УДК 004.6

Велійка Олег Іванович,
(аспірант ПВНЗ «Європейський університет»)
ORCID ID 0009-0001-2505-6997

Рибчинський Максим Олександрович,
(аспірант ПВНЗ «Європейський університет»)
ORCID ID 0009-0001-9192-4115

АВТОМАТИЗАЦІЯ ОБРОБКИ ДАНИХ: ОСНОВНІ ІНСТРУМЕНТИ ТА МЕТОДИ

Анотація. *Стаття присвячена ключовій ролі обробки даних у сучасному світі, підкреслюючи її фундаментальні концепції, аналіз і значення збору та використання даних. Окрему увагу присвячено інноваційним технологіям, таким як штучний інтелект (ШІ) і машинне навчання, з'ясовуючи їхню роль в обробці даних. Розглянуто різні моделі машинного навчання та їх практичне застосування в аналізі даних. Крім того, дано визначення великих даних та підкреслено необхідність ефективної обробки значних обсягів даних, висвітлюючи такі інструменти, як Hadoop і Mapreduce для цієї мети. Окрім цього у статті розглянуто технологію Інтернету речей (IoT) і його роль у зборі даних, окреслюючи пов'язані з цим проблеми та перспективи на майбутнє.*

Ключові слова: *автоматизована обробка даних, інструменти, методи управління даними, системи.*

Постановка проблеми. Автоматизована обробка даних означає використання комп'ютерних систем і програмного забезпечення для ефективної обробки великих обсягів даних. Цей підхід мінімізує втручання людини, забезпечуючи швидше й точніше керування даними . У суч.асну цифрову епоху підприємства щодня створюють величезні обсяги даних. Управління цими даними вручну є недоцільним, що призводить до впровадження автоматизованих систем. Ці системи спрощують операції, підвищують точність і економлять час, що робить їх незамінними в різних галузях промисловості. Традиційні методи обробки текстової інформації, з використанням ручної праці, вже не відповідають сучасним вимогам щодо швидкості та точності виконання. Ручна обробка документів потребує значних людських ресурсів, часу та може призводити до помилок через людський фактор. В умовах, коли своєчасність та точність інформації є критичними, такі методи стають неефективними. в сучасних умовах стрімкого зростання обсягів інформації та підвищених вимог до її обробки, використання технологій штучного інтелекту стає необхідним для забезпечення ефективності та точності.

Аналіз останніх досліджень і публікацій. Питанням автоматизації обробки даних та використання технологій ШІ для обробки документів присвячена значна кількість робіт. Так, у роботі [1] розглядаються методології та виклики, пов'язані з впровадженням чат-ботів, що використовують технології NLP. У праці [2] описано концепцію технології RAG, яка поєднує пошук інформації та генерацію тексту. У дослідженнях [4] описується концепція технології RAG, що поєднує пошук інформації та генерацію тексту. Це дає змогу моделям LLM, бути більш ефективними у створенні узагальнених звітів з великих обсягів даних. Технологія RAG

може значно підвищити точність та релевантність згенерованих текстів, що є критично важливим для автоматизації створення звітів. Однак не достатньо дослідженими лишаються питання можливості використання зазначених технологій без доступу до мережі інтернет.

Мета статті. Метою дослідження є теоретичне узагальнення основних інструментів автоматизації обробки даних в сучасних умовах.

Виклад основного матеріалу дослідження. Автоматизована обробка даних (ADP) стосується використання комп'ютерних систем і програмного забезпечення для ефективної та точної обробки, організації та керування даними. Він охоплює широкий спектр діяльності, пов'язаної з обробкою даних у цифровому форматі.

Системи ADP розроблені для автоматизації та оптимізації завдань, пов'язаних із даними, зменшуючи потребу в ручному втручанні та мінімізуючи ризик дорогих помилок. Вони широко використовуються в різних галузях, включаючи фінанси, охорону здоров'я, виробництво та логістику .За даними McKinsey [4], близько 50% роботи можна автоматизувати – їхнє останнє опитування підтверджує, що майже 31% підприємств уже повністю автоматизували принаймні одну функцію.

У звіті Zapier [5] говориться, що 90% працівників інтелектуальної сфери підтверджують, що автоматизація покращила життя людей на робочому місці. Те саме опитування стверджує, що 88% власників малого бізнесу кажуть, що автоматизація дозволяє їхній компанії конкурувати з більшими компаніями. Завдяки автоматизації повторюваних і трудомістких завдань, пов'язаних із даними, таких як введення, перевірка та обробка даних, автоматизація даних звільняє співробітників від зосередження на більш стратегічних і доданих заходах, дозволяючи їм підвищити ефективність і заощадити час.

