окремі визначення зосереджуються на використанні інформаційних та комунікаційних технологій і сучасної інфраструктури, інші підкреслюють людські ресурси та якість життя. Через відсутність єдиного комплексного визначення, термін «розумне місто» нерідко використовується для опису широкого спектру аспектів розвитку, зокрема інформаційних та комунікаційних технологій, освіти та загальної сталості розвитку міст. «Розумне місто» демонструє перспективи підвищення ефективності у шести ключових сферах:

- економіка;
- люди;
- управління;
- мобільність;
- навколишнє середовище;
- життя.

Воно розвивається завдяки стратегічному поєднанню ресурсів та дій, керованих проактивними, незалежними та поінформованими громадянами. «Розумні міста» пропонують значні економічні переваги, зокрема сприяння інноваціям, заохочення підприємництва, створення нових робочих місць та покращення конкурентної позиції міст. Вони також знижують витрати, одночасно підвищуючи ефективність комунальних послуг, виступаючи каталізатором економічного зростання. Стимулюючи швидкий розвиток, «розумні міста» сприяють зростанню ВВП, підвищують рівень зайнятості та залучають іноземні інвестиції – ключові фактори у відродженні міської економіки.

Інформаційно-технологічні проєкти класу «розумне місто» сприяють покращенню стандартів життя, підвищенню конкурентоспроможності міст та подоланню перешкод, як от бідність, соціальна ізоляція чи екологічні проблеми. У сучасних наукових дослідженнях «розумне місто» розглядають з різних точок зору, які охоплюють цифровізацію, мешканців, міську владу, установи та організації, бізнес тощо [2].

Література

1. Al-Msie'deen, Ra'Fat. "Smart city: Definitions, architectures, development life cycle, technologies, application domains, case studies, challenges and opportunities." (2024).
2. Eligüzel, İbrahim Miraç, and Nazmiye Eligüzel. "A Hybrid LDA and Fuzzy CRITIC-MABAC Model for Smart City Ranking." IEEE Access (2025).