

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: Дослідження засобів та методів машинного навчання для аналітичного
опрацювання великих даних

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

(підпис)

Прийма П.В.
(прізвище та ініціали)

Керівник

(підпис)

Кунанець Н.Е.
(прізвище та ініціали)

Нормоконтроль

(підпис)

Мацюк О.В.
(прізвище та ініціали)

Завідувач кафедри

(підпис)

Боднарчук І.О.
(прізвище та ініціали)

Рецензент

(підпис)

Яцишин В.В.
(прізвище та ініціали)

Тернопіль
2023

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)

« 23 » травня 2023 р.

**ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ**

на здобуття освітнього ступеня Магістр
(назва освітнього ступеня)

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

Студенту Приймі Павлу Васильовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження засобів та методів машинного навчання для аналітичного
опрацювання великих даних

Керівник роботи Кунанець Наталія Едуардівна, д.н.с.к., професор кафедри КН
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від « 22 » листопада 2022 року № 4/7-950

2. Термін подання студентом завершеної роботи 22 травня 2023р.

3. Вихідні дані до роботи Наукові публікації про методи та засоби машинного навчання для
аналітичного опрацювання великих даних

4. Зміст роботи (перелік питань, які потрібно розробити)

Вступ. 1 Стан та перспективи досліджень в галузі машинного навчання та опрацювання
великих даних. 2 Дослідження машинного навчання та аналітичного опрацювання великих
даних. 3 Аналіз результатів дослідження машинного навчання та аналітичного опрацювання
великих даних. 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 Застосування IoT. 6 IoT та архітектура великих даних.

7 Розширені характеристики великих даних. 8 Таксономія BDA. 9. Архітектура BDA.

10. Процес систематичного огляду літератури. 11. Ключові слова, які зазвичай
використовуються в дослідженнях BDA. 12 Часова тенденція публікації щодо машинного
навчання для аналітичного опрацювання великих даних. 13. Методи ML, які в основному
використовуються для BDA. 14 Висновки. 15 Завершальний слайд.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Мацюк О.В., доцент	10.04.2023	16.04.2023
Безпека в надзвичайних ситуаціях	Клепчик В.М., ст. викладач	17.04.2023	23.04.2023

7. Дата видачі завдання 14 листопада 2022 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням до кваліфікаційної роботи	14.11.2022-15.11.2022	Виконано
2.	Підбір наукових джерел про засоби та методи машинного навчання, великі дані та їх аналітичне опрацювання	16.11.2022-27.11.2022	Виконано
3.	Переклад та опрацювання наукових джерел про засоби та методи машинного навчання, великі дані та їх аналітичне опрацювання	28.11.2022-25.12.2022	Виконано
4.	Виконання дослідження щодо засобів та методів машинного навчання, аналітичне опрацювання великих даних	09.01.2023-12.03.2023	Виконано
5.	Оформлення розділу «Стан та перспективи досліджень в галузі машинного навчання та опрацювання великих даних»	13.03.2023-19.03.2023	Виконано
6.	Оформлення розділу «Дослідження машинного навчання та аналітичного опрацювання великих даних»	20.03.2023-26.03.2023	Виконано
7.	Оформлення розділу «Аналіз результатів дослідження машинного навчання та аналітичного опрацювання великих даних»	27.03.2023-02.04.2023	Виконано
8.	Виконання завдання до підрозділу «Охорона праці»	10.04.2023-16.04.2023	Виконано
9.	Виконання завдання до підрозділу «Безпека в надзвичайних ситуаціях»	17.04.2023-23.04.2023	Виконано
10.	Оформлення кваліфікаційної роботи	24.04.2023-30.04.2023	Виконано
11.	Нормоконтроль	01.05.2023-07.05.2023	Виконано
12.	Перевірка на плагіат	08.05.2023	Виконано
13.	Попередній захист кваліфікаційної роботи	15.05.2023	Виконано
14.	Захист кваліфікаційної роботи	24.05.2023	

Студент

(підпис)

Прийма П.В.

(прізвище та ініціали)

Керівник роботи

(підпис)

Кунанець Н.Е.

(прізвище та ініціали)

АНОТАЦІЯ

Дослідження засобів та методів машинного навчання для аналітичного опрацювання великих даних // Кваліфікаційна робота освітнього рівня «Магістр» // Прийма Павло Васильович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра комп'ютерних наук, група СНм-61 // Тернопіль, 2023 // С. 70, рис. – 15, табл. – 3, кресл. – 15, додат. – 1, бібліогр. – 75.

Ключові слова: Big Data Analytics, машинне навчання, Великі дані, Hadoop, MapReduce.

Кваліфікаційна робота присвячена розробці засобів та методів машинного навчання для аналітичного опрацювання великих даних. В першому розділі кваліфікаційної роботи освітнього рівня «Магістр» описано розвиток наукових досліджень в галузі аналітичного опрацювання великих даних. В комплексі розглянуто Інтернет речей та аналітичне опрацювання великих даних. Описано інформаційно-технологічні IoT-платформи та аналітичне опрацювання великих даних. Досліджено концепцію великих даних та їх аналітичне опрацювання.

В другому розділі кваліфікаційної роботи досліджено машинне навчання та аналітичне опрацювання великих даних. Описано методи машинного навчання для аналітичного опрацювання великих даних. Висвітлена методика аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних.

В другому розділі кваліфікаційної роботи досліджено машинне навчання та аналітичне опрацювання великих даних. Описано методи машинного навчання для аналітичного опрацювання великих даних.

ANNOTATION

Research of Machine Learning Tools and Methods for Analytical Processing of Big Data // Qualification work of the educational level "Master" // Pryima Pavlo Vasylovych // Ternopil National Technical University named after Ivan Pulyuy, Faculty of Computer Information Systems and Software Engineering, Department of Computer Science, SNnm-61 group // Ternopil, 2021 // P. 70, fig. - 15, tables - 3, chair. - 15, annexes - 1, references. - 75.

Key words: Big Data Analytics, Machine Learning, Big Data, Hadoop, MapReduce.

The qualification work is dedicated to the development of machine learning tools and methods for analytical processing of big data. The first chapter of the qualification work of the Master's level describes the development of scientific research in the field of analytical processing of big data. The Internet of Things and analytical processing of big data are considered in the complex. Information technology IoT platforms and analytical processing of big data are described. The concept of big data and its analytical processing have been studied.

In the second section of the qualification work, machine learning and analytical processing of big data were investigated. Machine learning methods for analytical processing of big data are described. The method of analysis of literary sources regarding means and methods of machine learning for analytical processing of big data is highlighted.

In the second section of the qualification work, machine learning and analytical processing of big data were investigated. Machine learning methods for analytical processing of big data are described.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

AI (англ. Artificial intelligence) – Штучний інтелект.

ANN (англ. Artificial Neural Networks) – Штучні нейронні мережі.

BD (англ. Big Data) – Великі дані.

BDA (англ. Big Data Analytics) – аналітичне опрацювання великих даних.

BI (англ. Business intelligence) – Бізнес-аналітика – позначення комп'ютерних методів та інструментів для організацій, що забезпечують переведення транзакційної ділової інформації в людинно-сприйнятну, а також засобів для масової роботи з такою обробленою інформацією.

ЕВ (англ. ExaByte) – Ексабайт – кратна одиниця вимірювання кількості інформації, що дорівнює 2^{60} стандартним (8-бітним) байтам або 1024 петабайтам.

EL (англ. Ensemble Learning) – Ансамблеве навчання.

ELM (англ. Extreme Learning Machines) – Екстремальне машинне навчання.

DBN (англ. Deep Belief Network) – Мережа глибокої довіри.

HMM (англ. Hidden Markov Model) – Прихована модель Маркова.

KNN (англ. K-Nearest Neighbors) – Метод k-найближчих сусідів.

ML (англ. Machine Learning) – Машинне навчання.

MLP (англ. MultiLayer Perceptron) – Багатошаровий перцептрон.

NB (англ. Naive Bayes) – Наївний Баєс.

SQL (англ. Structured Query Language) – Мова структурованих запитів.

SVD (англ. Singular-Value Decomposition) – Сингулярне розкладання.

SVM (англ. Support Vector Machines) – Метод опорних векторів.

ПК – Персональний комп'ютер.

ЗМІСТ

ВСТУП	7
1 СТАН ТА ПЕРСПЕКТИВИ ДОСЛІДЖЕНЬ В ГАЛУЗІ МАШИННОГО НАВЧАННЯ ТА ОПРАЦЮВАННЯ ВЕЛИКИХ ДАНИХ	9
1.1 Розвиток наукових досліджень в галузі аналітичного опрацювання великих даних	9
1.2 Інтернет речей та аналітичне опрацювання великих даних	11
1.3 Інформаційно-технологічні IoT-платформи та аналітичне опрацювання великих даних	13
1.4 Концепція великих даних та їх аналітичне опрацювання.....	17
1.5 Висновок до першого розділу	25
2 ДОСЛІДЖЕННЯ МАШИННОГО НАВЧАННЯ ТА АНАЛІТИЧНОГО ОПРАЦЮВАННЯ ВЕЛИКИХ ДАНИХ	26
2.1 Машинне навчання та аналітичне опрацювання великих даних	26
2.2 Методи машинного навчання для аналітичного опрацювання великих даних	29
2.3 Методика аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних	30
2.4 Висновок до другого розділу	36
3 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ МАШИННОГО НАВЧАННЯ ТА АНАЛІТИЧНОГО ОПРАЦЮВАННЯ ВЕЛИКИХ ДАНИХ	37
3.1 Інструменти аналітичного опрацювання великих даних	37
3.2 Результати досліджень в галузі аналітичного опрацювання великих даних	39
3.3 Часовий розподіл наукових публікацій щодо аналітичного опрацювання великих даних	43

3.4 Метрики оцінки, що використовуються в галузі аналітичного опрацювання великих даних	49
3.5 Ключові проблеми аналітичного опрацювання великих даних та перспективи майбутніх досліджень.....	51
3.6 Висновок до третього розділу	55
4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ	56
4.1 Медичні профілактичні заходи щодо збереження здоров'я та працездатності користувачів комп'ютерів та відеодисплейних терміналів	56
4.2 Організація оповіщення і зв'язку у надзвичайних ситуаціях техногенного та природного характеру.....	58
ВИСНОВКИ.....	60
ПЕРЕЛІК ДЖЕРЕЛ	62
ДОДАТКИ	

ВСТУП

Актуальність теми. Експоненційне зростання кількості пристроїв з інтегрованими датчиками та виконавчими механізмами, підключеними через Інтернет, спричиняє зростання обсягів великих даних накопичених засобами IoT-пристроїв. Спільна робота людей та машин на основі IoT-пристроїв підвищує операційні ефективність і продуктивність. Аналіз даних, отриманих з IoT-пристроїв покращує процеси прийняття рішень та підвищує якість життя. Еволюція розумного світу стала неминучою. Інтернет речей з'єднує фізичний світ з Інтернетом, який передає критичну інформацію швидше, ніж система, яка залежить від втручання людини. Дані, зібрані IoT-речами, величезні, а великі дані забезпечують швидше й ефективніше зберігання та обробку. Аналітика великих даних використовує інструменти аналізу для опрацювання величезних обсягів даних, продукованих IoT-пристроями, щоб допомогти в прийнятті ефективних рішень. Конвергенція IoT, аналітики великих даних і хмарних обчислювальних технологій формує обширний набір можливостей для дослідження. Зростання доступності цифрових технологій для кожного громадянина робить доступнішим безпрецедентні за обсягами набори даних. Можливість обробляти ці гігантські обсяги даних у режимі реального часу за допомогою інструментів аналітичного опрацювання великих даних (BDA) та алгоритмів машинного навчання (ML) є надзвичайно важливою. Однак велика кількість доступних безкоштовних інструментів BDA, платформ і інструментів інтелектуального аналізу даних ускладнює вибір відповідного інструменту для ефективного та правильного вирішення поставлених завдань. Тому засоби та методи машинного навчання для аналітичного опрацювання великих даних є актуальним напрямком сучасних досліджень.

Мета і задачі дослідження. Метою даної кваліфікаційної роботи освітнього рівня «Магістр» є підвищення рівня повноти подання інформації

щодо машинного навчання та аналітичного опрацювання великих даних. Для досягнення поставленої мети було потрібно виконати наступні завдання:

- Проаналізувати стан досліджень в даній предметній області.
- Дослідити існуючі на даний час методи машинного навчання та їх використання для аналітичного опрацювання великих даних.
- Проаналізувати інструменти аналітичного опрацювання великих даних.

Об’єкт дослідження процесу аналітичного опрацювання великих даних.

Предмет дослідження. методи аналітичного опрацювання великих даних та засоби їх практичної реалізації.

Наукова новизна одержаних результатів кваліфікаційної роботи полягає у тому, що проаналізовано методи машинного навчання. Розглянута практична методика аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних.

Практичне значення одержаних результатів. Сформовано перелік метрик оцінки, що використовуються в галузі аналітичного опрацювання великих даних.

Апробація результатів магістерської роботи. Основні результати проведених досліджень обговорювались на X науково-технічній конфіції «Інформаційні моделі, системи та технології» Тернопільського національного технічного університету імені Івана Пулюя (м. Тернопіль, 2022 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 75 найменувань та 2 додатків. Загальний обсяг кваліфікаційної роботи складає 70 сторінки, з них 45 сторінки основного тексту, який містить 15 рисунків та 3 таблиці.

1 СТАН ТА ПЕРСПЕКТИВИ ДОСЛІДЖЕНЬ В ГАЛУЗІ МАШИННОГО НАВЧАННЯ ТА ОПРАЦЮВАННЯ ВЕЛИКИХ ДАНИХ

1.1 Розвиток наукових досліджень в галузі аналітичного опрацювання великих даних

Нещодавно відбувся бурхливий розвиток наукових досліджень в галузі аналітики великих даних, який поєднує аспекти великих даних, Інтернету речей (IoT) та хмарних інформаційних технологій. Обширне коло дослідників працює над питаннями аналітики великих даних для IoT-пристроїв та систем.

Зокрема в [1] запропоновано трирівневу архітектуру інформаційних систем міського планування для побудови «Розумних міст», яка має обширну множину давачів для «розумних будинків», мереж транспортних засобів, розумних парковок, контролю метеоумов та управління постачанням води. В зазначеній науковій роботі розглянуто збір даних із давачів, попередню обробку, класифікацію та прийняття рішень за допомогою Hadoop із Spark. Автори [2] запропонували використовувати туманну та хмарну системи для IoT-пристроїв у «Розумних будинках». Досліджено поведінкову та прогнозну аналітику процесів споживання енергії. Аналітика в туманних вузлах збільшує можливості керування інтегрованим масивом потоків IoT-даних. Для формування зв'язків між пристроями відбувається одночасне використання пристроїв та проводився частковий аналіз шаблонів.

В [3] розглянута трирівнева архітектура з використанням туманних та хмарних інформаційних технологій із технологіями аналітичного опрацювання великих даних для даних IoT-пристроїв та систем. На туманному рівні відбувається моніторинг процесу хропіння пацієнтів разом з іншими факторами, зокрема, фізичною активністю, середовищем сну, фізіологічними параметрами. Контроль відбувався, щоб виявити виникнення

обструктивного апное сну. Зазначений метод моніторингу та аналітичного опрацювання великих даних надає медичному персоналу інформацію, необхідну для прийняття рішень та лікування пацієнтів. Автори [4] запропонували інформаційну модель на основі туманних обчислень, яка аналізувала дані серцевих характеристик пацієнтів і діагностувала їх стан. Стан пацієнта контролюється за допомогою множини медичних давачів, зокрема давач кисню інтегрований в організмі людини, давач ЕКГ, давач температури тощо. Дані передаються на підключений IoT-пристрій. Пристрої надсилають дані на туманний сервер для обробки. На цьому сервері дані збираються, захищаються, накопичуються та надсилаються на хмарний сервер для подальшої обробки. На хмарному сервері дані попередньо обробляються, фільтруються, стискаються, шифруються. При цьому автоматично формуються рекомендації щодо прийому ліків.

Структура інформаційно-технологічної системи для класифікації медичних даних за допомогою алгоритмів машинного навчання на основі туманних і хмарних обчислень була запропонована в [5]. Зокрема, запропонована п'ятирівнева архітектура, в якій туманні вузли знаходяться ближче до «розумних» IoT-пристроїв і надсилають дані в хмарну інформаційну систему, яка їх обробляє. MapReduce і Apache Spark реалізовані в хмарній системі. Міська система аналітичного опрацювання великих даних у сфері охорони здоров'я була запропонована з використанням даних метеорологічних сайтів, мобільного краудсорсингу та вимірювань засобами Інтернету речей. Дані індикаторів якості повітря надсилаються до периферійних хмарних сервісів та хмарної інфраструктури центру аналітичної обробки даних [6]. Запропонована архітектура на основі великих даних та Інтернету речей для системи моніторингу хронічних пацієнтів, яка забезпечує інтерфейси між пацієнтами, медичними працівниками та особами, які здійснюють догляд громадян. В дослідженні було використано давачі електрокардіографа, пульсоксиметра та тонометра.

Дані з датчиків інтегрованих до смартфонів пацієнтів можуть зберігатися в приватних хмарних сховищах, а потім передаватися в розподілену файлову систему Hadoop і оброблялися засобами Hadoop і Spark [7].