Автоматизація зменшує ризик людських помилок під час обробки даних, дозволяє швидше приймати рішення та підвищує точність даних.

Автоматизована обробка даних є ключовою для організацій, які хочуть ефективно керувати своїми даними та захищати їх.

Це зменшує ймовірність людських помилок у введенні та обробці даних, забезпечує кращу перевірку даних і може включати журнали аудиту та контроль доступу. Це також гарантує своєчасне оновлення, резервне копіювання та відновлення даних.

Розглянемо ключові методи обробки даних

Методи обробки даних необхідні для перетворення необроблених даних у значущу інформацію. Ці методи включають збір даних , очищення даних, перетворення даних і аналіз даних . Кожен крок має вирішальне значення для забезпечення точності та надійності оброблених даних.

- Збір даних. Першим кроком в обробці даних є збір даних із різних джерел. Це може включати бази даних, онлайн-форми, датчики тощо. Автоматизовані системи спрощують цей процес шляхом інтеграції з різними джерелами даних, забезпечуючи комплексний збір даних.
- Очищення даних: необроблені дані часто містять помилки, дублікати та невідповідності. Очищення даних включає виявлення та виправлення цих проблем, щоб забезпечити точність і надійність даних. Автоматизовані інструменти можуть ефективно впоратися з цим завданням, скоротивши час і зусилля, необхідні для ручного очищення.
- Перетворення даних: після очищення даних їх потрібно перетворити у відповідний формат для аналізу. Це може включати нормалізацію даних, агрегування даних і застосування різних перетворень, щоб підготувати їх до наступних кроків.
- Аналіз даних :Останнім кроком є аналіз даних для отримання значущої інформації. Автоматизовані системи обробки даних використовують розширені алгоритми та методи машинного навчання для швидкого й точного аналізу великих наборів даних (табл.1).

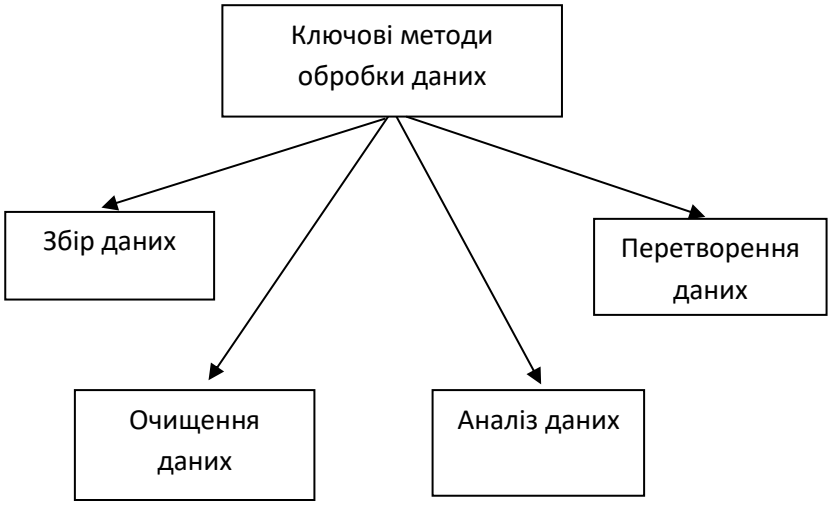

Рис.1 **Ключові методи обробки даних**

Джерело [5]

Для автоматизованої обробки даних доступні різні інструменти. Ці інструменти варіюються від програмних додатків до комплексних платформ, призначених для обробки різних аспектів обробки даних.

- Apache Hadoop: структура, яка дозволяє розподіллено обробляти великі набори даних між кластерами комп'ютерів за допомогою простих моделей програмування.
- Apache Spark: уніфікований механізм аналітики з відкритим кодом для обробки великих даних із вбудованими модулями для потокової передачі, SQL, машинного навчання та обробки графіків.
- Talend: надає платформу інтеграції даних з відкритим вихідним кодом, яка дозволяє користувачам підключати, перетворювати та керувати даними з різних джерел.
- RapidMiner: наукова платформа даних, яка забезпечує інтегроване середовище для підготовки даних, машинного навчання, глибокого навчання, аналізу тексту та прогнозної аналітики.
- KNIME: програмне забезпечення з відкритим вихідним кодом для створення наукових програм і служб даних, які використовуються для аналізу та інтерпретації даних.

Використовуючи ці інструменти, компанії можуть автоматизувати різні аспекти обробки даних, від збору даних до аналізу, забезпечуючи ефективне та точне керування даними.

