УДК 004.6
DOI https://doi.org/10.32838/2663-5941/2021.5/13

***Захарчин Н.Г.***
Національний лісотехнічний університет України

***Захарчин Н.Р.***
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

# РІСТ СТРУКТУРОВАНИХ ТА НЕСТРУКТУРОВАНИХ ДАНИХ ТА УПРАВЛІННЯ НИМИ: ЗАГАЛЬНІ АСПЕКТИ

*Зважаючи на інтенсивний розвиток людства в інформаційну еру, стає зрозуміло, що перед великими компаніями, установами й підприємствами постає питання важливості ефективного оперування великими даними (Big Data), до яких вони мають доступ. Попри те, що правильне використання таких даних є надзвичайно важливим завданням, воно також є одним із найважчих у контексті роботи з великими даними у зв'язку з динамікою та постійною мінливістю даних, їх неоднозначністю та різноманіттям. У статті розглядаються основні проблеми, що виникають під час роботи зі структурованими й неструктурованими даними на тлі їх стрімкого глобального зростання. Наведено й детально проаналізовано статистику збільшення обсягів і масштабності даних у світі за минулі роки, а також передбачення та прогнози, побудовані на її основі, щодо майбутніх тенденцій росту інформації. Ґрунтуючись на раніше проведених дослідженнях, зроблено висновки про неминуче стрімке зростання кількостей даних і в наступні роки. У статті докладно розглянуто поняття структурованих і неструктурованих даних, їх характеристики, переваги й недоліки обох видів, їх практичне застосування. Представлено найбільш вживані методи й інструменти для роботи з великими даними, акцентовано на основних випадках їх використання. Важливим моментом дослідження, який вартий уваги в ракурсі практичного застосування, є різниця у відповідності того чи іншого інструменту до опрацювання різних видів даних. Так, зазвичай засоби, котрі призначені для структурованих даних, не можуть застосовуватися під час роботи з неструктурованими. Кількість неструктурованих даних, якими Інтернет-користувачі оперують у сучасному світі, значно перевищує кількість структурованих, окрім того, більше користі суспільству, підприємництву й бізнесу потенційно здатні принести саме неструктуровані дані завдяки своїй багатоформатності. Саме через це в статті наголошено на тому, що великі дані, більшою мірою неструктуровані, потребують пошуку й знаходження нових результативних підходів, які б дозволяли суб'єктам підприємництва отримувати більше користі з тієї ж кількості інформації.*

***Ключові слова:*** *великі дані, бази даних, зростання даних, підприємство, Інтернет речей.*

**Постановка проблеми.** XXI століття – час постійних змін, інновацій і розвитку. «Майбутнє близько» – такі слова найкраще схарактеризують буремну добу сьогодення. Технології розвиваються не по роках, а по місяцях, щодня відбувається новий прорив у галузях науки й техніки. Тому абсолютно логічним наслідком таких факторів є величезна кількість інформації, яка зберігається, обробляється, переглядається та редагується людством щодня. Інформація існує в різноманітних формах і проявах – інтернет-сайти, книги, статті, фільми, сторінки в соцмережах тощо. Однак найбільшою проблемою є те, що можливості людини щодо сприйняття та обробки інформації були й залишаються досить вузькими, оскільки обсяги даних у світі ростуть, але людство не встигає досягти відповідного рівня прогресу, щоб це усвідомити й опрацювати. Це своєю чергою ставить завдання теоретичної систематизації обробки поточної інформації.

**Аналіз останніх досліджень і публікацій.** Питання функціонування та використання структурованих і неструктурованих даних досліджує А.С. Еберенду (Madonna University, Nigeria) у своїй праці "Unstructured Data: an overview of the data of Big Data" [1]. У роботі, зокрема, зазначається, що сучасні інституції намагаються скористатись можливостями використання неструктурованих даних, аналізуючи й досліджуючи їх. Авторка констатує, що збір та аналіз неструктурованих даних дає підприємствам актуальну інформацію про свою діяльність, допомагає бути інноваційними, конкурентоспроможними, підвищувати продуктивність.

В українській науці окремі елементи аналізу структурованих даних містяться в роботі

А.Ю. Яцишина [2]. Він розглядає питання проєктування гібридних сховищ даних з урахуванням структурованості даних, одночасно вводячи поняття мультибазових сховищ даних як розширення гібридних сховищ даних. Водночас слід зазначити, що досліджень, які розкривають проблеми росту даних, не достатньо.

**Постановка завдання.** Метою роботи є проведення аналізу й оцінки стану розвитку структурованих і неструктурованих даних і наявних на ринку інструментів управління ними.