1.2 Інтернет речей та аналітичне опрацювання великих даних

Інтернет речей (IoT) – це мережі фізичних об’єктів в які інтегровано датчики, програмне забезпечення, технології зв’язку та обміну даними з іншими об’єктами в Інтернет [8]. Об’єктом в Інтернеті речей може бути будь-яка сутність, якій можна призначити адресу Інтернет-протоколу (IP) і яка може передавати дані через Інтернет [9]. Спродиковані IoT-пристроями дані надсилаються в хмарне сховище даних, де ними можна ділитися з іншими пристроями та аналізувати (див. рисунок 1.1).

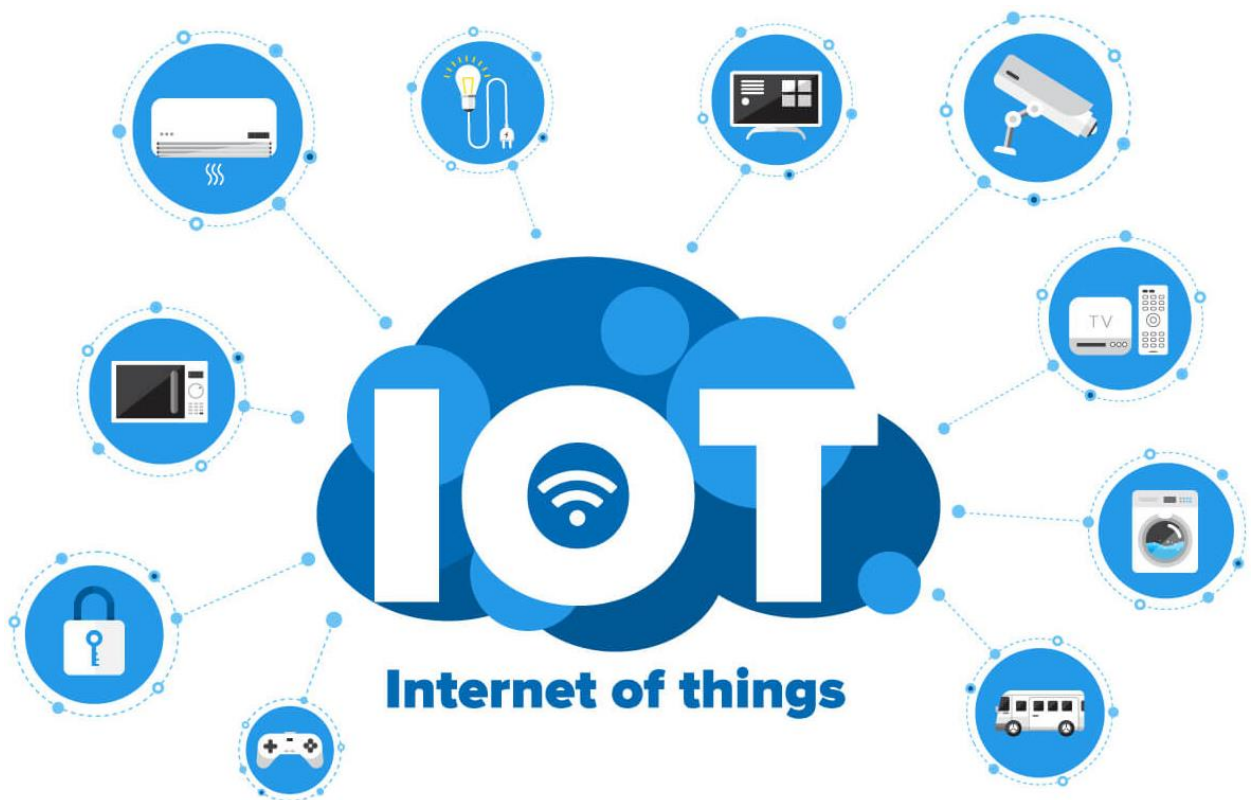


Рисунок 1.1 – Застосування IoT

Інтернет речей набув значного поширення в повсякденному житті і це матиме значний вплив у майбутньому, оскільки отримає повний контроль над якістю життя. Це дозволяє людям жити розумніше завдяки використанню розумних пристроїв для автоматизації будинків, зокрема:

- розумні годинники;
- розумні мобільні телефони;
- розумна побутова техніка;
- розумні дверні замки;
- розумні пристрої безпеки тощо.

Розумні інформаційні технології використовуються в розбудові розумних міст з оптимізованими системами керування дорожнім рухом за допомогою давачів, які автоматично оновлюють інформацію про транспортні потоки, системами управління повітрям, які оновлюють дані про забруднення в режимі реального часу та прогнозують якість повітря, раціоналізованим автоматичним збором сміття та інтелектуальними системами паркування в густонаселених містах [9]. Високоякісні медичні послуги надають переносні монітори стану здоров'я, зокрема «розумний» одяг, «розумні» браслети, медичні носимі прилади, які використовуються для моніторингу частоти серцевого ритму та пульсу, а також для моніторингу стану здоров'я або подання тривоги у разі виникнення потреби невідкладної медичної допомоги [10]. У промисловості IoT покращує швидкість та якість виробництва [11]. Розумне сільське господарство за допомогою зрошувальних систем на основі Інтернету речей використовується для моніторингу якості ґрунту та його збагачення [12]. IoT допомагає впоратися зі стихійними лихами, особливо у випадку лісових пожеж [13]. IoT-пристрої генерують мільярди або трильйони записів даних, які передаються в хмарні сховища для зберігання та обробки в режимі реального часу. Великі за обсягами набори згенерованих даних аналізуються та виділяються корисні інформаційні шаблони за допомогою технологій Big Data.

1.3 Інформаційно-технологічні IoT-платформи та аналітичне опрацювання великих даних

Інтернет речей (IoT) охоплює обширний перелік розумних пристроїв, підключених до мереж за допомогою бездротових технологій. Це робить інформаційно-технологічну структуру доволі складною. На даний час узгодженої стандартної інформаційно-технологічної IoT-архітектури. Базова архітектура IOT – це трирівнева архітектура [14], яка складається з:

- рівня сенсорики;
- мережевого рівня;
- рівня застосунків.

На рівні сприйняття відбувається збір інформації та даних про середовище від давачів і виконавчих механізмів [15]. Давачі можуть перетворювати інформацію в дані, а приводи виконавчих механізмів можуть приймати рішення та виконувати дії на основі отриманих даних. Мережевий рівень дозволяє підключатися до множини різнотипових мереж, серверів і «розумних» пристроїв. Прикладний рівень надає користувачам набір послуг, зокрема, увімкнення та вимкнення пристроїв, смарт-годинників, розумних систем безпеки тощо [16]. П'ятирівнева архітектура (див. рисунок 1.2) складається з:

- сенсорного рівня;
- мережевого рівня;
- рівня обробки;
- рівня застосунків;
- бізнес-рівня.

Мережевий рівень передає дані від рівня сприйняття до рівня обробки через бездротові мережі. Рівень обробки є проміжним рівнем програмного забезпечення. Він отримує з транспортного рівня, зберігає, аналізує та обробляє дані.



Рисунок 1.2 – Узагальнена архітектура IOT-систем

Бізнес-рівень передбачає аналіз результатів на основі бізнес-моделей. Водночас він здійснює подання даних у виді графіків, діаграм.

Дані з IoT-пристроїв збираються за допомогою давачів та приводів. На рисунку 1.3 подано IoT та архітектуру Big Data. Пристрої повинні мати один компонент для зв'язку за допомогою бездротових або дротових мереж. Використовуються [17]:

- поширені протоколи зв'язку – wifi, zigbee, z-wave, LoRaWAN, bluetooth;
- протоколи стільникових мереж – cat-m, nb-iot;
- протоколи цифрового відео – 4G LTE і 5G.

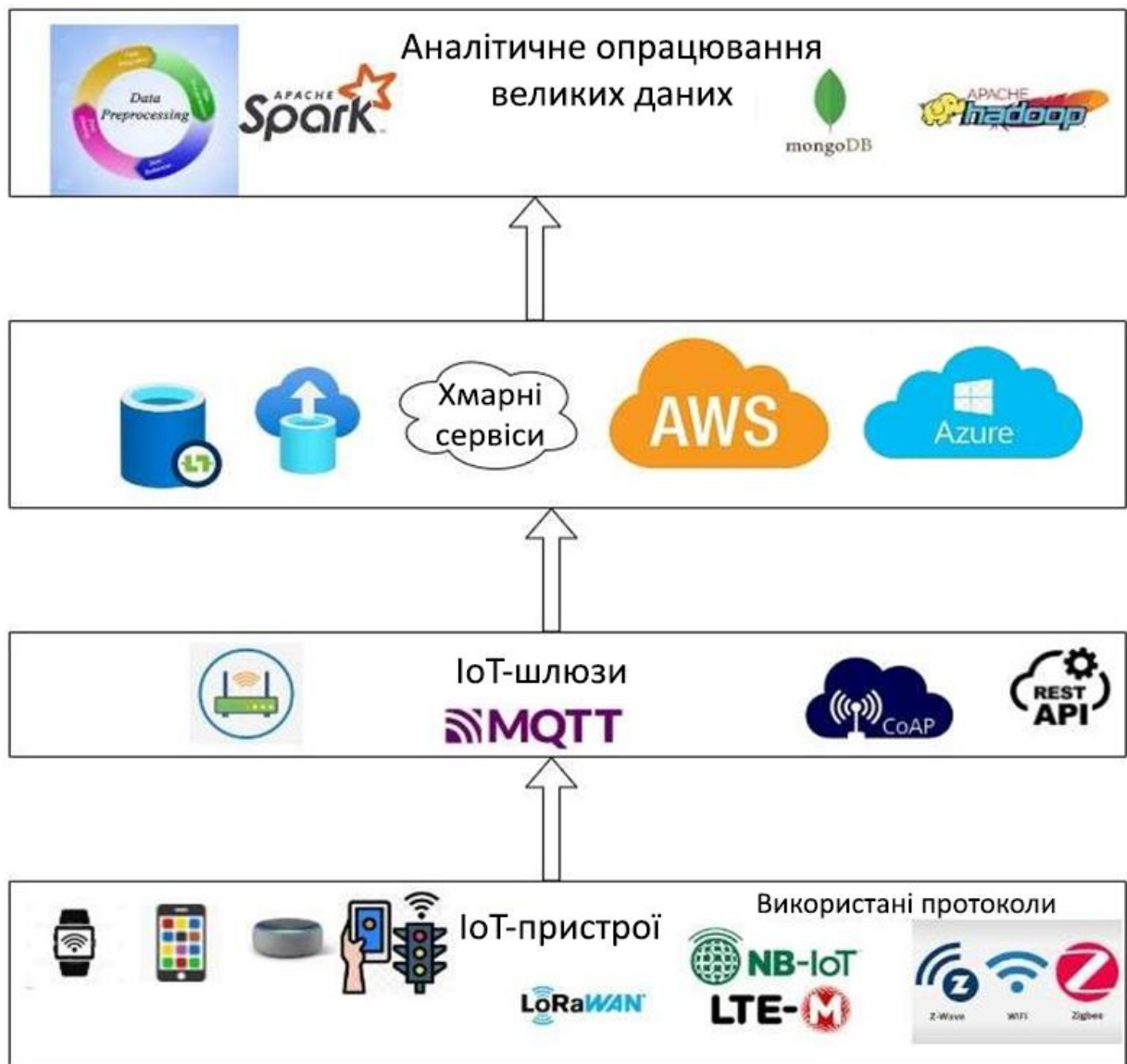


Рисунок 1.3 – IoT та архітектура великих даних

Коли IoT-пристрій не може запустити Інтернет-протокол безпосередньо, повідомлення передається на інший пристрій – шлюз, який обробляє та пересилає повідомлення в Інтернет. Для цього використовуються протоколи шлюзів [18]:

- MQTT;
- REST;
- CoAP.

Потім дані надсилаються в хмару для зберігання, обробки та доступу. Хмарні програми працюють на платформах AWS, Azure тощо.

Згенеровані великі дані мають інший формат через їх розмір, швидкість, з якою вони накопичуються, характер і якість даних. Великі дані зберігаються в озерах даних, які підтримують різні типи даних. Потім дані попередньо обробляються за допомогою інструментів інтелектуального аналізу даних і програмного забезпечення для підготовки даних, що включає очищення даних, інтеграцію, вибір і перетворення даних. Попередньо оброблені дані потім піддаються аналітичній обробці даних за допомогою методів аналізу даних для отримання знань, пошуку закономірностей і прийняття рішень.

Різні типи даних, продуковані IoT-пристроями, аналізуються за допомогою інструментів аналізу даних. Для IoT-даних можна використовувати різні типи аналітичних інструментів, зокрема:

- описовий;
- діагностичний;
- прогнозний;
- прескриптивний аналіз.

Описовий аналіз – це найпоширеніший і найпростіший метод, який отримує значущу інформацію з величезних даних, знаходячи відповіді на такі запитання, як «що», «де», «скільки» і «хто», і його часто подають у формі кругових діаграм, стовпчастих діаграм, таблиць тощо.

Діагностичний аналіз встановлює залежності та знаходить закономірності, відповідаючи на запитання, наприклад, «чому щось сталося?». Він також надає детальну інформацію.

Прогнозний аналіз – це техніка, яка використовує результати описової та діагностичної аналітики, відповідає на запитання про те, що ймовірно станеться, щоб знайти кластери, винятки та передбачити майбутні тенденції.

Прескриптивний аналіз – це розширена техніка, яка визначає, які дії потрібно виконати за допомогою методів машинного навчання.

Зростання IoT стало неминучим, оскільки люди віддають перевагу розумним будівлям, розумним кампусам і розумним містам. Інтернет речей змінив багато сфер, головним чином сектор охорони здоров'я, сільське господарство, промисловість, виробництво тощо. Дані Інтернету речей стали важливим джерелом великих даних, які використовуються для дослідження та аналізу даних, покращення процесу прийняття рішень і, зрештою, підвищення якості життя.

1.4 Концепція великих даних та їх аналітичне опрацювання

Згідно з [19], *«BD – це збір даних у величезних обсягах завдяки досягненням у технологічних інструментах і платформах, які підтримують високошвидкісний збір, зберігання та аналіз даних»*. Концепція BD, наведена Дугом Лейні, цитована в Ref. [20] брендуються за обсягом, швидкістю та різноманітністю, визнаним як 3Vs.

Великі дані – це набір великих за обсягом різноманітних наборів даних, який з часом зростає в геометричній прогресії, що ускладнює ефективне зберігання та обробку традиційними інструментами керування даними. Він може бути:

- структурованим;
- неструктурованим;
- напівструктурованим

Обсяги наборів даних вимірюються в діапазоні від терабайтів до зетабайтів.

Великі дані можна охарактеризувати чотирма властивостями (англ. V) [13] (див. рисунок 1.4):

- обсягом;
- швидкістю;
- різноманітністю;

– достовірністю.

Обсяг описує розмір даних, який надзвичайно зростає. Наприклад, передбачуваний глобальний мобільний трафік становив 6,2 EB (ексабайт) у 2016 році та 65 EB до кінця 2021 року, а до кінця 2027 року він досягне 288 EB. Швидкість означає швидкість, з якою дані накопичуються з джерел даних, як машини, мережі, мобільні пристрої та платформи соціальних мереж.



Рисунок 1.4 – Характеристики великих даних

Користувачі Whatsapp надсилають понад 100 мільярдів повідомлень на день, а в Google здійснюється 3,5 мільярда пошукових запитів на день. Різновид вказує на характер даних, які можуть бути структурованими, неструктурованими та напівструктурованими. Дані можуть бути структуровані у:

- Google-таблицях;
- файлах Excel;
- SQL-даних.

Неструктуровані у:

- текстових даних;
- повідомлень соціальних мереж;
- зображень;
- аудіо- та відеоповідомлень.

Напівструктуровані у файлах HTML, JSON або XML. Достовірність характеризує якість даних, на яку впливають непослідовність і невизначеність даних [21]. Дані з низькою достовірністю містять безглузді та шумні дані, тоді як дані з високою достовірністю є цінними та важливими для загальних результатів.

Однак більшість досліджень [22] розширюють концепцію Дага Лейні до п'яти ключових характеристик (5V), а саме (див. рисунок 1.5):

- об'єм;
- швидкість;
- різноманітність;
- значення;
- достовірність.



Рисунок 1.5 – Розширені характеристики великих даних

Визначення BD постійно змінюється відповідно до прогресу в технології, ємності для зберігання даних, швидкості передачі даних та інших можливостей інформаційних систем [20]. Перша «V» – обсяг, позначає розмір даних, який експоненціально збільшується з часом. Стверджується, що галузь охорони здоров'я генерує великі обсяги даних в електронних медичних записах порівняно з більшістю галузей [22]. Друге «V» – швидкість, означає швидкість, з якою дані генеруються та отримуються з різних галузей. Третє «V» – різноманітність, позначає множинність і неоднорідність даних. На думку деяких дослідників, четверте значення «V» є найважливішою та незамінною характеристикою всіх 5V BD, оскільки воно здатне перетворювати галузеві дані на частину цінної інформації. П'яте «V» – достовірність, дуже схоже на гарантію якості даних. Це дає певну достовірність знань у певному секторі.

Аналіз даних – це процес, який використовує методи для аналізу наборів даних і виявлення закономірностей і цінної інформації для підвищення ефективності процесів та прийняття рішень [23]. BDA передбачає використання передових аналітичних інструментів і методів для великих за обсягом структурованих, напівструктурованих або неструктурованих наборів даних. BDA передбачає збір даних із різних джерел, обробку, очищення та аналіз даних. BDA використовується при аналізі великих за обсягом наборів даних отриманих з різних джерел, покращенні елементів ланцюгів поставок, фінансових операціях завдяки підвищенню ефективності бізнес-процесів, знанні поведінки споживачів, аналізі настроїв та управлінні ризиками [24].