Автоматизована обробка даних пропонує численні переваги, що робить її цінним активом для організацій у різних галузях.

- Ефективність: автоматизовані системи можуть обробляти великі обсяги даних набагато швидше, ніж ручні методи. Така ефективність дозволяє підприємствам обробляти більше даних за менший час, підвищуючи загальну продуктивність.
- Точність: зводячи до мінімуму втручання людини, автоматизована обробка даних зменшує ризик помилок. Ця точність гарантує, що дані є надійними та заслуговують довіри.
- Економія: автоматизація зменшує потребу в ручній праці, що призводить до значної економії коштів. Організації можуть ефективніше розподіляти ресурси, зосереджуючись на стратегічних ініціативах, а не на рутинних завданнях обробки даних.

• Масштабованість: автоматизовані системи обробки даних можуть легко масштабуватися для обробки зростаючих обсягів даних. У міру розширення бізнесу ці системи можуть витримувати збільшене навантаження даних без шкоди для продуктивності.

• Обробка в реальному часі: багато автоматизованих систем обробки даних пропонують можливості обробки в реальному часі. Це дає організаціям доступ до актуальної інформації та своєчасне прийняття рішень.

Ефективні методи обробки даних мають вирішальне значення для максимізації переваг автоматизації. Ці методи включають пакетну обробку, обробку в реальному часі та паралельну обробку.

• Пакетна обробка (передбачає обробку даних великими пакетами через заплановані проміжки часу. Цей метод підходить для завдань, які не вимагають миттєвих результатів, наприклад для створення звітів).

• Обробка в реальному часі (передбачає обробку даних у міру їх створення. Цей метод ідеально підходить для програм, які вимагають негайного зворотного зв'язку, таких як системи моніторингу та онлайн-транзакції).

• Паралельна обробка (передбачає поділ великих наборів даних на менші фрагменти та їхню одночасну обробку. Цей метод використовує потужність кількох процесорів для прискорення обробки даних).

Ефективна обробка даних залежить від різноманітних інструментів, призначених для оптимізації та автоматизації процесу. Ці інструменти допомагають ефективно керувати даними та маніпулювати ними, гарантуючи, що організації можуть швидко отримувати цінну інформацію. Серед поширених інструментів обробки даних:

• Apache Hadoop - фреймворк із відкритим кодом, який дозволяє розподіллено обробляти великі набори даних між кластерами комп'ютерів.

• Apache Spark - швидка кластерна обчислювальна система загального призначення, яка забезпечує можливості обробки даних у пам'яті.

• Microsoft Azure Data Factory - хмарна служба інтеграції даних, яка дозволяє створювати, планувати та керувати конвеєрами даних.

• Amazon Web Services (AWS) Glue - повністю керована служба вилучення, трансформації та завантаження (ETL), яка спрощує підготовку та завантаження даних для аналітики .

• Google Cloud Dataflow - повністю керована служба для виконання конвеєрів Apache Beam для потокової та пакетної обробки.

Автоматизована обробка даних означає використання програмного забезпечення та алгоритмів для виконання завдань, які в іншому випадку вимагали б ручного втручання. Такий підхід зменшує ризик людської помилки, підвищує ефективність і дозволяє швидше приймати рішення. Автоматизована обробка даних може включати:

• Приймання даних: автоматичний збір та імпорт даних із різних джерел у централізоване сховище.

• Перетворення даних: автоматичне перетворення вихідних даних у формат, придатний для аналізу.

• Аналіз даних: автоматичний аналіз даних для виявлення закономірностей, тенденцій і аномалій.

• Візуалізація даних: автоматичне створення візуальних представлень даних для сприяння розумінню та прийняттю рішень

Великі мовні моделі (LLM) є іще одним з найважливіших інструментів під час обробки природної мови (NLP), що дають змогу автоматично розуміти, аналізувати та генерувати текст на основі великої кількості вхідних даних. Великі мовні моделі (LLM) – це одна з

найшвидше розвиваючихся галузей штучного інтелекту. Це глибоко навчальні моделі, які здатні розуміти і генерувати текст, перекладати мови, писати різні типи контенту і навіть вести розмови, імітуючи людське спілкування. Ключові особливості LLM:

• Глибоке навчання (LLM навчаються на величезних наборах текстових даних, що дозволяє їм опанувати складні мовні паттерни і зв'язки).

• Здатність розуміти і генерувати текст (LLM можуть розуміти сенс тексту, перекладати його на інші мови, писати різні типи контенту, включаючи новини, статті, вірші і навіть коди програм).

• Масштабність (LLM працюють з величезними наборами даних, що дозволяє їм бути більш точними і ефективними в обробці інформації).