Для досягнення зазначеної мети визначено такі основні завдання дослідження: розглянути рівень зростання великих даних у світі й навести прогнози щодо цього рівня в майбутньому, викласти основні характеристики й проаналізувати переваги й недоліки структурованих і неструктурованих даних, порівняти зручність роботи з ними й наявність процесуальних інструментів.

Практична значущість результатів дослідження полягає в можливостях для великих компаній систематизації обробки поточної інформації з різних джерел, що постійно доповнюються та змінюються, за допомогою описаних у статті інструментів.

**Виклад основного матеріалу дослідження.**

**Ріст даних у світі.** Британський математик та аналітик Клайв Гамбі у 2006 році зазначав: «Дані – це нова нафта». Водночас всесвітня датасфера (*англ.* datasphere) – загальний обсяг даних, створених, фіксованих і відтворених у світі – дорівнював приблизно 160 ексабайт. Отже, дані, зростаючи експоненційно, стали на рівні з нафтою одним із найцінніших світових ресурсів [3].

На початку минулого десятиліття міжнародна компанія-постачальник маркетингових досліджень IDC (International Data Corporation) оцінила обсяг нових даних, згенерованих у 2010-му році, у 1.2 зетабайт (1,2 трильйона гігабайт), причому у 2009-му році ця цифра становила 0.8 зетабайт. Передбачалося, що у 2020-му році обсяг новостворених даних зросте до 35 зетабайт. А прогноз на 2025-й рік становить аж 175 зетабайт [4]. 49% цього обсягу становитимуть дані в хмарі, і 90 зетабайт із всього обсягу становитимуть дані, генеровані пристроями Інтернету речей [5].

Зростання даних у світі показано на рис. 1.

Погляньмо ближче на статистику щодо росту даних. 2020-й рік завдяки пандемії COVID-19 зробив значний внесок у цифровізацію світу, роблячи технології невід'ємною частиною щоденного життя. Поглянувши на дані, маємо можливість глибше усвідомити швидкість, з якою розвивається суспільство. Так, дослідження "Data never sleeps 8.0" сервісу Domo надає статистику щодо даних, які генеруються у світі за хвилину [5]. Його результати показали, що кожної хвилини протягом 2020-го року користувачі Facebook обмінювались 150 000-ма повідомленнями, клієнти Amazon замовляли 6 659 посилок, а загальна сума, витрачена на покупки в інтернеті за ту ж хвилину, становить 1 млн доларів [5]. 5 мільярдів людей кожного дня мають справу з даними. До 2025-го року ця кількість збільшиться до 6 мільярдів і становитиме 75% від усього населення планети. Немало таких взаємодій відбувається завдяки пристроям Інтернету речей [5].

За прогнозами, подальший ріст даних буде спричиненим саме завдяки вбудованим пристроям (embedded devices). Людство все більше й більше залежить від розумних девайсів: смартквартири, розумні годинники, фітнес-браслети, камери спостереження, банківські картки тощо. Незважаючи на те, що кожна окрема операція, виконана будь-яким вбудованим пристроєм, несе дуже мало інформації, цих операцій – безліч, а отже, інформація генерується швидко й у всій можливій різноманітності. Прогнози повідомляють, що до 2025-го року дані, генеровані вбудованими пристроями, становитимуть 20% від усіх даних [8].

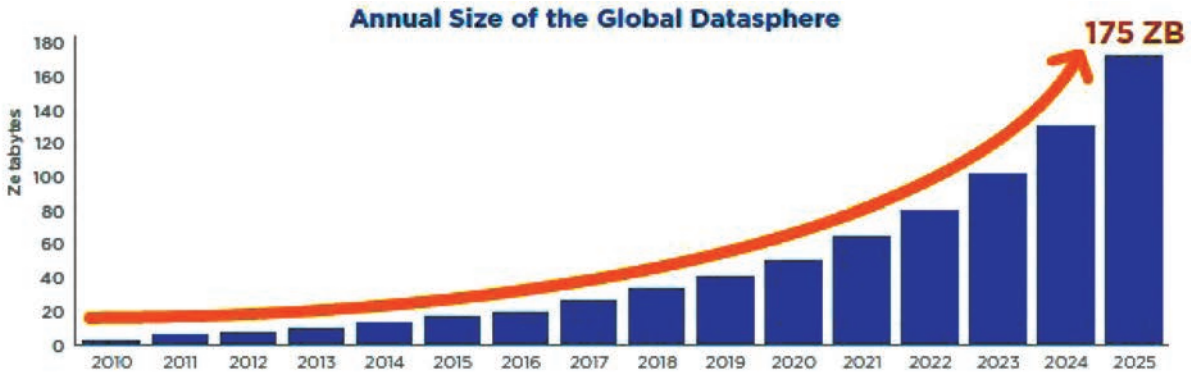


**Рис. 1. Зростання даних у світі згідно з дослідженням IDC DataAge 2025 [6]**

Поняття, які використовуються для класифікації великих даних – це структуровані й неструктуровані дані.