Згідно [25], BDA застосовує передові аналітичні методи та техніки для великих за обсягом наборів даних. Подібним чином, BDA можна визначити як процес збору, систематизації та ретельного вивчення BD для прогнозування, пошуку шаблонів та виявлення знань разом з іншою інформацією в BD [26]. BDA практично включає дві речі, великі дані та

аналітичне опрацювання, і те, як вони об'єдналися, щоб створити одну з прогресивних сучасних тенденцій у бізнес-аналітиці (BI). BDA [26] складається з (див. рисунок 1.6):

- описової аналітики;
- прогнозової аналітики;
- прескриптивної аналітики.

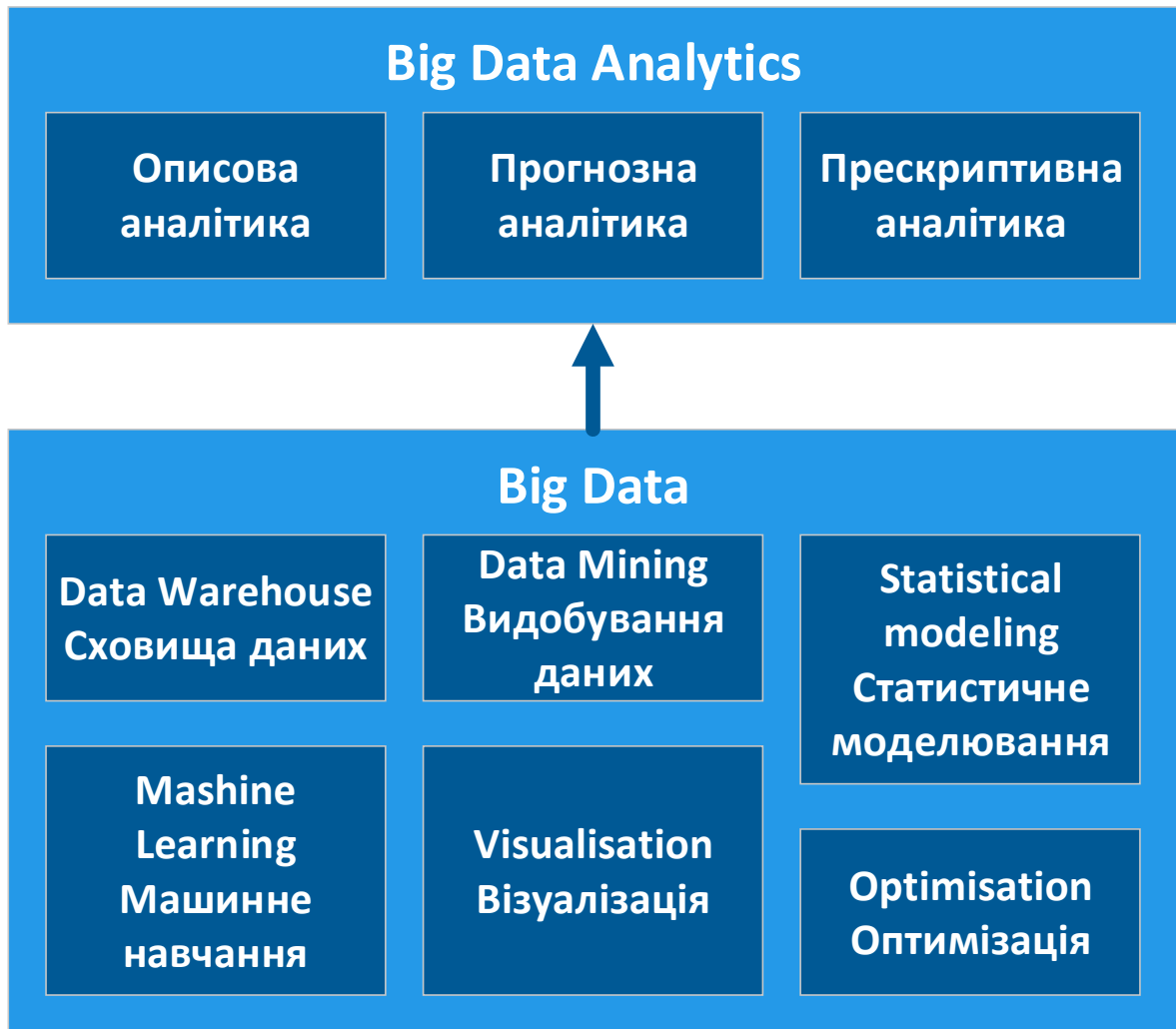


Рисунок 1.6 – Таксономія BDA

Таким чином, BDA використовує методи аналізу даних, щоб виявити закономірності в журналах викликів, мобільних банківських транзакціях і створеному користувачами онлайн-контенті. Основні правила BDA

сформовані на основі інженерії знань, математики, людинно-машинних інтерфейсів, статистики, інформаційних технологій та інформатики.

Наразі BDA – це нова технологія великих даних, яка набула поширення у різних сферах людської діяльності, компаніях, географічних регіонах, а також серед окремих осіб, щоб допомогти їм приймати керовані даними рішення для досягнення поставлених цілей [27]. Впродовж останнього часу BDA можна використовувати за допомогою декількох аналітичних платформ та множини інструментів. Зокрема, що базуються на:

- SQL-запитах;
- кластеризації фактів;
- інтелектуальному аналізу даних;
- статистичному аналізу;
- методах обробки природної мови;
- візуалізації даних;
- AI;
- ML;
- текстовій аналітиці;
- MongoDB;
- Hadoop;
- MapReduce.

Слід відмітити, що доступні для обробки 5V великих даних платформи та інструменти значно покращилися впродовж останнього періоду часу. Загалом ці технології та інструменти не є надто дорогими, і значна частина доступного програмного забезпечення поширюється з відкритим кодом. На рисунку 1.7 показана базова прикладна теоретична архітектура BDA [20].

Алгоритми ML домінують в аналізі, візуалізації та моделюванні великих даних. ML дозволяє машинам навчатися на основі наборів даних у базовому визначенні, застосовувати знання та розуміння прихованих даних і робити прогнози. ML можна класифікувати на чотири класи, а саме:

- контрольоване навчання;
- неконтрольоване навчання;
- напівконтрольоване навчання;
- навчання з підкріпленням.

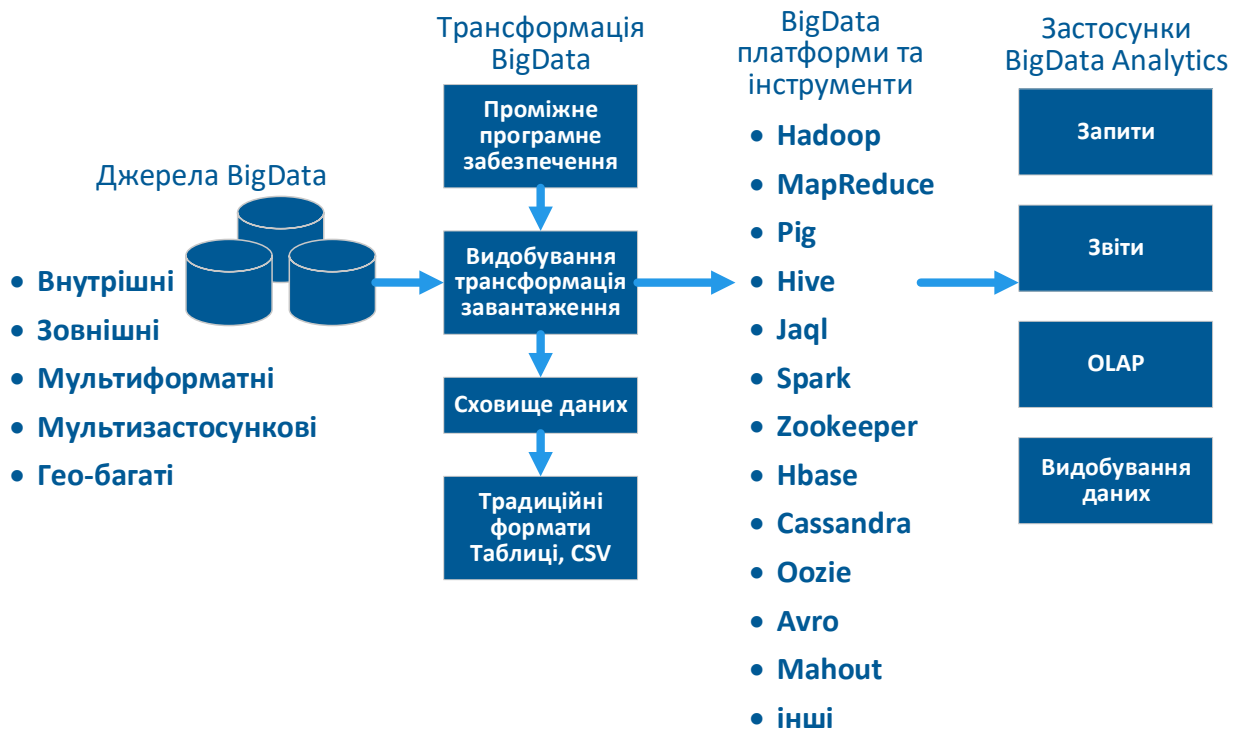


Рисунок 1.7 – Теоретична архітектура BDA

Класи алгоритмів ML:

- класифікація;
- регресія;
- кластеризація;
- зменшення розмірності;
- ранжування.

Зокрема, алгоритми:

- метод опорних векторів (SVM);
- штучні нейронні мережі (ANN);
- наївний Баєс (NB);
- Tensor Auto-Encoder (TAE);

- ансамблеве навчання (EL);
- метод k-найближчих сусідів (KNN);
- прихована модель Маркова (HMM);
- сингулярне розкладання (SVD);
- радіальна нейронна мережа базової функції (RBF-NN);
- аналіз головних компонентів (PCA);
- генеративні змагальні мережі (GAN);
- обробка природної мови (NLP);
- рекурентні нейронні мережі (RNN);
- двонаправлена стробована рекурентна сутність (Bi-GRU);
- узагальнений дискримінантний аналіз (GDA);
- мережа Deep Q (DQN);
- нейронна мережа загальної регресії (GRNN);
- нейронна мережа прямого зв'язку (FNN);
- довготривала короткочасна пам'ять (LSTM);
- мережа глибокого автокодування (DAN);
- багатошаровий перцептрон (MLP);
- екстремальне машинне навчання (ELM);
- мережа глибокої довіри (DBN).

У науковій літературі повідомляється про успішне використання алгоритмів ML у різних сферах застосування, зокрема:

- Фінанси та фондовий ринок.
- Прогнозування енергетичних систем та виявлення несправностей.
- Охорона здоров'я.
- Викладання і навчання.
- Сільське господарство (врожайність, викиди та виявлення хвороб).
- Транспорт.
- Петрологія.

1.5 Висновок до першого розділу

В першому розділі кваліфікаційної роботи освітнього рівня «Магістр» описано розвиток наукових досліджень в галузі аналітичного опрацювання великих даних. В комплексі розглянуто Інтернет речей та аналітичне опрацювання великих даних. Описано інформаційно-технологічні IoT-платформи та аналітичне опрацювання великих даних. Досліджено концепцію великих даних та їх аналітичне опрацювання.

2 ДОСЛІДЖЕННЯ МАШИННОГО НАВЧАННЯ ТА АНАЛІТИЧНОГО ОПРАЦЮВАННЯ ВЕЛИКИХ ДАНИХ

2.1 Машинне навчання та аналітичне опрацювання великих даних

Великі обсяги даних генеруються щодня в різних сферах людської діяльності, зокрема, соціальних мереж, інженерії, комерції, біомолекулярних досліджень, фізіології тощо [28]. Обсяги цифрових даних, створені з різних цифрових платформ і пристроїв, у всьому світі зростають неймовірними темпами. Станом на 16 грудня 2020 року обсяг щоденних даних у всьому світі становив 59 зетабайт. Очікується, що у 2024 році він досягне 149 зетабайт [29], оскільки ми переходимо у майбутнє, яке ще більше керуватиметься даними.

Збільшення обсягів даних є основною ознакою «великих даних», терміну, який став загальноприйнятною назвою у дослідницьких спільнотах, організаціях та Інтернеті.

Нещодавно Великі дані (BD) і новітні методи та засоби, зокрема Big Data Analytics (BDA), змінили спосіб функціонування організацій і підприємств, створивши нові важливі перспективи для підприємств, фахівців і наукових кіл [26]. Окрім підприємств і дослідницьких установ, урядові та неурядові організації регулярно генерують великі за обсягом та складні унікальні набори даних [30]. Тому отримання важливої інформації із цих доступних великих даних стало критично важливим для організацій у всьому світі. Аналіз літературних джерел свідчить, що швидко та ефективно отримувати корисну інформацію з BD досить складно [19].

Незважаючи на те, що більшість алгоритмів штучного інтелекту (AI) і машинного навчання (ML) і їхні платформи для виконання BDA є безкоштовними, вони вимагають нового набору навичок, який є незвичайним для більшості практиків у цій галузі та IT-відділів організацій[31]. Отже,

інтеграція цих інструментів і платформ у внутрішні та зовнішні дані організації на загальній платформі є складним завданням.

Крім того, наявність кількох алгоритмів машинного навчання створює труднощі для правильного вибору з них, тобто «пошуку голки в стозі сіна». Таким чином, необхідно провести всебічний порівняльний аналіз BDA в різних галузях з алгоритмами ML. Крім того, залежно від джерела даних або галузі, великі дані мають інший формат:

- структурований;
- напівструктурований;
- неструктурований.

Автор [22] довів, що алгоритми ML працюють по-різному в залежності від формату вхідних даних. Зокрема, алгоритм ML може використовуватись з високою точністю до структурованого набору даних, ніж до напівструктурованого або неструктурованого набору даних. У [22] було продемонстровано що алгоритм ML може працювати по-різному в різних завданнях ML. Наприклад, той самий алгоритм, здатний виконувати завдання регресії або класифікації, може працювати краще в класифікації, ніж у регресії.

Дослідження в даній кваліфікаційній роботі призначене для:

- Допомоги дослідникам, IT-відділам та фахівцям оцінити правильні інструменти й алгоритми BDA під час аналізу великих даних.
- Новим дослідникам BDA прийняти обґрунтоване рішення та зробити корисний внесок у наукову спільноту.
- Результати слугуватимуть керівництвом для вдосконалення методів та інструментів, які поєднують великі дані та когнітивні обчислення.

Дослідження потрібно провести за напрямками:

- Детальна оцінка стану сучасних досліджень BDA з методами ML.
- Розгляд елементів таксономії BDA, зокрема, обсягів даних, походження досліджень, завдань та методів ML, показників оцінки.

– Стисле представлення важливих особливостей порівнюваних методів BDA та ML.

– Аналіз потенційних викликів, тенденцій досліджень та можливостей для перспективних досліджень BDA.

Обсяги даних, експоненціально зростають завдяки активному розвитку інформаційних технологій. Вибухове зростання накопичуваних обсягів даних спричинено [32]:

– Збільшенням кількості інтелектуальних пристроїв, що генерують дані, із давачами та виконавчими механізмами, які підключені глобально через хмару.

– Збільшенням кількості користувачів Інтернету.

– Зростанням процесів інтеграції елементів віртуальної реальності та доповненої реальності.

– Стільниковим зв'язком 5G-мереж.

– Збільшенням транзакцій електронної комерції тощо.

З 2017 по 2022 рік глобальний трафік Інтернет-протоколу зріс на 26% [33]. У 2022 році кількість підключених до Інтернету пристроїв у три рази перевищила населення планети. Смартфони становлять 44% від загального IP-трафіку. Обсяги продукованих цифровим світом даних досягнуть бронтобайту в найближче десятиліття.

Ми живемо в еру великих за обсягом даних. При цьому інформація є новою нематеріальною цінністю. Вибухове зростання обсягів даних створило проблеми з їх збором, зберіганням, пошуком, обробкою та представленням через обсяг, різноманітність і швидкість даних. Виявлення цінності або корисних шаблонів у великих даних створює багато проблем, які потребують величезних обчислювальних ресурсів, технологічної інфраструктури та кваліфікованих аналітиків даних. Інтернет речей (IoT) – це одна з технологічних революцій цієї епохи, що створює серйозний виклик у здатності використовувати великі обсяги даних. IoT і Big Data – це дві

незалежні інформаційні технології. При цьому IoT генеруватиме великі обсяги даних, а Big data підвищуватиме ефективність зберігання та обробки даних. В процесі досліджень, проведених в кваліфікаційній роботі освітньо-наукового рівня «Магістр» доцільно висвітлити зв'язок між IoT, як джерелами даних, ML та Big Data Analytics. При цьому слід проаналізувати дослідницьку перспективу в цій галузі. Зазначені зв'язки потрібно проаналізувати з точки зору перетворення величезних даних у більш зрозумілі та значущі моделі даних та знань. Водночас потрібно проаналізувати інформаційно-технологічні платформи, доступні для Big Data Analytics IoT-даних.

2.2 Методи машинного навчання для аналітичного опрацювання великих даних

Величезна кількість наукових публікацій в галузі BDA ускладнює практикам і дослідникам не відставати від розробок у цій галузі. Тому деякі опубліковані результати досліджень намагалися узагальнити різні застосунки ML у BDA та їх вдосконалення, щоб допомогти початківцям вибрати правильний алгоритм ML для BDA.

Однак обширне коло досліджень було звужено до BDA в конкретних галузях, зокрема [34]:

- охорона здоров'я;
- якість навколишнього середовища;
- Інтернет речей (IoT);
- сільське господарство;
- інформаційна безпека.