Сфери застосування LLM:

• Переклад мов (LLM значно покращили якість машинного перекладу, зробивши його більш точним і природним).

• Створення контенту (LLM можуть писати новини, статті, вірші, сценарії, рекламні тексти і інші типи контенту).

• Обробка природної мови (LLM використовуються в різних системах обробки природної мови, включаючи чат-боти, віртуальних асистентів, пошукові системи і інші).

• Навчання і освіта (LLM можуть допомагати у навчанні і освіті, надаючи інформацію, перекладаючи тексти, генеруючи питання і відповіді).

• Розробка програмного забезпечення (LLM можуть допомагати в розробці програмного забезпечення, генеруючи код і виправляючи помилки) [1].

Моделі, такі як GPT-3, BERT, Llama3 та інші, базуються на архітектурі трансформерів, що забезпечує ефективну обробку контексту та взаємодії між словами в тексті. Використання таких моделей значно покращує точність та релевантність результатів у різних завданнях, від машинного перекладу до створення узагальнених звітів [5]. Великі мовні моделі (LLM) здатні навчатися на великих обсягах даних, що дає змогу їм адаптуватися до різних контекстів і забезпечувати високу якість текстів.

Використовуючи ці інструменти та методи, організації можуть оптимізувати свої робочі процеси обробки даних і отримати максимальну користь від своїх даних. Для ефективної автоматизованої обробки даних організації повинні відповідати певним вимогам. Це включає забезпечення якості даних шляхом збору точних, повних і надійних даних. Високоякісні дані мають вирішальне значення для отримання значущої інформації. Крім того, організаціям необхідно впровадити надійні заходи безпеки даних , включаючи шифрування, контроль доступу та регулярні перевірки безпеки, щоб захистити дані від несанкціонованого доступу та злому.

Крім того, організації повинні зосередитися на інтеграції даних, щоб створити уніфіковане уявлення шляхом інтеграції даних із різних джерел. Це вимагає сумісності між різними системами та безперебійного потоку даних. Їм також потрібно вибрати правильні рішення для зберігання даних, такі як хмарне сховище, сховища даних і бази даних для розміщення великих обсягів даних. Нарешті, дотримання правових і нормативних вимог, пов’язаних з обробкою даних, зокрема законів про конфіденційність даних, галузевих стандартів і організаційної політики, має важливе значення.

Майбутнє автоматизованої обробки даних виглядає багатообіцяючим, з кількома тенденціями, які формують галузь:

• Інтеграція штучного інтелекту та машинного навчання: інтеграція штучного інтелекту та машинного навчання з автоматизованою обробкою даних розширить можливості аналізу даних , дозволяючи точніше прогнозувати та отримувати знання.

- Аналітика великих даних: зростаючий обсяг даних підштовхне попит на передові аналітичні інструменти та методи обробки й аналізу великих даних.
- Автоматизація Data Science: автоматизація Data Science оптимізує робочі процеси та скоротить час, необхідний для аналізу даних.
- Етика даних: зростаюча важливість етики даних призведе до розробки рамок і вказівок для забезпечення етичної практики обробки даних.[2].

Наведені у статті технології надають потужні інструменти для автоматизації обробки текстових даних та створення звітів. Однак, для оцінювання їхньої ефективності та виявлення практичних переваг і недоліків, потрібно провести низку експериментальних досліджень.

Висновки та пропозиції: Автоматизована обробка даних є важливою складовою сучасного управління даними. Він пропонує численні переваги, зокрема ефективність, точність і економію коштів. Застосовуючи передові методи та інструменти обробки даних, організації можуть покращити свої можливості обробки даних і досягти кращих результатів. Однак важливо вирішити проблеми, пов'язані з автоматизованою обробкою даних, щоб повністю реалізувати її потенціал.

Оскільки технології продовжують розвиватися, майбутнє автоматизованої обробки даних виглядає яскравим. Завдяки інтеграції штучного інтелекту та машинного навчання компанії можуть розраховувати на ще більш досконалі рішення для обробки даних, які стимулюватимуть інновації та зростання.

ЛІТЕРАТУРА

1.Lin C. C., Huang A. Y. Q., Yang S. J. H. A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). Sustainability 2023, 15(5). DOI: https://doi.org/10.3390/su15054012.

2.Wyndham, A., (2024). 10 Large Language Models That Matter to the Language Industry. Data & Indexes, Technology [online]. Avaliable at: https://slator.com/10-large-language-models-that-matter-to-the-language-industry [Accessed: 29 July 2024].