Структуровані дані – це дані, які відповідають раніше створеним моделям і завдяки цьому легко піддаються аналізу. Їх можна представити в табличному форматі з відношеннями між різними рядками й колонками. До прикладу, файли Excel або дані, які можна внести в бази даних SQL. Отже, структуровані дані залежать від існування певної моделі, за допомогою якої їх можна зберегти й обробити [9].

Основними перевагами структурованих даних є:

1. Легкі під час використання алгоритмами машинного навчання. Організованість структурованих даних дозволяє легко керувати ними й добувати їх.

2. Легкі під час використання бізнес-користувачами (клієнтами). Пересічний користувач має можливість працювати зі структурованими даними з розумінням теми, до якої вони належать: немає потреби глибоко розуміти типи даних чи їх взаємозв'язок.

3. Доступ до більшої кількості інструментів. Історично склалося, що структуровані дані використовувались набагато довше, ніж неструктуровані, адже це був єдиний можливий варіант. Отже, існує більше інструментів для роботи зі структурованими даними, випробуваними часом.

Недоліки структурованих даних зосереджені переважно на тому, що їм бракує гнучкості [10]. Так, зазвичай, структуровані дані – це стисла інформація, організована в суворий спосіб; їх можна уявити у вигляді таблиці, де кожна клітинка містить певне значення. Дані з такою заздалегідь визначеною структурою найчастіше можуть використовуватись лише в один спосіб.

Неструктуровані дані – це інформація, яка не має наперед визначеної моделі або ж не організована в наперед визначений спосіб. Кажучи простішими словами, це абсолютна більшість інформації, з якою людям трапляється працювати [9]. Типові приклади неструктурованих даних:

– текстові файли (файли Word, електронні листи, презентації тощо);
– дані із соціальних мереж;
– вебсайти;
– мобільні дані (текстові повідомлення, локації);
– медіа (цифрові фотографії, аудіо- й відеофайли).

Плюси неструктурованих даних:

1. Довільний первинний формат. Оскільки неструктуровані дані зберігаються у своєму власному форматі, вони спеціально не визначаються, доки не виникне потреба. Це приводить до більшої кількості варіантів їх використання (можна оперувати даними в будь-який спосіб і з будь-якою метою). Також завдяки цьому аналітики можуть обробляти лише необхідні частини всієї інформації.

2. Швидкість накопичення. Немає потреби заздалегідь визначати дані, а отже, збирати їх можна швидко й легко.

3. Зберігання в хмарних озерах даних (Cloud data lakes). Хмарні озера даних дозволяють масово зберігати неструктуровані дані, а також визначають ціни за зберігання як «плата лише в разі безпосереднього використання», що допомагає скоротити витрати [10].

Найбільшим недоліком неструктурованих даних є те, що для їх аналізу потрібна попередня підготовка в цій галузі. Так, звичайний користувач не може працювати з неструктурованими даними в їх первинному вигляді через їх невизначену природу. Використання неструктурованих даних вимагає розуміння теми, а також того, як дані пов'язані, і того, як використати ці зв'язки на благо компанії [10]. Крім необхідності попередньої підготовки, неструктуровані дані вимагають ще й наявності спеціальних інструментів для роботи з ними.

Окрім очевидної різниці в зберіганні структурованих і неструктурованих даних (у реляційній базі даних і поза нею відповідно), основною відмінністю між ними є міра складності аналізу цих видів даних. Як вже було зазначено, аналіз структурованих даних є процесом, який легко зреалізувати, для цього існують дієві інструменти.

А як щодо неструктурованих даних? Так, за потреби користувачі можуть проводити в певному сенсі автоматизований аналіз цих даних, наприклад, виконувати пошук вмісту в текстових файлах. Проте цього занадто мало для того, щоб переважити відсутність впорядкованої внутрішньої моделі даних. На жаль, компанії отримують не так багато користі від потенційно важливої інформації, наприклад, даних із соцмереж, блогів, взаємодії з клієнтами. Незважаючи на те, що засоби для аналітики неструктурованих даних уже присутні на ринку, жоден із них не є явним лідером, здатним спричинити прорив у роботі з ними. Прикладами вже наявних інструментів є MongoDB – документо-орієнтована система керування базами даних, оптимізована для зберігання документів, і Apache Giraph, що обробляє великі дані за допомогою графів.