Автори [35] зосереджені на огляді, викликів і підходів до BDA. У [36] подано аналізу ризику BD. На відміну від цього, в [37] описано BDA у різних сферах, але зосереджено увагу на ефективності моделей та вартості

обчислень. Крім того, робота [38] надала огляд змісту, обсягу та результатів аналітичного опрацювання великих даних, а також обговорила її майбутню еволюцію.

Згідно з опублікованими даними, було проведено відносно небагато заходів для вдосконалення машинного навчання та аналітичного опрацювання великих даних у різних галузях, зокрема:

- охороні здоров'я;
- сільському господарстві;
- енергетиці;
- машинобудуванню тощо.

Жодні з опублікованих результатів досліджень не розглядали зазначені проблеми в достатній мірі. Однак великі дані спричиняють збурення в кожному економічному секторі. Отже, буде несправедливо звужувати аналітичне опрацювання великих даних в одній або декількох галузях. Крім того, під час проведеного аналізу опублікованих результатів досліджень було виявлено недоліки:

- Процес відбору паперу в деяких документах.
- Більшість досліджень не враховували походження документів.
- Не продемонстровано чітких статистичних даних щодо застосування платформ BDA та інструментів моделювання.
- Жодне з проаналізованих результатів досліджень не розглядало цілі документів.

2.3 Методика аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних

Згідно [39] методика аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних прагне швидко та легко продемонструвати конкретну проблему чи набір

пов'язаних тем, а також підкреслити, де є прогалини в науковій літературі та можливості для подальших досліджень. Методика аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних повинна бути простіша, ніж повноцінний огляд літератури. Це тому, що, на відміну від огляду літератури, який зосереджується на синтезі результатів кількох досліджень для розробки висновків щодо широкої галузі дослідження, методика аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних зосереджується на одному предметі чи темі. Важливо пам'ятати, що огляд літератури та формат методики аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних однакові. Так само немає суттєвої різниці між етапами методики аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних та оглядом літератури [39]. Однак єдина різниця між ними полягає в тому, що один підхід ширший, а інший вужчий. Методика аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних ефективніша, оскільки її стислий формат дозволяє легко проаналізувати ці теми в науковій літературі, дозволяючи більшій кількості практиків отримувати ефективні результати. На рисунку 2.1 подано процес опрацювання літературних джерел [40], який дотримується п'яти вказівок, зокрема:

- стратегія пошуку;
- критерії відбору;
- процес відбору дослідження;
- забезпечення якості;
- якісний і кількісний аналіз.

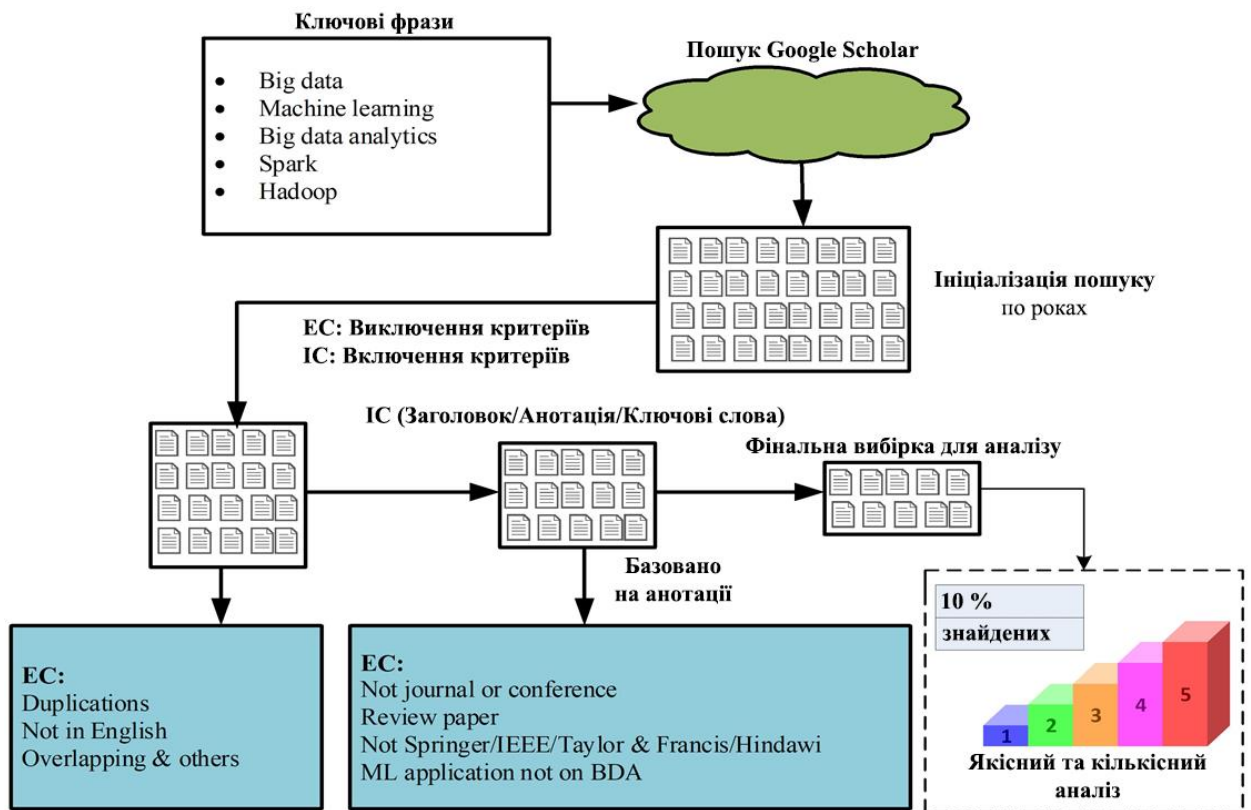


Рисунок 2.1 – Процес систематичного огляду літератури

Розглянемо детально, що досягається на кожному етапі. Google Scholar було прийнято як центральну платформу пошукової системи для збору релевантних статей завдяки відкритому доступу та гнучкості обмеження дати публікацій. Доцільно завантажувати лише відповідні статті журналів і конференцій [40]. У пошуку слід використовувати п'ять основних фраз, зокрема, «великі дані», «машинне навчання», «аналітика великих даних», «Apache Spark» і «Hadoop», поданих англійською мовою. Однак потрібно отримувати декілька пов'язаних запитів до п'яти ключових фраз, використовуючи тенденції Google.

Наведемо множину пошукових фраз, які були прийняті в дослідженнях як допоміжні слова, зокрема «великі дані та аналітика даних», «аналітика великих даних», «бізнес-аналітика», «бізнес-аналітика великих даних», «аналітика великих даних», «data analytics», «analytics», «Hadoop», «Hadoop spark», «Hadoop hive», «big data Hadoop», «deep learning», «deep machine learning», «spark Apache tutorial», «Scala» [40].

Опубліковані результати досліджень розглядають аналітику великих даних із застосуванням машинного навчання в різних сферах людської діяльності. Понад півтори тисячі наукових публікацій у журналах і на конференціях, було завантажено авторами [40] за допомогою пошуку за ключовими словами в Google Scholar. Завантажені документи перевірялися на кілька етапів, а остаточно відібрані статті переглядалися на основі запропонованої таксономії. На основі проведеного аналізу можна передбачити інструменти, які найчастіше використовуються для аналітичного опрацювання великих даних, тенденції та перспективні напрямки дослідницької роботи протягом. Робота [40] допоможе дослідникам і фахівцям галузі отримати цінну базу для подальших досліджень, щоб зрозуміти повний контекст аналітичного опрацювання великих даних із машинним навчанням і його застосування в різних галузях.

На рисунку 2.3 показано тенденції аналітики великих даних від Google Trends [40].

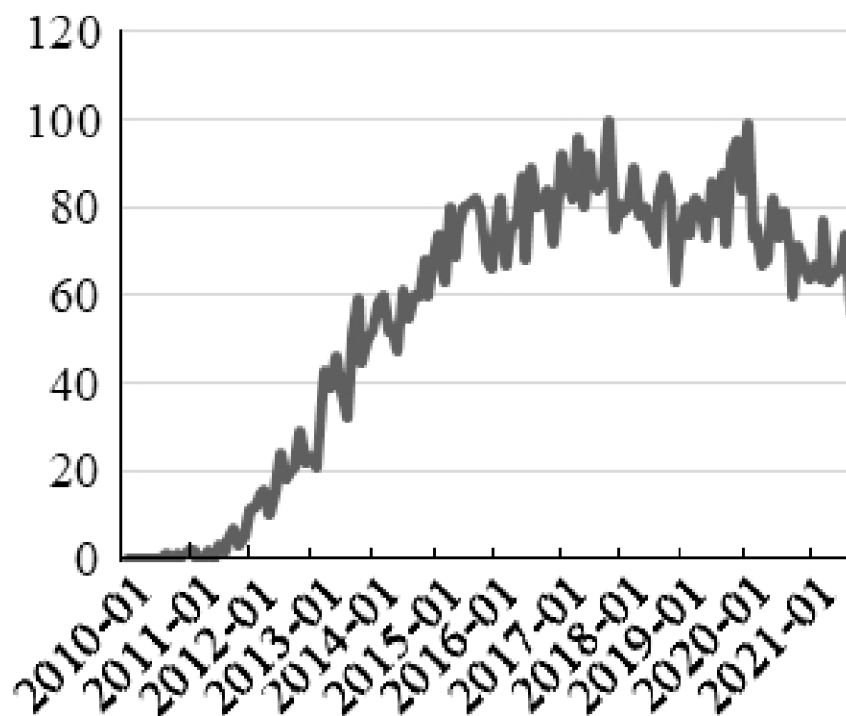


Рисунок 2.3 – Тенденція пошукових запитів Google щодо аналітичного опрацювання великих даних [40]

Статті доцільно розглядати на основі узгоджених критеріїв включення-виключення за авторами. Критерії включення:

- стаття написана англійською мовою;
- стаття має стосуватися великих даних та DBA;
- стаття опублікована між 2010–2022 роками;
- стаття має бути опублікована у журналі або на конференції.

Критерії виключення:

- статті, не опубліковані протягом 2010–2021 років;
- статті, опубліковані не в журналі чи на конференції.

Крім того з якісного та кількісного аналізу були виключені:

- рецензійні статті про BDA;
- статті, не опубліковані в Elsevier, Springer, Taylor & Francis, IEEE та Hindawi;
- ML-застосунки, у яких набір даних і інструменти не підпадають під концепцію великих даних.

Спочатку всі статті, що стосуються засобів та методів машинного навчання для аналітичного опрацювання великих даних, доцільно ретельно відібрати на етапі первинного скринінгу. На підставі поданих критеріїв виключення та включення, потрібно перевірити завантажені статті (див. рис. 2.2). Невідповідні статті, тобто статті, не опубліковані англійською мовою, дублікати, роботи, що перекриваються, доцільно виключити. Далі потрібно перевірити решту на основі назви, анотації, видавництва та типу публікації, а статті, які не були пов'язані із дослідженням, слід відкинути. Нарешті, потрібно відфільтрувати статті на основі анотацій за допомогою логічного оператора «І» за всіма визначеними термінами пошуку на останньому етапі скринінгу.

При проведенні оцінки якості (QA) робіт після аналізу та оцінювання тез вибраних робіт. Деякі з критеріїв забезпечення якості:

- Чи зрозуміла мета дослідника статті?

- Чи ефективно застосовується методологія?
- Чи безперечно пояснюються результати?
- Чи існує зв'язок між вступом, результатами та висновком?

У результаті комплексної перевірки описаної в [41] було відібрано сто сорок статей, що стосуються досліджуваної галузі, із півтора тисяч завантажених спочатку статей. Із ста сорока статей понад сімдесят проаналізовано якісно, понад шістдесят проаналізовано кількісно. Оцінка якості відіграє суттєву роль у процедурі систематичного огляду літератури.

2.4 Висновок до другого розділу

В другому розділі кваліфікаційної роботи досліджено машинне навчання та аналітичне опрацювання великих даних. Описано методи машинного навчання для аналітичного опрацювання великих даних. Висвітлена методика аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних.

3 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ МАШИННОГО НАВЧАННЯ ТА АНАЛІТИЧНОГО ОПРАЦЮВАННЯ ВЕЛИКИХ ДАНИХ

3.1 Інструменти аналітичного опрацювання великих даних

Інструменти аналітичного опрацювання великих даних використовуються для обробки масивних структурованих, неструктурованих і напівструктурованих даних та отримання знань, бізнес-прогнозування, ефективного прийняття рішень і візуалізації шаблонів. Розглянемо множину популярних інструментів.

Apache Hadoop з відкритим кодом і є одним із найпопулярніших інструментів аналізу даних. HDFS (розподілена файлова система Hadoop) – це компонент зберігання даних, який використовується для зберігання різноманітних даних, зокрема текст, файли xml, json, аудіо-файли, зображення та відео, шляхом поділу даних на частини та збереження в кластерах окремих серверів [42]. Він базується на Java і має високу швидкість, оскільки завдання розділені та виконуються одночасно на розподілених серверах. Оскільки дані зберігаються на багатьох розподілених серверах, резервне копіювання даних доступне, навіть якщо один сервер вийде з ладу. Він масштабований і економічно ефективний.

Apache Spark – це розподілена система з відкритим кодом, яка обробляє дані за допомогою апаратної оперативної пам'яті [43]. Швидкість обробки даних Spark в сотні разів перевищує швидкість Hadoop. Spark зручний для розробників, оскільки для створення програм можна використовувати різні мови, зокрема:

- Java;
- Python;
- R;

–Scala тощо.

Багато організацій, зокрема «Finra», «Yelp», «Zillow», «gumgum», використовують spark. Він став одним із найпопулярніших фреймворків розподіленої обробки великих даних.

MongoDB – це база даних з відкритим вихідним кодом, орієнтована на документи NoSQL, яка сумісна з багатьма інформаційними технологіями, платформами та мовами програмування [44], зокрема:

- Python;
- Ruby;
- JavaScript.

Він зберігає дані в документах json, які відповідають будь-якому типу мови програмування. Деякі з його функцій, наприклад динамічна схема, динамічні запити, багаті оновлення та легке агрегування, роблять його потужним інструментом для аналітичного опрацювання великих даних. Організації, які використовують MongoDB;

- Facebook;
- eBay;
- Google тощо.

Apache Cassandra – це розподілена СУБД NoSQL з відкритим вихідним кодом, яка використовується для обробки даних на різнотипових серверах [45], вона:

- може обробляти великі за обсягом набори та колекції даних;
- підтримує прості транзакції;
- не має єдиної точки відмови;
- має розподілене зберігання даних;
- підтримує масштабованість;
- розгортається в горизонтальному масштабі.

Використовує CQL (структуровану мову cassandra) для зв'язку з базою даних. Cassandra використовують:

- Facebook;
- Honeywell;
- Yahoo;
- Accenture тощо.

Tableau – це програмне забезпечення, яке використовується для візуалізації та дослідження великих даних, яке ефективно поєднує дані з широкого переліку джерел [46]. Основною перевагою Tableau є автоматизована звітність, яка створює звіти без коду, її легко налаштувати, співпрацю в режимі реального часу та просту інтеграцію. Організації, які використовують Tableau – це Verizon, ZS associates тощо.

3.2 Результати досліджень в галузі аналітичного опрацювання великих даних

Поданий в [40] огляд наукових літературних джерел показує, що поточні дослідження BDA можна класифікувати за п'ятьма різними темами, а саме:

- основна область BDA для обробки масштабу;
- управління шумом і нечіткістю даних;
- аспекти конфіденційності та безпеки;
- інженерія даних;
- стик BDA і науки про дані.

На основі онтології, запропонованої в [26], кластеризовано тип засобів аналітичного опрацювання великих даних на три групи, зокрема:

- описова аналітика BDA (група «А»);
- прогнозна аналітика BDA (група «В»);
- BDA та прескриптивна аналітика (група «С»).

Було помічено, що шістдесят відсотків розглянутих документів [47] базувалися на прогнозній аналітиці BDA, двадцять відсотків – на

прескриптивній аналітиці BDA [48], десять відсотків – на описовій аналітиці BDA [47], сім відсотків – A+B [49] і три відсотки – B+C [50]. Декілька досліджень BDA використовували прескриптивну аналітику – це можна пояснити тим фактом, що прескриптивна аналітика великих даних знаходиться на ранній стадії. Однак автор [51] стверджував, що розробка алгоритмів злиття інформації для об'єднання керованого людиною навчання, і керованого давачами навчання, є шляхом до прескриптивної аналітики великих даних. Кліланд [52] стверджував, що швидка еволюція BDA змінює декілька секторів людської діяльності, зокрема, галузі охорони здоров'я та медицини. Галузь охорони здоров'я – це сфера, де щоденно генеруються масивні історичні дані з різних джерел, зокрема, відповідність документів і нормативні вимоги, ведення записів і догляд за пацієнтами [53].

Тому не дивно побачити такий обширний перелік опублікованих результатів досліджень в царині охорони здоров'я – 30%, за якою слідують виявлення аномалій – 11%, кібербезпека, конфіденційність даних та Інтернет речей – 5%, а також автомобільний транспорт і транспорт – 5%. Результати аналізу опублікованих відомостей щодо досліджень в царині аналітичного опрацювання великих даних подані в таблиці 3.1.