3.Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., Cui, B., (2024). Retrieval-Augmented Generation for AI-Generated Content: A Surve. Arxiv[online]. Avaliable at: https://arxiv.org/abs/2402.19473 [Accessed: 29 July 2024].

4.Чала О. О. Фізико технологічна база для побудови математичної моделі прогнозування дефектв у підкладках функціональних компонентів МОЕМС /О. О. Чала, І. Ш. Невлюдов, В. В. Невлюдова //VII Міжнародна науково-практична конференція «Напівпровідникові матеріали, інформаційні технології та фотовольтаїка»:Тези доповідей. − Кременчук: Кременчуцький національний університет імені Михайла Остроградського, 2022. – С. 32-33

5.Чала О. О. Дефектоутворення, як основа Defect Engineering в МЕМС таМОЕМС // Технология приборостоения. 2020. № 1. С. 78–81.

6.Nevlyudov Igor Structural diagram of automated quality control process of silicon wafers during their surface shaping / I. Nevlyudov, I Botsman, S. Tesliuk // The V International Science Conference «Theoretical and scientific bases of development of scientific thought»,February 16 – 19, 2021, Rome, Italy. PP 612-615.

7.McKinsey [A FUTURE THAT WORKS: AUTOMATION, EMPLOYMENT, AND PRODUCTIVITY https://www.mckinsey.com/~/media/mckinsey/featured%20insights/Digital%20Disruption/Harnessig%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works-Executive-summary.ashx

8.Zapier report: The 2021 state of business automation https://zapier.com/blog/state-of-business-automation-2021/

REFERENCES

1.Lin C. C., Huang A. Y. Q., Yang S. J. H. A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). Sustainability 2023, 15(5). DOI: https://doi.org/10.3390/su15054012.

2.Wyndham, A., (2024). 10 Large Language Models That Matter to the Language Industry. Data & Indexes, Technology [online]. Avaliable at: https://slator.com/10-large-language-models-that-matter-to-the-language-industry [Accessed: 29 July 2024].

3.Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., Cui, B., (2024). Retrieval-Augmented Generation for AI-Generated Content: A Surve. Arxiv[online]. Avaliable at: https://arxiv.org/abs/2402.19473[Accessed: 29 July 2024].

4.Chala O. O. Fizyko tekhnolohichna baza dlia pobudovy matematychnoi modeli prohnozuvannia defektv u pidkladkakh funktsionalnykh komponentiv MOEMS /O. O. Chala, I. Sh. Nevliudov, V. V. Nevliudova // VII Mizhnarodna naukovo-praktychna konferentsiia «Napivprovidnykovi materialy, informatsiini tekhnolohii ta fotovoltaika»:Tezy dopovidei. − Kremenchuk: Kremenchutskyi natsionalnyi universytet imeni Mykhaila Ostrohradskoho, 2022. – S. 32-33

5.Chala O. O. Defektoutvorennia, yak osnova Defect Engineering v MEMS taMOEMS // Tekhnolohyia pryborostoenyia. 2020. № 1. S. 78–81.

6.Nevlyudov Igor Structural diagram of automated quality control process of silicon wafers during their surface shaping / I. Nevlyudov, I Botsman, S. Tesliuk // The V International Science Conference «Theoretical and scientific bases of development of scientific thought»,February 16 – 19, 2021, Rome, Italy. PP 612-615.

7.McKinsey [A FUTURE THAT WORKS: AUTOMATION, EMPLOYMENT, AND PRODUCTIVITY https://www.mckinsey.com/~/media/mckinsey/featured%20insights/Digital%20Disruption/Harnessig%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works-Executive-summary.ashx

8.Zapier report: The 2021 state of business automation https https://zapier.com/blog/state-of-business-automation-2021

Oleg Ivanovich Veliyka,
(postgraduate student of PVNZ "European University")
ORCID ID 0009-0001-2505-6997

Maksym Oleksandrovych Rybchinsky,
(postgraduate student of PVNZ "European University")
ORCID ID 0009-0001-9192-4115

TITLE OF THE ARTICLE

The article focuses on the key role of data processing in today's world, emphasizing its fundamental concepts, analysis and the importance of data collection and use. Particular attention is paid to innovative technologies such as artificial intelligence (AI) and machine learning, clarifying their role in data processing. Various models of machine learning and their practical application in data analysis are considered. In addition, the definition of big data is given and the need to efficiently process large amounts of data is highlighted, highlighting tools such as Hadoop and Mapreduce for this purpose. In addition, the article examines the Internet of Things (IoT) technology and its role in data collection, outlining the related challenges and future prospects.

Keywords*: automated data processing, tools, data management methods, systems.*