Таблиця 3.1 – Результати аналізу опублікованих відомостей щодо досліджень в царині аналітичного опрацювання великих даних

Область використання	Застосування	Обсяг даних
1	2	3
Катастрофи	Запропоновано модель прогнозування шкоди від сильного дощу за допомогою ML та BD	528500 точок даних
Енергетика	Прогнозований офшорний ВЕС на основі алгоритмів BD і ML	396000 спостережень

Продовження таблиці 3.1

1	2	3
Охорона здоров'я	Прогнозована хвороба в охороні здоров'я за допомогою ML і BD	20320848 записів
Електронна комерція	Прогнозований попит споживачів на продукт за допомогою BD	35200 спостережень
Оперативне реагування	Оптимізоване виявлення аномалій за допомогою BD	710 MB
Кібербезпека	Виявлено та класифіковано шкідливі пакети команд і відповідей у мережі SCADA	64100 сутності
Набір даних про погоду	Запропоновано вдосконалення змінної відстані для оцінки потоку BD часових рядів	34435268 сутності

Обсяги даних проаналізовані в опублікованих відомостях щодо досліджень в царині аналітичного опрацювання великих даних подані в таблиці 3.2.

Таблиця 3.2 – Результати аналізу опублікованих відомостей щодо досліджень в царині аналітичного опрацювання великих даних

Мета	Обсяг даних
1	2
Системи й мережі прогнозованого стільникового трафіку	300 мільйонів записів
Економіка, мережева безпека NLP і дистрибутив із збереженням ядра SVM	2405500 спостережень
Прогнозований обсяг виробництва електроенергії ML та BD	NA
Прогнозована затримка поїзда за допомогою BDA	NA

Продовження таблиці 3.2

1	2
Модель виявлення обличчя на основі швидкої мережі глибокої згортки	600000 прикладів
BDA для комплексної оцінки кредитного ризику	2284 записів
Методика каталогізації та узагальнення фільмів	53000 записів
Чотиривимірна (4D) модель представлення на основі слів, використовуючи рівень взаємодії з відповідями на питання та рівень гіпервзаємодії	NA
Відкриття запропонованої відповіді на дискусійних форумах за допомогою ML і BDA	NA
Підвищена продуктивність вбудованої архітектури розумного міста на основі BDA	3.1 GB
Всестороння експериментальна оцінка продуктивності між фреймворками Spark і MapReduce	600 GB

Примітка: NA – не вказано, тобто автори не вказали галузь, з якої надійшли їхні дані.

Деякі опубліковані результати досліджень вказали, що розмір даних з точки зору простору для зберігання коливається від 708 МБ [54] до 600 ГБ [48], тоді як з точки зору кількості спостережень, він коливається від тисячу сімсот вісімдесят [55] до трьох мільярдів записів [56]. Виходячи з обсягів даних, можна стверджувати, що дослідження [54] не має відношення до великих даних. Однак ми вважаємо, що розмір даних, який використовується лише в дослідженні, не класифікує його як дослідження великих даних. Водночас інструменти та платформи, що використовуються для емпіричного аналізу, також мають значення.

3.3 Часовий розподіл наукових публікацій щодо аналітичного опрацювання великих даних

На рисунку 3.1 показано часову тенденцію публікацій щодо машинного навчання для аналітичного опрацювання великих даних [40]. По горизонталі подано роки, а по вертикалі показано кількість опублікованих наукових статей.

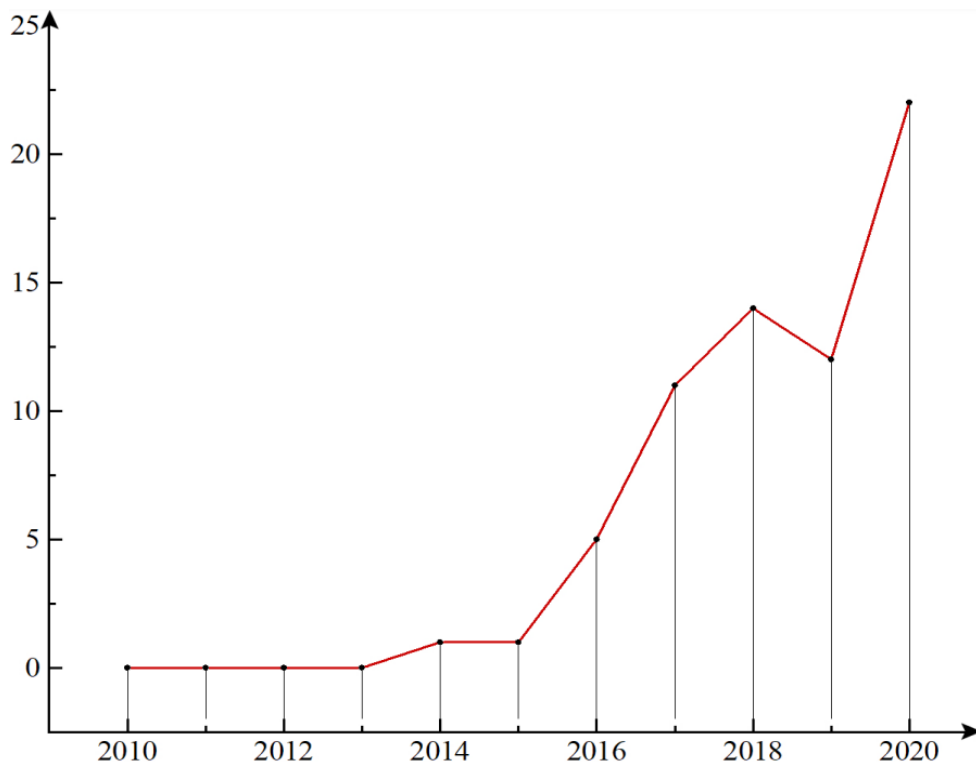


Рисунок 3.1 – Часова тенденція публікації щодо машинного навчання для аналітичного опрацювання великих даних

Незважаючи на те, що огляд обмежив пошук літератури між 2010–2020 рр., помічено, що дослідницька робота щодо застосування машинного навчання в BDA почала привертати увагу фахівців та дослідників протягом останніх п'яти років і з тих пір поступово зростає [20]. Незважаючи на це, BDA існує вже порівняно давно, проте в останній період часу воно лише зайняло належне місце, яке стало популярним. Крім того, можна зробити

висновок, що суттєве зростання BD у поточні роки [29] привернуло увагу дослідників до вивчення переваг, які можна ефективно отримати від доступності даних для прийняття науково-обґрунтованих рішень.

На рисунку 3.2 показано розподіл видавців проаналізованих наукових статей.

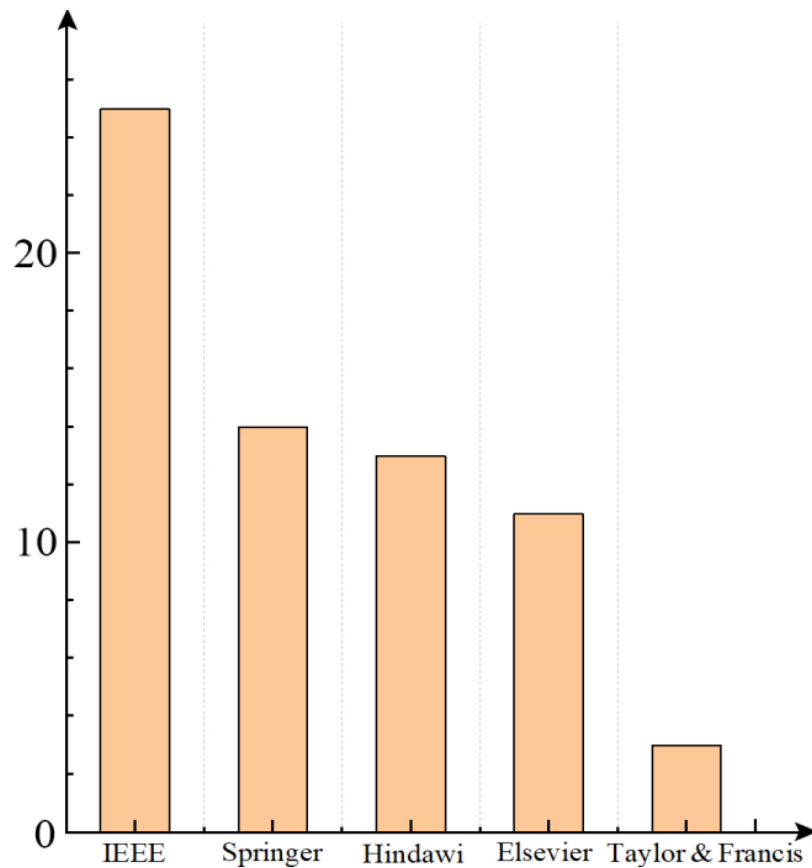


Рисунок 3.2 – Видавництва, що публікують матеріали щодо машинного навчання для аналітичного опрацювання великих даних

З семидесяти статей сорок відсотків були опубліковані в IEEE, по майже двадцять відсотків у Springer та Hindawi, а сімнадцять та три відсотки статей були опубліковані в Elsevier і Taylor & Francis відповідно. Результати свідчать про те, що відомі видавництва побачили потребу надати науковому співтовариству аналітику великих даних із програмами машинного навчання.

4.2 Інструменти платформ великих даних для аналітичного опрацювання

У таблиці 3.3 показано популярні інформаційно-технологічні платформи та інструменти DBA.

Таблиця 3.3 – Платформи та інструменти BDA

Платформи та інструменти BD	Кількість документів	Відсоток (%)
Flink	1	2.33
Apache Mahout	1	2.33
HiBench	1	2.33
H2O	1	2.33
MATLAB	5	11.63
MapReduce	6	13.95
Apache Hadoop	13	30.23
Apache Spark	15	34.88

З семидесяти статей шістдесят п'ять відсотків вказали інструменти BDA, використані для експериментального аналізу. Незважаючи на те, що Hadoop вважається найпотужнішим і найпопулярнішим інструментом для BDA [19], результати оцінювання свідчать протилежне. Слід відмітити, що Spark є найбільш використовуваним інструментом (35%) для DBA серед дослідників у цій галузі, за яким слідує Hadoop (30%). Цей висновок можна пояснити тим, що Spark є швидшим і легшим у використанні для аналізу великих даних, ніж Hadoop MapReduce. Крім того, фахівці вважають, що Spark забезпечує більшу швидкість обробки, ніж Hadoop [48].

Водночас, Hadoop не пропонує конвеєрну передачу даних, і його непросто використовувати. Наприклад, Гу та Лі [57] відзначили, що він не

підходить для повторюваних операцій через суттєву вартість перезавантаження даних диска при кожному повторенні. Аналогічно, автори [58] повідомили, що Spark є ефективнішим при обробці великих обсягів даних, ніж Hadoop. Отже, ці фактори можуть сприяти його (Spark) поширенню серед дослідників у сфері аналітичного опрацювання великих даних.

З іншого боку, порівняльне дослідження показує, що Spark споживає більше пам'яті під час роботи, ніж Hadoop, оскільки він завантажує всі процеси в пам'ять і деякий час зберігає їх у кешах [48]. Тому доцільно здійснювати вибір між цими двома платформами, спираючись на їхні характеристики. Наприклад, вартість, простота використання, обмеження пам'яті, відмовостійкість, рівень продуктивності, тип обробки даних і безпека вказують на відповідність поточному проекту та поточним і перспективним потребам організації. Підводячи підсумок, можна сказати, що з відкритим вихідним кодом фреймворків Spark і Apache ринок значно розширився, оскільки все більше фірм і дослідників оптимальні найкращі способи для впровадження цих платформ.

На рисунку 3.3 показано методи ML, які в основному використовуються для BDA. Отримані результати свідчать про те, що штучний інтелект і ML продовжуватимуть розвиватися, оскільки все більше компаній і галузей прагнуть трансформувати свій повсякденний бізнес, максимізуючи прибуток і мінімізуючи ризики. Встановлено, що SVM є найпоширенішим алгоритмом машинного навчання при аналітиці великих даних через його здатність працювати:

- практично без попереднього знання набору даних;
- з високою розмірністю та ризиком надмірного обтікання.

Алгоритм дерева рішень (DT) хоч і простий, але є третім найбільш використовуваним алгоритмом ML у BDA. Відповідно до [55], DT легко зрозуміти. Крім того, вони перевіряють модель за допомогою статистичних

тестів, зокрема, ентропія або приріст інформації, що сприяє її популярності в BDA.

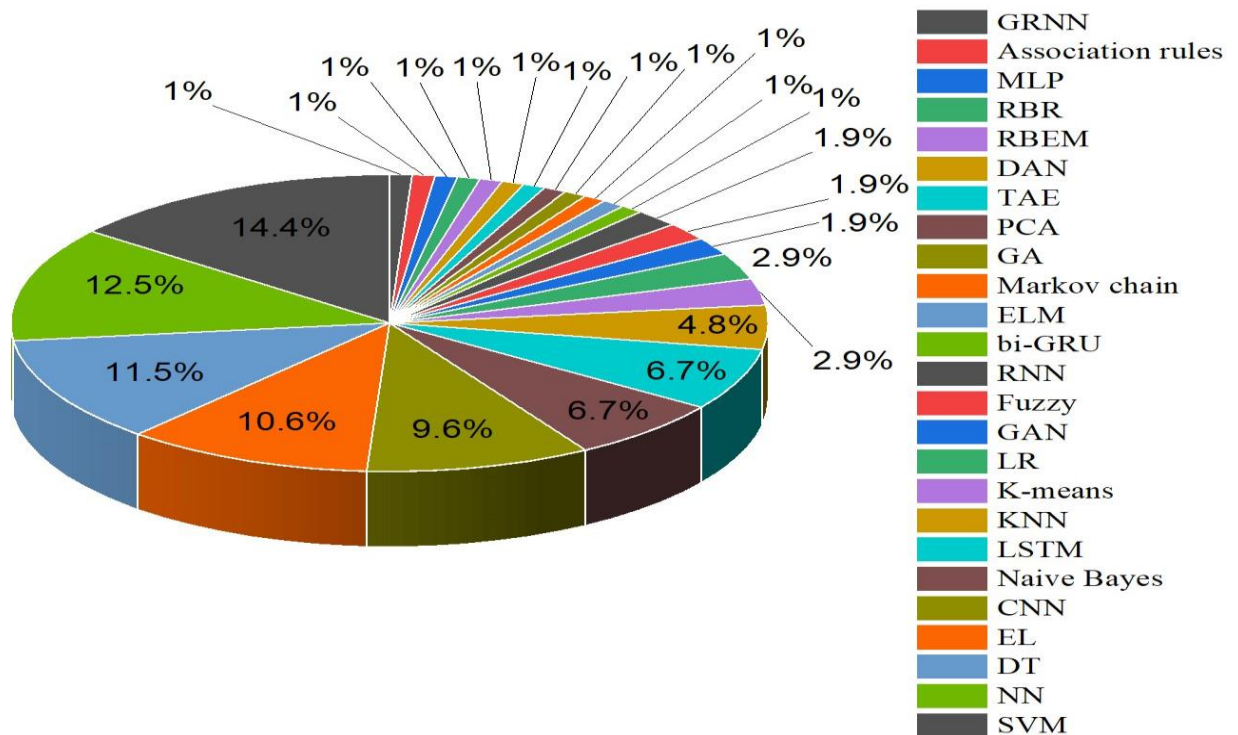


Рисунок 3.3 – Методи ML, які в основному використовуються для BDA [40]

В більшості досліджень [59] використовувалися глибокі нейронні мережі (DNN), зокрема, LSTM і згорткові нейронні мережі (CNN), як показано на рис. 3.3. Цей результат можна віднести до збереження пам'яті LSTM і вирішення проблеми зникнення градієнта. CNN може автоматично помічати та видобувати відповідну внутрішню структуру з набору даних часового ряду для створення детальних вхідних функцій, використовуючи операції згортки та об'єднання. Крім того, алгоритми CNN і LSTM стійкі до перешкод і точні для класифікації часових рядів.

Розглядаючи ситуацію, коли ці техніки були гібридними з іншими техніками DNN, можна сказати, що BDA і DL, залежать один від одного і демонструють взаємовигідну асоціацію. Великі обсяги даних дозволяють методам DL краще узагальнювати, таким чином, даючи значиміші та цінніші результати в галузі BDA. Отже, згідно [60], методи ML, так само як DNN,

варто вивчити в подальших дослідженнях, щоб дослідити характеристики платформ. Знову ж таки, GAN, запропонований Яном Гудфеллоу в 2014 році, розглядається як надійний алгоритм глибокого навчання [61], але на даний час йому приділено мало уваги в BDA. Таким чином, майбутні дослідження можуть віртуально вивчити цю техніку, щоб перевірити її здатність у BDA.

Крім того, слід відзначити, що в дослідженнях [47] застосовуються методи ансамблевого навчання, зокрема, випадковий ліс, посилення та пакетування, щоб підвищити спроможності методів EL у BDA. Гібридизація алгоритму ML є чудовою технікою для компенсації недоліків окремого алгоритму [22]. Однак було виявлено, що лише деякі дослідження з розглянутих наукових публікацій прийняли це [59]. З рисунку 3.3 видно, що більшість алгоритмів ML застосовуються до BDA.

Однак дві сфери потребують подальшого вдосконалення:

- великі обчислювальні витрати, пов'язані з більшістю алгоритмів ML;
- вартість зв'язку для різноманітних комп'ютерних вузлів у паралельних обчисленнях.

На рисунку 3.4 показано розподіл проаналізованих наукових публікацій по країнах.

Більшість досліджень було проведено в Китаї (37%), за яким йдуть США (18%). Хоча автор [62] повідомляє, що більше досліджень аналітики BDA розвивається в Китаї, ніж у США. Цей результат не є несподіванкою, оскільки переважна більшість опублікованих результатів досліджень показує, що в Китаї найбільше користувачів мобільних телефонів, за ним йдуть Індія та США. Результат не є несподіванкою, оскільки населення Китаю становить 18,5% від загального населення світу. Таким чином, можна зробити висновок, що Китай продукує більше електронних даних від користувачів мобільних телефонів, ніж будь-яка інша країна. Отже, необхідні додаткові дослідження великих даних. Цікаво, що в основних опублікованих документах Африка не вказана.

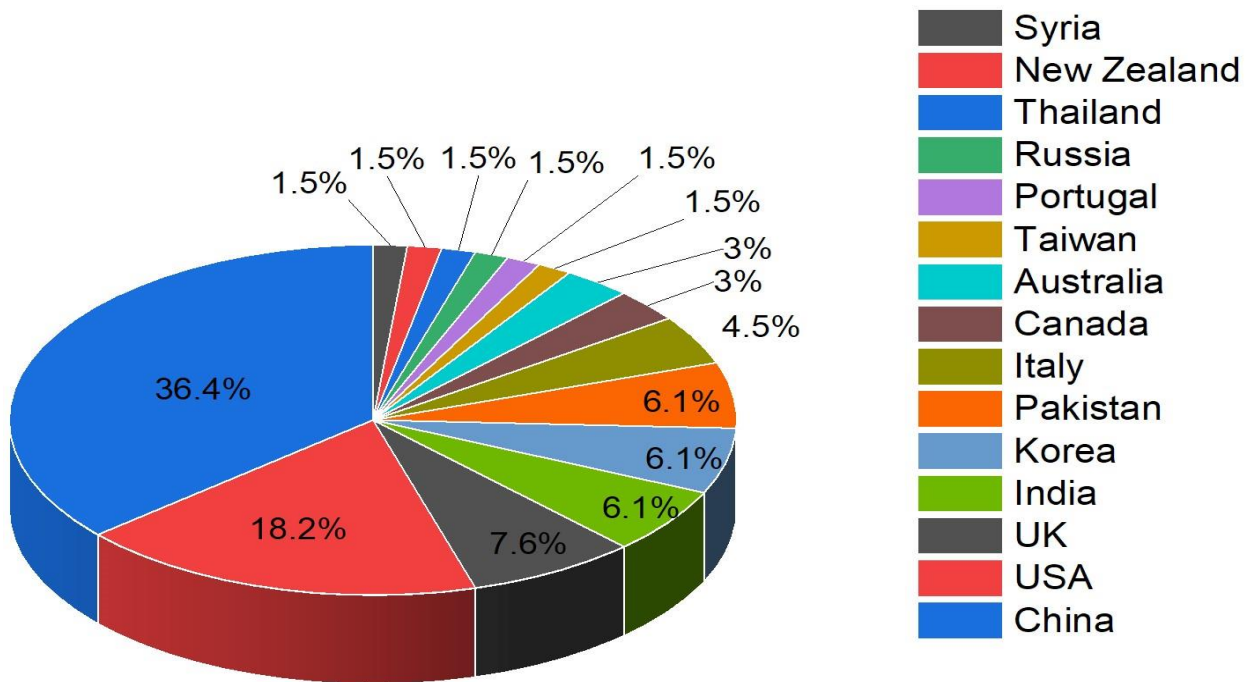


Рисунок 3.4 – Поширення джерела публікації [40]

Тому доцільно заохочувати аналітику великих даних, отриманих на континенті, для покращення бізнес-рішень.

3.4 Метрики оцінки, що використовуються в галузі аналітичного опрацювання великих даних

Для вимірювання продуктивності моделей машинного навчання можна використовувати кілька показників оцінювання залежно від завдання ML [22]. Деякі з найпоширеніших: середньоквадратична помилка (MSE), середньоквадратична помилка (RMSE), середня абсолютна помилка (MAE), точність, точність, відкликання, площа під кривою (AUC), і F-бал. Щоб отримати докладніші відомості та визначення різних показників оцінювання, читачі можуть звернутися до Ref. [22]. На рисунку 3.5 показано розподіл метрик помилок, що використовуються в BDA.

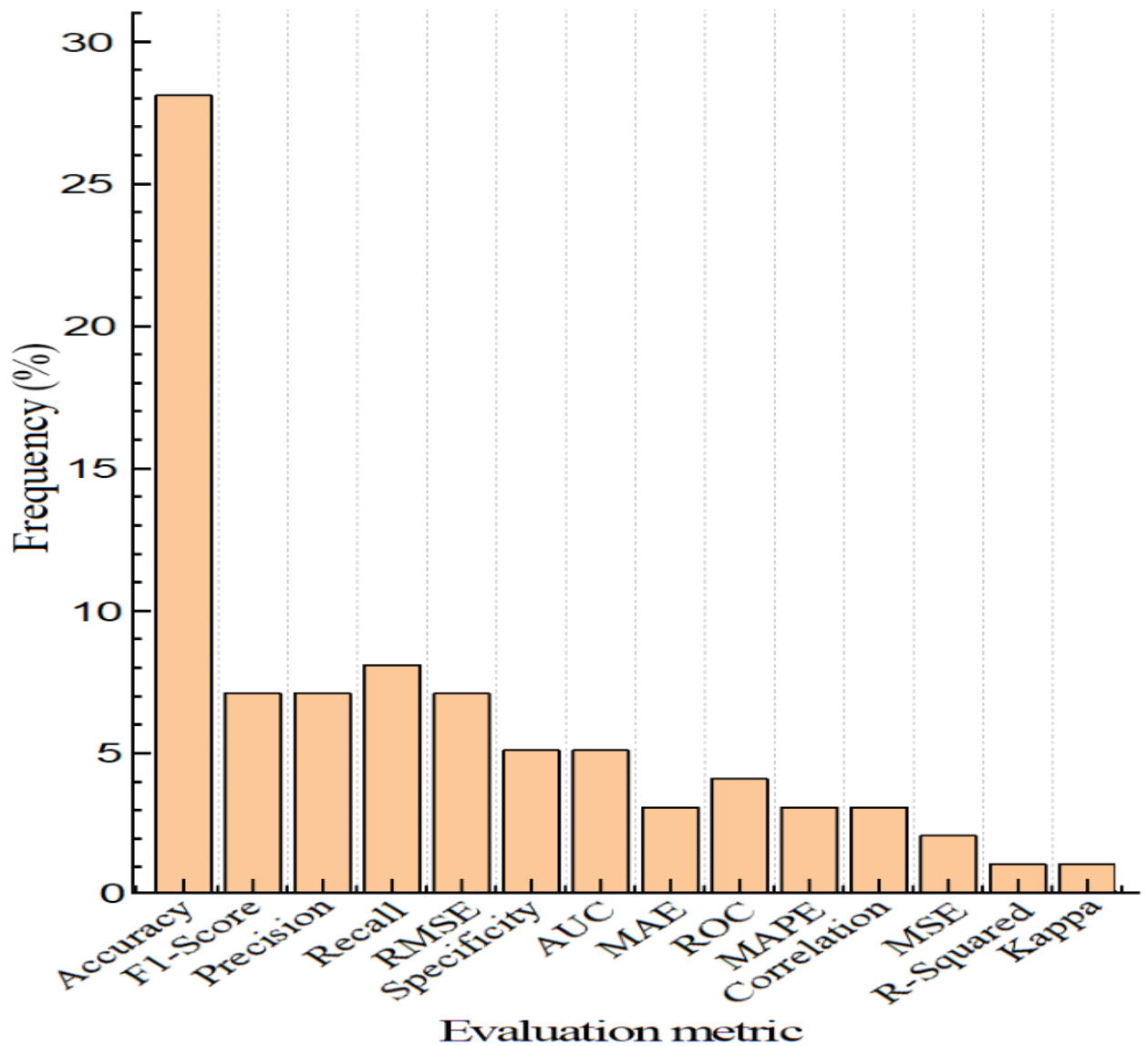


Рисунок 3.5 – Розподіл показників оцінювання [40]

Було помічено, що точність (37%) була найбільш використовуваним показником, оскільки більшість наукових публікацій були прив'язані прогностичній аналітиці BD. Таким чином, можна стверджувати, що поєднання показників точності та помилок пропонує кращу оцінку моделі ML [22]. У дослідженні [59] поєднано два або більше показників для оцінки моделей і структури ML.

3.5 Ключові проблеми аналітичного опрацювання великих даних та перспективи майбутніх досліджень

IoT-пристрої генерують великі за розміром набори, масиви та колекції даних, які створюють багато проблем з точки зору чотирьох характеристик великих даних, зокрема швидкості, правдивості, обсягу та різноманітності.

Зберігання та аналітичне опрацювання IoT-даних є серйозною проблемою, оскільки дані, створені IoT-пристроями, збільшуються з високою швидкістю. Водночас ємність систем, що обробляють великі дані, обмежена [63]. Багато хмарних IoT-платформ, які пропонують постачальники хмарних послуг Amazon AWS, Microsoft Azure і Google Cloud, пропонують багатофункціональні служби зберігання даних для великих за обсягом набори та колекції IoT-даних.

Візуалізація даних є важливим аспектом аналітичного опрацювання великих даних, отриманих за допомогою IoT-пристроїв, оскільки дані, створені IoT-пристроями, є структурованими, напівструктурованими або неструктурованими гетерогенними даними. Дані повинні бути підготовлені для ефективної візуалізації та аналізу для прийняття ефективних рішень.

Безпека пристроїв – це складне багатofакторне завдання, яке постає перед галуззю аналітичного опрацювання великих даних, оскільки безпека пристроїв, підключених до мережі, залежить від постачальника, і якщо захист слабкий, це має значний вплив. Кібербезпека буде скомпрометована та забезпечить доступ до центрів обробки даних, пристроїв і призведе до втрати особистих даних. Безперервна подача живлення для IoT-пристроїв є важливим питанням для безперервної, безперебійної та стабільної роботи.

IoT-дані надають обширний перелік можливостей для досліджень у будь-якій області, де використовуються аналітичне опрацювання великих даних, IoT-пристрої разом із давачами та приводами. IoT-дані генеруються та аналітично опрацьовуються в різних сферах:

- охорона здоров'я;
- навколишнє середовище;
- Розумні міста;
- енергетика;
- транспорт;
- виробництво;
- сільське господарство тощо.

Обсяг досліджень в галузі аналітичного опрацювання великих даних на даний час є доволі обширним, а проблеми, що виникають у сфері зберігання, безпеки та візуалізації, також різноманітні. Універсальним стандартом для аналітичного опрацювання великих даних є критичні вимоги до часу, яка матиме величезний вплив на рівень людського життя.

Виокремимо перелік невирішених проблем, пов'язаних з аналітичним опрацюванням великих даних [35]:

- Етичні проблеми. Було помічено, що існує декілька етичних проблем, зокрема, авторське право – повторне використання даних без дозволу, конфіденційність і участь конкуруючої організації, пов'язаної з великими даними. Цей результат підтверджує доповідь [64].

- Навички є ще одним важливим викликом. Аналітика великих даних передбачає командну роботу експертів і кваліфікованого персоналу з різних галузей, зокрема, інформатики, біології, математики, медицини, економіки тощо, для виконання певного завдання. Крім того, інфраструктури, зокрема, програмне та апаратне забезпечення – бажано суперкомп'ютери для аналітичної обробки великих даних, роблять цю галузь коштовною для малих і середніх підприємств і організацій або у країнах з малим і середнім доходом.

- На даний час аналітика великих даних стала популярною темою для обговорення серед вчених і професіоналів галузі в усьому світі [65] через появу нових складних технологій, інструментів і платформ. Проте ряд

досліджень проаналізованих в [35] вказали на складність, пов'язану з обробкою цієї величезної та безпрецедентної кількості даних, що складається з різних типів даних та поточкових даних [66]. Таким чином, великі дані є багатовимірними, різноманітними, гігантськими, складними, неповними, аморфними, шумними та помилковими, що ускладнює попередню обробку даних у BDA. Однак дуже важливо, щоб моделі машинного навчання ефективно працювали з високою точністю. Тим не менш, можна підтвердити, що ця проблема все ще існує в BDA сьогодні.

В [40] було визначено кілька відкритих місць, які необхідно вирішити в рамках перспективних досліджень:

- Аналітичне опрацювання великих даних у соціальних мережах застосовувалася на політичній арені країн, що розвиваються, для вдосконалення стратегій передвиборчої кампанії. Наприклад, кампанія Обама запропонувала модель класифікації окремих виборців, яка привела їхню стратегію кампанії до ефективного результату [67]. Однак немає впевненості, що ці моделі можна перенести на вибори та іншу політичну діяльність і поведінку в слаборозвинутих країнах і країнах у всьому світі. Оскільки на індивідуальні почуття та вирази в соціальних мережах природньо впливає обширний перелік культурних точок зору.

- Мережі мобільного стільникового зв'язку є сектором, якому приділяється мало уваги в царині аналітичного опрацювання великих даних, але вони створюють і передають великі обсяги даних на регулярній основі.

Однак DBA може підвищити продуктивність MCN і максимізувати прибутки операторів [68]. Отже, є можливість для подальших досліджень у цій галузі. Подібним чином, вивчення різних міждомених наборів даних для прогнозування стільникового трафіку та передачі знань у «Розумних містах» є перспективним напрямком у майбутньому.

Деякі дослідження розглядали управління трафіком на основі аналізу великих даних. Однак великий відсоток цих робіт зосереджено на одному

джерелі даних. Із збільшенням кількості джерел даних у транспортній галузі очікується, що майбутні дослідження поєднуватимуть дані з різних джерел даних, наприклад, інформація про погоду, камери відеоспостереження та інші джерела даних, пов'язані з транспортом.

Помічено, що розробники платформ BDA зазвичай надають очікуваним користувачам обмеження – параметри, які можна налаштувати, щоб підвищити ефективність і продуктивність своїх інфраструктур. Тим не менш, складність і велика розмірність цих обмежень роблять ручне налаштування BDA на цих структурах трудомістким, складним і непродуктивним [69]. Автори [70] зазначають, що Hadoop має понад 190 конфігураційних обмежень, які можуть суттєво вплинути на продуктивність і часові характеристики. Це призводить до ряду проблем у створенні високопродуктивних моделей для фреймворків БД. Тому майбутні дослідження можуть вивчити методи, щоб зробити процес конфігурації автоматизації більш гнучким.

Незважаючи на те, що BDA стрімко зростає в останні роки дослідження [71] виявили відсутність гендерного балансу в адаптації BDA в більшості галузей. Тобто чоловіки, порівняно з жінками, домінують щодо позитивного наміру використовувати BDA. Водночас жінки чинять більший опір змінам, ніж чоловіки, приймаючи BDA в медичних та суміжних організаціях. Отже, оскільки BDA буде розвиватись у майбутньому слід проводити більше пропагандистських та орієнтаційних досліджень, щоб сприяти гендерному балансу в BDA.

Щоб полегшити роботу з великими даними, надалі необхідно вирішувати виявлені проблеми, узгоджувати неповні та різноманітні джерела даних, шумні та помилкові дані, які впливають на ефективність аналітики даних. Тому розробникам засобів аналітичного опрацювання великих даних необхідно високо та ефективно автоматизувати етап попередньої обробки

даних, наприклад, очищення даних, вибірка та стиснення, де це можливо, з меншими людськими зусиллями.

3.6 Висновок до третього розділу

В третьому розділі кваліфікаційної роботи описано інструменти аналітичного опрацювання великих даних. Проаналізовано результати досліджень в галузі аналітичного опрацювання великих даних. Висвітлено часовий розподіл наукових публікацій щодо аналітичного опрацювання великих даних. Проаналізовано метрики оцінки, що використовуються в галузі аналітичного опрацювання великих даних. Розглянуто ключові проблеми аналітичного опрацювання великих даних та перспективи майбутніх досліджень.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1 Медичні профілактичні заходи щодо збереження здоров'я та працездатності користувачів комп'ютерів та відеодисплейних терміналів

Кваліфікаційна робота освітньо-наукового рівня «Магістр» присв'ячена дослідженню засобів та методів машинного навчання для аналітичного опрацювання великих даних. Більшість робіт щодо аналітичного опрацювання великих даних відбувається з використанням ПК. Тому в контексті охорони праці доцільно розглянути медичні профілактичні заходи щодо збереження здоров'я та працездатності користувачів ПК та відеодисплейних терміналів. Заходи з охорони праці користувачів ПК необхідно розглядати в трьох основних аспектах [72]:

- соціальному;
- психологічному;
- медичному.

У соціальному плані розв'язання цих проблем пов'язане з оптимізацією умов життя, праці, відпочинку, харчування, побуту, розвитком культури, транспорту.

Значне місце у профілактиці розладів здоров'я належить психології праці. Тому заходи, пов'язані з формуванням раціональних колективів, у яких відсутня психологічна несумісність, сприяють зменшенню нервово-психічного перенапруження, підвищенню працездатності та ефективності праці.

Особливої значущості у користувачів відеодисплейних терміналів набуває психоемоційний стрес, який більшою або меншою мірою проявляється у кожного з них.

Існує перелік профілактичних заходів для користувачів ПК, що включає як складові первинної профілактики здоров'я (професійний відбір),

так і вторинної, яка направлена на зниження ймовірності розвитку перевтоми та перенапруження. Ці комплексні заходи спрямовані на відновлення функціонального стану зорового та опорно-рухового апарату.

Медичні профілактичні заходи щодо збереження здоров'я та підвищення працездатності користувачів комп'ютерів:

- медичні огляди;
- раціональне та профілактичне харчування;
- спеціальні вправи;
- самомасаж;
- психофізіологічне розвантаження.

Режим праці і відпочинку передбачають з урахуванням специфіки праці робітників. У першу чергу забезпечують оптимальні режими особам, що працюють з підвищеними фізичними і нервово-емоційними навантаженнями, в умовах монотонності із впливом небезпечних і шкідливих виробничих факторів.

Лікувально-профілактичне обслуговування включає попереджувальні та періодичні медичні огляди працюючих, лікувально-профілактичне харчування і проведення лікувально-профілактичних заходів щодо запобігання захворюванням працюючих [73]. Санітарно-побутове обслуговування передбачає забезпечення працюючих санітарно-побутовими приміщеннями і пристроями та їх функціонування згідно з чинними нормами і правилами. Професійний відбір за окремими спеціальностями передбачає визначення професійної (фізіологічної та психофізіологічної) придатності працюючих до безпечного виконання робіт.

В даному параграфі було розглянуто медичні профілактичні заходи щодо збереження здоров'я та працездатності користувачів комп'ютерів та відеодисплейних терміналів тому, що кваліфікаційна робота присв'ячена дослідженню засобів та методів машинного навчання для аналітичного опрацювання великих даних.

4.2 Організація оповіщення і зв'язку у надзвичайних ситуаціях техногенного та природного характеру

На даний час функціонування обчислювальних та даних центрів для аналітичного опрацювання великих даних в Україні відбувається в умовах війни, що супроводжуються виникненням різнотипових надзвичайних ситуацій. Правовою основою організації оповіщення населення загрози чи виникненні надзвичайних ситуацій є Конституція України, Кодекс цивільного захисту України, постанова Кабінету Міністрів України від 27.09.2017 № 733 «Про затвердження Положення про організацію оповіщення про загрозу виникнення або виникнення надзвичайних ситуацій та зв'язку у сфері цивільного захисту» [74], відповідні розпорядження обласної державної адміністрації та інші акти.

Одним із основних завдань Цивільного захисту України як державної системи органів управління, сил і засобів, які створені для організації і забезпечення захисту населення від наслідків надзвичайних ситуацій техногенного, природного та воєнного характеру є оповіщення населення про загрозу і виникнення надзвичайних ситуацій у мирний час і особливий період та постійне інформування його про наявну обстановку.

Система централізованого оповіщення представляє собою комплекс організаційно-технічних заходів, апаратури і технічних засобів оповіщення, засобів та каналів зв'язку, мереж дротового, радіо, телевізійного мовлення, призначених для своєчасного доведення сигналів та інформації з питань цивільного захисту до центральних і місцевих органів виконавчої влади, підприємств, установ, організацій і населення. Для зосередження уваги громадян перед передачею мовної інформації вмикаються сирени, інші сигнальні засоби. Їх звук означає попереджувальний сигнал «УВАГА ВСІМ».

Телерадіокомпанії незалежно від форми власності та радіотрансляційні вузли операторів телекомунікацій оприлюднюють повідомлення про загрозу

виникнення або виникнення надзвичайних ситуацій, а також іншу інформацію з питань цивільного захисту (відомості про надзвичайні ситуації, що прогнозуються або виникли, межі їх поширення і наслідки, а також способи та методи захисту від них) на безоплатній основі.

Переривання трансляції програм мовлення для оповіщення населення здійснюється в автоматичному режимі за допомогою спеціальних технічних засобів, встановлених у апаратних телерадіокомпаній та на пунктах управління обласної державної адміністрації (в чергових службах органів місцевого самоврядування) [75].

У разі неможливості переривання трансляції програм мовлення з пунктів управління обласних державних адміністрацій (чергових служб органів місцевого самоврядування) оповіщення населення здійснюється безпосередньо з радіотрансляційних вузлів, апаратних телерадіокомпаній відповідно до спільних інструкцій, які розробляються місцевими органами виконавчої влади або органами місцевого самоврядування за участю телерадіокомпаній. За рівнями системи оповіщення поділяються на загальнодержавну автоматизовану систему централізованого оповіщення, територіальні автоматизовані системи централізованого оповіщення, місцеві автоматизовані системи централізованого оповіщення, а також спеціальні, локальні та об'єктові системи оповіщення.

Територіальна автоматизована система централізованого оповіщення функціонує в регіонах, областях та локаціях для забезпечення прийому сигналів та інформації від загальнодержавної автоматизованої системи централізованого оповіщення, оповіщення осіб керівного складу місцевих органів виконавчої влади, а також органів місцевого самоврядування, підприємств, установ, організацій, органів управління та сил цивільного захисту і населення через місцеві автоматизовані системи централізованого оповіщення та інші системи оповіщення у разі загрози виникнення або виникнення надзвичайних ситуацій.

ВИСНОВКИ

В першому розділі кваліфікаційної роботи освітнього рівня «Магістр»:

- Описано розвиток наукових досліджень в галузі аналітичного опрацювання великих даних.
- В комплексі розглянуто Інтернет речей та аналітичне опрацювання великих даних.
- Описано інформаційно-технологічні IoT-платформи та аналітичне опрацювання великих даних.
- Досліджено концепцію великих даних та їх аналітичне опрацювання.

В другому розділі кваліфікаційної роботи:

- Досліджено машинне навчання та аналітичне опрацювання великих даних.
- Описано методи машинного навчання для аналітичного опрацювання великих даних.
- Висвітлена методика аналізу літературних джерел щодо засобів та методів машинного навчання для аналітичного опрацювання великих даних.

В третьому розділі кваліфікаційної роботи:

- Описано інструменти аналітичного опрацювання великих даних.
- Проаналізовано результати досліджень в галузі аналітичного опрацювання великих даних.
- Висвітлено часовий розподіл наукових публікацій щодо аналітичного опрацювання великих даних.
- Проаналізовано метрики оцінки, що використовуються в галузі аналітичного опрацювання великих даних.
- Розглянуто ключові проблеми аналітичного опрацювання великих даних та перспективи майбутніх досліджень.

У розділі «Охорона праці та безпека в надзвичайних ситуаціях» проаналізовано медичні профілактичні заходи щодо збереження здоров'я та

працездатності користувачів комп'ютерів та відеодисплейних терміналів. Описано процеси організації оповіщення і зв'язку у надзвичайних ситуаціях техногенного та природного характеру.

ПЕРЕЛІК ДЖЕРЕЛ

- 1 F. Balali, J. Nouri, A. Nasiri, and T. Zhao, “Data Analytics,” *Data Intensive Ind. Asset Manag.*, pp. 105–113, 2020, doi: 10.1007/978-3-030-35930-0_7.
- 2 A. Yassine, M. S. Hossain, G. Muhammad, S. Singh, and M. Shamim Hossain, “IoT Big Data Analytics for Smart Homes with Fog and Cloud Computing Smart Meters Big Data View project A Vision System for Date Fruit Harvesting Robot View project IoT Big Data Analytics for Smart Homes with Fog and Cloud Computing”, *Future Generation Computer Systems*, pp.563-573, 2019, doi: 10.1016/j.future.2018.08.040.
- 3 Duda, O., Kunanets, N., Martsenko, S., Matsiuk, O., Pasichnyk, V., *Building secure Urban information systems based on IoT technologies. CEUR Workshop Proceedings 2623*, pp. 317-328. 2020.
- 4 S. S. Gill, R. C. Arya, G. S. Wander, and R. Buyya, “Fog-Based Smart Healthcare as a Big Data and Cloud Service for Heart Patients Using IoT,” *Lecture Notes Data Engineering and Communication Technology*, vol. 26, pp. 1376–1383, 2019, doi: 10.1007/978-3-030-03146-6_161.
- 5 Bodnarchuk I., Duda O., Kharchenko A., Kunanets N., Matsiuk O., Pasichnyk V. Choice method of analytical information-technology platform for projects associated to the smart city class. *ICTERI 2020 ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer Proceedings of the 14th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume I: Main Conference* p.317-330.
- 6 .M. Chen, J. Yang, L. Hu, M. Shamim Hossain, and G. Muhammad, “Urban Healthcare Big Data System Based on Crowdsourced and Cloud-Based Air Quality Indicators,” *IEEE Communications Magazine*, vol. 56, no. 11, pp. 14–20, 2018, doi: 10.1109/MCOM.2018.1700571.

- 7 Duda, O., et al, Selection of Effective Methods of Big Data Analytical Processing in Information Systems of Smart Cities. CEUR Workshop Proceedings 2631, pp. 68-78. 2020.
- 8 Y. I. Gonzales et al., “The Internet of Things (IoT): An Overview Related papers An overview of the Internet of Things for people with disabilities” vol. 5, pp. 71–82, 2015.
- 9 Duda O., Matsiuk O., Kunanets N., Pasichnyk V., Rzhenskyi A., Bilak Y., Formation of Hypercubes Based on Data Obtained from Systems of IoT Devices of Urban Resource Networks, International Journal of Sensors, Wireless Communications and Control (2020) 10: 1. ISSN 2210-3287.
- 10 M. Wu and J. Luo, “Wearable Technology Applications in Healthcare: A Literature Review,” Online J. Nurs. Informatics Contrib., vol. 23, no. 3, pp. 1–10, 2019.
- 11 Duda, O., Palka, O., Pasichnyk, V., Matsiuk, O., Kunanets, N., & Tabachyshyn, D. (2020, September). Existing City Assessment Systems. In 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT) (Vol. 2, pp. 238-241). IEEE.
- 12 E. Said Mohamed, A. A. Belal, S. Kotb Abd-Elmabod, M. A. El-Shirbeny, A. Gad, and M. B. Zahran, “Smart farming for improving agricultural management,” Egypt. J. Remote Sens. Sp. Sci., vol. 24, no. 3, pp. 971–981, Dec. 2021, doi: 10.1016/J.EJRS.2021.08.007.
- 13 K. Sharma, D. Anand, M. Sabharwal, P. K. Tiwari, O. Cheikhrouhou, and T. Frikha, “A Disaster Management Framework Using Internet of Things-Based Interconnected Devices,” Math. Probl. Eng., vol. 2021, 2021, doi: 10.1155/2021/9916440.
- 14 S. N. Swamy, D. Jadhav, and N. Kulkarni, “Security threats in the application layer in IOT applications,” Proceedings of International Conference on IoT Soc. Mobile, Analytics and Cloud, I-SMAC 2017, pp. 477–480, Oct. 2017, doi: 10.1109/I-SMAC.2017.8058395.

- 15 N. M. Kumar and P. K. Mallick, “The Internet of Things: Insights into the building blocks, component interactions, and architecture layers,” in *Procedia Computer Science*, vol. 132, pp. 109–117, 2018 doi: 10.1016/j.procs.2018.05.170.
- 16 L. Nastase, “Security in the Internet of Things: A Survey on Application Layer Protocols,” in *Proceedings - 2017 21st International Conference on Control Systems and Computer, CSCS 2017*, pp. 659–666, 2017, doi: 10.1109/CSCS.2017.101.
- 17 L. Oliveira, J. J. P. C. Rodrigues, S. A. Kozlov, R. A. L. Rabêlo, and V. H. C. De Albuquerque, “MAC layer protocols for internet of things: A survey,” *Futur. Internet*, vol. 11, no. 1, pp. 1–42, 2019, doi: 10.3390/fi11010016.
- 18 P. Desai, A. Sheth, and P. Anantharam, “Semantic Gateway as a Service Architecture for IoT Interoperability,” in *Proceedings - 2015 IEEE 3rd International Conference on Mobile Services, MS 2015*, pp. 313–319, 2015, doi: 10.1109/MobServ.2015.51.
- 19 J. Zakir, T. Seymour, and K. Berg, Big data analytics, *Issues Inf. Syst.*, vol. 16, no. 2, pp. 81–90, 2015.
- 20 R. Raja, I. Mukherjee, and B. K. Sarkar, A systematic review of healthcare big data, *Sci. Program.*, vol. 2020, p. 5471849, 2020.
- 21 R. H. Hariri, E. M. Fredericks, and K. M. Bowers, “Uncertainty in big data analytics: survey, opportunities, and challenges,” *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0206-3.
- 22 I. K. Nti, A. F. Adekoya, and B. A. Weyori, A comprehensive evaluation of ensemble learning for stockmarket prediction, *J. Big Data*, vol. 7, no. 1, p. 20, 2020.
- 23 K. Vassakis, E. Petrakis, and I. Kopanakis, “Big data analytics: Applications, prospects and challenges,” in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 10, pp. 3–20, 2018.

- 24 A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, Apr. 2015, doi: 10.1016/J.IJINFOMGT.2014.10.007.
- 25 P. Russom, Introduction to big data analytics, <https://vivomente.com/wp-content/uploads/2016/04/bigdata-analytics-white-paper.pdf>, 2011.
- 26 Z. H. Sun, L. Z. Sun, and K. Strang, Big data analytics services for enhancing business intelligence, *J. Comput. Inf. Syst.*, vol. 58, no. 2, pp. 162–169, 2018.
- 27 T. T. Le, W. X. Fu, and J. H. Moore, Scaling tree-based automated machine learning to biomedical big data with a feature set selector, *Bioinformatics*, vol. 36, no. 1, pp. 250–256, 2020.
- 28 B. Aragona and R. De Rosa, Big data in policy making, *Math. Popul. Stud.*, vol. 26, no. 2, pp. 107–113, 2019.
- 29 A. Holst, Amount of information globally 2010–2024, <https://www.statista.com/statistics/871513/worldwidedata-created/>, 2020.
- 30 B. K. Sarkar, Big data for secure healthcare system: A conceptual design, *Complex Intell. Syst.*, vol. 3, no. 2, pp. 133–151, 2017.
- 31 G. Kaur, P. Tomar, and P. Singh, Design of cloud-based green IoT architecture for smart cities, in *Internet of Things and Big Data Analytics Toward Next-Generation Intelligence*, N. Dey, A. E. Hassanien, C. Bhatt, A. S. Ashour, and S. C. Satapathy, eds. Cham, Germany: Springer, 2018, pp. 315–333.
- 32 K. Shafique, B. A. Khawaja, F. Sabir, S. Qazi, and M. Mustaqim, “Internet of things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT Scenarios,” *IEEE Access*, vol. 8, pp. 23022–23040, 2020, doi: 10.1109/ACCESS.2020.2970118.
- 33 I. Cisco Systems, “Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper,” Cisco Forecast Methodol., pp. 2017–2022, 2019.
- 34 K. Y. Ngiam and I. W. Khor, Big data and machine learning algorithms for health-care delivery, *Lancet Oncol.*, vol. 20, no. 5, pp. e262–e273, 2019.

- 35 J. F. Qiu, Q. H. Wu, G. R. Ding, Y. H. Xu, and S. Feng, A survey of machine learning for big data processing, *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, pp. 67, 2016.
- 36 B. Ale, Risk analysis and big data, *Saf. Reliab.*, vol. 36, no. 3, pp. 153–165, 2016.
- 37 O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, Efficient machine learning for big data: A review, *Big Data Res.*, vol. 2, no. 3, pp. 87–93, 2015.
- 38 D. Z. Chong and H. Shi, Big data analytics: A literature review, *J. Manag. Anal.*, vol. 2, no. 3, pp. 175–201, 2015.
- 39 L. Collins, Mini literature review: A new type of literature review article, [https://www.emeraldgroup publishing.com/archived/products/journals/call for papers. htm%3Fid%3D5730](https://www.emeraldgrouppublishing.com/archived/products/journals/call_for_papers.htm%3Fid%3D5730), 2021.
- 40 Nti, Isaac Kofi, et al. "A mini-review of machine learning in big data analytics: Applications, challenges, and prospects." *Big Data Mining and Analytics* 5.2 (2022): 81-97.
- 41 Amandeep Singh, K., and T. V. Ananthan. "Research Challenges on Big Internet of Things Data Analytics." *Journal of Computational and Theoretical Nanoscience* 16.5-6 (2019): 2113-2116.
- 42 N. Akhtar, F. Parwej, and Y. Perwej, “A Perusal of Big Data Classification and Hadoop Technology ” *Science and Education*, vol. 4, no. 1, pp. 26–38, 2017, doi: 10.12691/iteces-4-1-4.
- 43 M. Assefi, E. Behraves, G. Liu, and A. P. Tafti, “Big data machine learning using apache spark MLlib,” in *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, vol. 2018–Janua, pp. 3492–3498, 2017, doi: 10.1109/BigData.2017.8258338.
- 44 E. G. Caldarola and A. M. Rinaldi, “Big data: A survey: The new paradigms, methodologies and tools,” in *DATA 2015 - 4th International*

Conference on Data Management Technologies and Applications, Proceedings, pp. 362–370, 2015, doi: 10.5220/0005580103620370.

45 A. MadhaviLatha and G. V Kumar, “Streaming data analysis using apache cassandra and zeppelin,” *IJISSET-International Journal of Innovative Science, Engineering & Technology*, vol. 3, no.10, 2016.

46 L. Nair, L. Nair, S. Shetty, and S. Shetty, “Interactive visual analytics on Big Data: Tableau vs D3.js,” *Journal of e-Learning and Knowledge Society*, vol. 12, no. 4, 2016.

47 C. Choi, J. Kim, J. Kim, D. Kim, Y. Bae, and H. S. Kim, Development of heavy rain damage prediction model using machine learning based on big data, *Adv. Meteorol.*, vol. 2018, p. 5024930, 2018.

48 N. Ahmed, A. L. C. Barczak, T. Susnjak, and M. A. Rashid, A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench, *J. Big Data*, vol. 7, no. 1, p. 110, 2020.

49 W. Gu, K. Foster, J. Shang, and L. R. Wei, A gamepredicting expert system using big data and machine learning, *Expert Syst. Appl.*, vol. 130, pp. 293–305, 2019.

50 K. P. Zhu, G. C. Li, and Y. Zhang, Big data oriented smart tool condition monitoring system, *IEEE Trans. Ind. Inform.*, vol. 16, no. 6, pp. 4007–4016, 2020.

51 A. Bousdekis, N. Papageorgiou, B. Magoutas, D. Apostolou, and G. Mentzas, Sensor-driven learning of time-dependent parameters for prescriptive analytics, *IEEE Access*, vol. 8, pp. 92383–92392, 2020.

52 B. Cleland, J. Wallace, R. Bond, M. Black, M. Mulvenna, D. Rankin, and A. Tanney, Insights into antidepressant prescribing using open health data, *Big Data Res.*, vol. 12, pp. 41–48, 2018.

53 M. Giacalone, C. Cusatelli, and V. Santarcangelo, Big data compliance for innovative clinical models, *Big Data Res.*, vol. 12, pp. 35–40, 2018.

54 K. A. Jallad, M. Aljnidi, and M. S. Desouki, Anomaly detection optimization using big data and deep learning to reduce false-positive, *J. Big Data*, vol. 7, no. 1, p. 68, 2020.

55 F. Celli, F. Cumbo, and E. Weitschek, Classification of large DNA methylation datasets for identifying cancer drivers, *Big Data Res.*, vol. 13, pp. 21–28, 2018.

56 D. Chrimes and H. Zamani, Using distributed data over HBase in big data analytics platform for clinical services, *Comput. Math. Methods Med.*, vol. 2017, p. 6120820, 2017.

57 L. Gu and H. Li, Memory or time: Performance evaluation for iterative operation on hadoop and spark, in *Proc. 10th Int. Conf. High Performance Computing and Communications & 2013 IEEE Int. Conf. Embedded and Ubiquitous Computing*, Zhangjiajie, China, 2013, pp. 721–727.

58 Y. Samadi, M. Zbakh, and C. Taddonki, Performance comparison between Hadoop and Spark frameworks using HiBench benchmarks, *Concurr. Comput.: Pract. Exp.* vol. 30, no. 12, p. e4367, 2018.

59 M. Chen, Y. X. Hao, K. Hwang, L. Wang, and L. Wang, Disease prediction by machine learning over big data from healthcare communities, *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

60 D. Nallaperuma, R. Nawaratne, T. Bandaragoda, A. Adikari, S. Nguyen, T. Kempitiya, D. De Silva, D. Alahakoon, and D. Pothuhera, Online incremental machine learning platform for big data-driven smart traffic management, *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4679–4690, 2019.

61 A. Fonseca and B. Cabral, Prototyping a GPGPU neural network for deep-learning big data analysis, *Big Data Res.*, vol. 8, pp. 50–56, 2017.

62 S. Srivastava, Top 10 countries & regions leading the big data adoption in 2019, <https://www.analyticsinsight.net/top-10-countries-regions-leading-the-big-data-adoption-in-2019/>, 2020.

- 63 H. Cai, B. Xu, L. Jiang, and A. V. Vasilakos, "IoT-Based Big Data Storage Systems in Cloud Computing: Perspectives and Challenges," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 75–87, Feb. 2017, doi: 10.1109/JIOT.2016.2619369.
- 64 A. K. U. Haq, A. Khattak, N. Jamil, M. A. Naeem, and F. Mirza, Data analytics in mental healthcare, *Sci. Program.*, vol. 2020, p. 2024160, 2020.
- 65 Y. Samadi, M. Zbakh, and C. Tadonki, Comparative study between Hadoop and Spark based on Hibenx benchmarks, in *Proc. 2nd Int. Conf. Cloud Computing Technologies and Applications*, Marrakech, Morocco, 2016, pp. 267–275.
- 66 C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, Big data analytics: A survey, *J. Big Data*, vol. 2, no. 1, p. 21, 2015.
- 67 R. J. Dalton, The potential of big data for the crossnational study of political behavior, *Int. J. Sociol.*, vol. 46, no. 1, pp. 8–20, 2016.
- 68 Y. He, F. R. Yu, N. Zhao, H. X. Yin, H. P. Yao, and R. C. Qiu, Big data analytics in mobile cellular networks, *IEEE Access*, vol. 4, pp. 1985–1996, 2016.
- 69 M. Y. Li, Z. Q. Liu, X. H. Shi, and H. Jin, ATCS: Auto-tuning configurations of big data frameworks based on generative adversarial nets, *IEEE Access*, vol. 8, pp. 50485–50496, 2020.
- 70 M. Khan, Z. W. Huang, M. Z. Li, G. A. Taylor, P. M. Ashton, and M. Khan, Optimizing hadoop performance for big data analytics in smart grid, *Math. Probl. Eng.*, vol. 2017, p. 2198262, 2017.
- 71 M. Shahbaz, C. Y. Gao, L. L. Zhai, F. Shahzad, and M. R. Arshad, Moderating effects of gender and resistance to change on the adoption of big data analytics in healthcare, *Complexity*, vol. 2020, p. 2173765, 2020.
- 72 Основні правила дотримання охорони праці при роботі на персональних ЕОМ. URL: <https://www.victorija.ua/dovidnik/osnovni-pravyla-dotrymannya-ohorony-pratsi-pry-roboti-na-personalnyh-eom.html>.

73 КУРС ЛЕКЦІЙ. ОХОРОНА ПРАЦІ В ГАЛУЗІ. URL: <https://www.uzhnu.edu.ua/uk/infocentre/get/36621>.

74 Постанова Кабінету Міністрів України від 27.09.2017 № 733 «Про затвердження Положення про організацію оповіщення про загрозу виникнення або виникнення надзвичайних ситуацій та зв'язку у сфері цивільного захисту». URL: <https://zakon.rada.gov.ua/laws/show/733-2017-%D0%BF#Text>.

75 Організація оповіщення і зв'язку. URL: <https://guns.odessa.gov.ua/guns-opovwennya-naselennya/organ-zac-ya-opov-wennya-zv-yazku/>.

ДОДАТКИ

Тези конференції

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

МАТЕРІАЛИ

Х НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

«ІНФОРМАЦІЙНІ МОДЕЛІ,
СИСТЕМИ ТА ТЕХНОЛОГІЇ»



7–8 грудня 2022 року

ТЕРНОПІЛЬ
2022

УДК 001
М34

ПРОГРАМНИЙ КОМІТЕТ

Голова: Сергій Лупенко– докт. техн. наук, 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, асистент Марія Стадник; асистент Наталія Шаблій; ст. викладач Ліліана Джиджора.

Матеріали X науково-технічної конфіції «Інформаційні моделі, системи та технології»
М34 Тернопільського національного технічного університету імені Івана Пулюя,
(Тернопіль, 7–8 грудня 2022 р.). – Тернопіль : Тернопільський національний технічний
університет імені Івана Пулюя, 2022. –162 с.

Адреса оргкомітету: ТНТУ ім. І. Пулюя, м. Тернопіль, вул. Руська, 56, 46001,
тел. (0352) 52-41-33, факс (0352) 254983.
E-mail: confis2022@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 IX науково-технічної конференції «Інформаційні моделі, системи та технології» (Тернопіль, 7–8 грудня 2022 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.

© Тернопільський національний технічний
університет імені Івана Пулюя, 2022

О. Кравчук РОЗРОБКА ТЕЛЕГРАМ БОТІВ НА PYTHON	
O. Kravchuk DEVELOPMENT OF TELEGRAM BOTS IN PYTHON	29
Н. Лісовий, А. Ставицька, А. Гіжовський АНАЛІТИЧНЕ ОПРАЦЮВАННЯ ВЕЛИКИХ ЗА ОБСЯГОМ ДАНИХ	
N. Lisovyi, A. Stavyska, A. Hizhovskiy LARGE DATA VOLUMES ANALYTICAL PROCESSING	30
Н. Шаблій, П. Марценюк СИСТЕМИ МОНІТОРИНГУ СТАНУ ДОВКІЛЛЯ	
N. Shabliy, P. Martseniuk ENVIRONMENTAL STATE MONITORING SYSTEMS	31
Р. Маслій СИСТЕМА БЕЗПЕКИ ДЛЯ ІОТ З ВИКОРИСТАННЯМ SIEM ТЕХНОЛОГІЙ	
R. Maslii SECURITY SYSTEM FOR IOT USING SIEM TECHNOLOGIES	32
А. Блавицький, С. Мацюк, С. Криськова ЖИТТЄВИЙ ЦИКЛ ПЛАТЕЖУ	
A. Blavitskiy, S. Matsiuk, S. Kryskova PAYMENT LIFE CYCLE	33
М. Мокрицький, Ю. Скоренький ДОСЛІДЖЕННЯ ВРАЗЛИВОСТЕЙ НЕЙРОІНТЕРФЕЙСІВ	
M. Mokrytskyi, Yu. Skorenkyu STUDY OF BRAIN-COMPUTER INTERFACES VULNERABILITY	34
Г. Мушинська, Л. Дмитроца АНАЛІТИКА ОПТИМІЗАЦІЇ ЧАТ-БОТА	
H. Mushynska, L. Dmytrotsa CHAT BOT OPTIMIZATION ANALYTICS	35
К. Николин РОЗВІДКА ВІДКРИТИХ ДЖЕРЕЛ ІНФОРМАЦІЇ ДЛЯ ВИЯВЛЕННЯ ЗАГРОЗ БЕЗПЕКИ БІЗНЕСУ	
K. Nykolyn OPEN SOURCE INTELLIGENCE FOR IDENTIFYING BUSINESS SECURITY THREATS	36
Т. Патральський ТРАНСФОРМАЦІЯ ДАНИХ У НАСТРОЮВАНІ ІНФОРМАЦІЙНІ ЗВІТИ ТА ІНФОРМАЦІЙНІ ПАНЕЛІ LOOKER STUDIO	
T. Patralskiy DATA TRANSFORMATION INTO CUSTOMIZABLE INFORMATION REPORTS AND INFORMATION PANELS LOOKER STUDIO	37
Ю. Петришин СИСТЕМИ МЕНЕДЖМЕНТУ, МОДЕЛЬ ISO 27001	
Yu. Petryshyn MANAGEMENT SYSTEMS, ISO 27001 MODEL	38
П. Прийма, А. Зав'ялова, В. Дуда ІНТЕРНЕТ РЕЧЕЙ, «ВЕЛИКІ ДАНІ» ТА АНАЛІТИКА. СТАН ТА ПЕРСПЕКТИВИ ДОСЛІДЖЕНЬ	
P. Pryima, A. Zavialova, V. Duda THE INTERNET OF THINGS, BIG DATA AND ANALYTICS. RESEARCH STATUS AND PROSPECTS	39
П. Прийма, А. Зав'ялова, В. Дуда ІНСТРУМЕНТИ АНАЛІТИЧНОГО ОПРАЦЮВАННЯ «ВЕЛИКИХ ДАНИХ»	
P. Pryima, A. Zavialova, V. Duda TOOLS FOR BIG DATA ANALYTICAL PROCESSING	40

УДК 004.9

П. Прийма, А. Зав'ялова, В. Дуда

(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ІНТЕРНЕТ РЕЧЕЙ, «ВЕЛИКІ ДАНІ» ТА АНАЛІТИКА. СТАН ТА ПЕРСПЕКТИВИ ДОСЛІДЖЕНЬ

UDC 004.9

P. Pryima, A. Zavialova, V. Duda**THE INTERNET OF THINGS, BIG DATA AND ANALYTICS. RESEARCH STATUS AND PROSPECTS**

Розвиток, доступність та запровадження інноваційних цифрових технологій практично у всіх сферах людської діяльності призводить до генерації безпрецедентно великих наборів та колекцій даних. Цифрові дані, сформовані з використанням різних цифрових платформ і пристроїв, у всьому світі зростають неймовірними темпами [1]. Експоненційне зростання кількості інтегрованих в фізичне середовище пристроїв з датчиками та виконавчими механізмами, підключеними через мережу Інтернет, спричиняє відповідне зростання обсягів даних, отриманих завдяки інформаційним технологіям на основі інтернету речей (IoT, англ. Internet of Things). Сумісна робота людей та машин на основі IoT підвищує їх операційну ефективність та загальну продуктивність людинно-машинних систем.

Аналітичне опрацювання даних з використанням IoT-пристроїв спрощує процеси прийняття рішень, підвищує їх ефективність та, як наслідок, підвищує якість життя [2]. Це призвело до чергового етапу інформаційно-технологічної еволюції та формування засад «розумного» світу. Інтернет речей дозволяє поєднати фізичний світ з Інтернетом, який дозволяє передавати критично-важливу інформацію швидше, ніж інформаційні системи, яка залежить від втручання людей. Дані, зібрані IoT-пристроями, характеризуються великими обсягами, мінливістю, швидкоплинністю. Інформаційні технології «Великих даних» (англ. Big Data) забезпечують їх швидке та ефективніше зберігання і аналітичне опрацювання. Аналітика великих даних (BDA, англ. Big Data Analytics) застосовує інструменти аналітичного опрацювання до великих за обсягом потоків та наборів даних, створених IoT-пристроями, щоб допомогти в прийнятті ефективних та оперативних рішень. Конвергенція IoT, BDA і хмарних інформаційно-технологічних платформ відкриває обширний перелік напрямків наукових досліджень.

На даний час доступні хмарні програмно-алгоритмічні засоби, які надають можливості для ефективного аналітичного опрацювання «Великих даних» в режимі реального часу. Зокрема, це інструменти для аналізу великих за обсягом даних. Почастки вони сформовані з використанням алгоритмів машинного навчання (ML, англ. Machine Learning). Однак обширний перелік загальнодоступних та безкоштовних інструментів BDA, хмарних інформаційно-технологічних платформ та інструментів інтелектуального аналізу даних зазвичай ускладнює вибір засобів для достатньо ефективного виконання поставлених завдань. Тому потребують детальнішого та системного дослідження концепція IoT, особливості застосування IoT-пристроїв, характеристики «Великих даних», зв'язок між IoT та BDA, хмарні інструменти, що використовуються для BDA та прикладні задачі аналітичного опрацювання «великих даних» для потреб IoT-систем.

Література

1. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes). URL: <https://www.statista.com/statistics/871513/worldwide-data-created/>.
3. Nti, Isaac Kofi, et al. «A mini-review of machine learning in big data analytics: Applications, challenges, and prospects.» Big Data Mining and Analytics 5.2 (2022): 81–97.

УДК 004.9

П. Прийма, А. Зав'ялова, В. Дуда

(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ІНСТРУМЕНТИ АНАЛІТИЧНОГО ОПРАЦЮВАННЯ «ВЕЛИКИХ ДАНИХ»

UDC 004.9

P. Pryima, A. Zavialova, V. Duda**TOOLS FOR BIG DATA ANALYTICAL PROCESSING**

На даний час щодня генеруються великі за обсягом набори та колекції даних практично у всіх сферах людської діяльності. Зокрема дані надходять від соціальних мереж, інженерії, виробництва, транспорту, комерції, галузі охорони здоров'я, біомолекулярних досліджень, фізіології тощо. «Великі дані» (BD, англ. Big Data) та інноваційні методи та підходи їх аналітичного опрацювання, зокрема Big Data Analytics (BDA), змінили спосіб функціонування установ, підприємств та організацій, сформувавши при цьому обширний перелік нових перспективних напрямків досліджень для фахівців та наукової спільноти [1]. Окрім виробничих підприємств і дослідницьких установ, урядові та неурядові організації на даний час регулярно генерують великі за обсягом унікальні набори та колекції даних. Тому видобування та отримання значущої інформації із доступних «Великих даних» стало життєво важливим для підприємств, установ та організацій стало критично актуальним у всьому світі.

Інструменти аналітичного опрацювання «Великих даних» (англ. Big Data) використовуються для обробки великих за обсягом, структурованих, неструктурованих і напівструктурованих даних з метою видобування знань, бізнес-прогнозування, підвищення ефективності процесів прийняття рішень, візуалізації шаблонів тощо.

Apache Hadoop є одним із найпопулярніших інструментів аналізу даних. Він поширюється з відкритим кодом. HDFS (розподілена файлова система Hadoop) – це вискоелективний компонент зберігання даних, який використовується для зберігання різноманітних та різноманітних даних, зокрема тексту, xml або json файлів, аудіофайлів, зображень та відео. Зберігання відбувається завдяки поділу даних на частини та збереження в кластерах товарних серверів [2]. Заснований на Java, Apache Hadoop характеризується високою швидкістю, оскільки окремі завдання розділяються та виконуються одночасно на розподілених серверах. Оскільки дані зберігаються на множині розподілених серверів то резервне копіювання даних доступне, навіть при виході з ладу одного окремого сервера.

Apache Spark – це розподілена інформаційна система з відкритим кодом, яка обробляє дані з використанням апаратної оперативної пам'яті. Водночас швидкість обробки даних засобами Spark відчутно перевищує швидкість Hadoop [3]. Spark зручний для системних архітекторів та розробників програмного забезпечення, оскільки для створення програм можна використовувати різні мови програмування, зокрема java, python, R, scala тощо. На даний час обширний перелік організацій, зокрема Finra, Yelp, Zillow, gumgum, використовували Spark. Тому він став практично одним із найпопулярніших фреймворків розподіленої обробки «Великих даних».

Література

1. Islam, AYM Atiquil, et al. «Performance-based evaluation of academic libraries in the big data era.» *Journal of Information Science* 47.4 (2021): 458–471.
2. Akhtar, Nikhat, Firoj Parwej, and Yusuf Perwej. «A perusal of big data classification and hadoop technology.» *International Transaction of Electrical and Computer Engineers System (ITECES), USA* 4.1 (2017): 26–38.
4. Cao, Jian, et al. «Personalized flight recommendations via paired choice modeling.» *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017.