Крім того, кількість неструктурованих даних у світі становить 80% і значно переважає кількість структурованих [9]. Це і є найбільшою проблемою неструктурованих даних – вони є важливим матеріалом для компаній, щоб будувати на їх базі нові оптимальні рішення для бізнесу. А отже, без відповідних інструментів для роботи з неструктурованими даними підприємства позбавлені величезних можливостей.

Великі дані залучають ціле різноманіття інструментів, технік і фреймворків для роботи. Для інструментів зберігання та обробки даних вимагаються такі характеристики, як масштабованість і наявність швидкісного доступу до величезних обсягів інформації. Розглянемо основні засоби, в яких наявні названі характеристики.

Першим інструментом для зберігання великих даних є реляційні бази даних, точніше, системи керування базами даних (англ. DBMS, Database Management System). Усі вони базуються на використанні мови SQL. Значна частина реляційних баз даних, призначених для великих даних, називається аналітичними базами даних масової паралельної обробки. Вони здатні швидко обробляти масивні обсяги даних (переважно структурованих) із мінімальними вимогами щодо моделювання, крім того, можуть масштабуватися до розміру багатьох терабайтів і петабайтів даних [11]. Реляційні бази даних використовуються для зберігання та обробки лише структурованих даних, тобто їх застосування доволі обмежене. Прикладами таких інструментів є системи MySQL, Microsoft SQL Server та інші.

Іншим підходом до зберігання великих даних є використання розподілених файлових систем – систем, що зберігають дані на великій кількості серверів. Вони є одним з інструментів для роботи з неструктурованими даними. Розподілені файлові системи дозволяють програмам отримувати доступ до відокремлених, ізольованих файлів так само, як це робиться з локальними, надаючи таким чином програмістам можливість працювати з файлами з будь-якого комп'ютера чи мережі. Найпопулярнішим серед розподілених файлових систем є Hadoop. Кластери Hadoop здатні масштабуватися до великих розмірів – петабайт і навіть ексабайт даних, тому компанії можуть не обмежуватися опрацюванням лише вибіркових наборів даних [11]. Також перевагою розподілених файлових систем є здатність забезпечувати прозорість і видимість даних навіть за несправності сервера чи диску.

Третім типом інструментів для роботи з даними є бази даних NoSQL – тип баз даних, розроблений інтернет-компаніями на початках 2000-х років. Цей тип не протилежний до реляційних баз на основі SQL, а радше доповнює їх, уможливлюючи співіснування. Типовими характеристиками баз даних NoSQL є спрощування моделей даних і відсутність стандартної мови запитів [11]. На відміну від реляційних баз даних бази даних NoSQL надають широкий спектр моделей, за допомогою яких можна зберігати дані, – немає визначеної структури, яка вимагається. Крім того, ці бази значно швидші й гнучкіші у використанні, що робить їх конкурентною опцією для роботи з неструктурованими даними [12].

Існує ще кілька можливостей для зберігання та обробки великих даних. Використання тої чи іншої залежить від програми, обсягу даних, які використовуватиме програма, складності алгоритмів майнінгу тощо.

**Висновки.** Незважаючи на наявність досить ефективних засобів для роботи з неструктурованими даними на ринку, їх недостатньо для рівнозначного зіставлення зі швидкістю прогресу людства в генеруванні даних. Саме тому для компаній, підприємств та урядів надзвичайно важливим завданням є пошук ефективних знарядь для зберігання та опрацювання великих даних, особливо неструктурованих, добуваючи з них максимальну користь для себе, наприклад, вподобання користувачів, статистику замовлень, відгуки із соцмереж тощо. Крім того, проблемою великих даних є їх стрімкий ріст. Загалом результати дослідження показують, що існує потреба в пошуку нових інструментів для ефективної роботи з неструктурованими даними.

**Список літератури:**

1. Eberendu A.C. Unstructured Data: an overview of the data of Big Data. *International Journal of Computer Trends and Technology*. 2016. No. 38 (1). P. 46–50.
2. Яцишин А.Ю. Проектування гібридних сховищ даних з врахуванням структурованості даних. *Управління розвитком складних систем.* 2012. Вип. 9. С. 59–65.
3. Greaton T. What's causing the exponential growth of data? 2019. URL: https://insights.nikkoam.com/articles/2019/12/whats_causing_the_exponential.
4. Press G. 6 Predictions About Data In 2020 And The Coming Decade. 2020. URL: https://www.forbes.com/sites/gilpress/2020/01/06/6-predictions-about-data-in-2020-and-the-coming-decade/?sh=3e23597a4fc3.

5. Data Age 2025. The digitization of the world. URL: https://www.seagate.com/gb/en/our-story/data-age-2025/.

6. Reinsel D., Gantz J., Rydning J. The Digitization of the World From Edge to Core. 2018. URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

7. Data Never Sleeps 8.0. URL: https://www.domo.com/learn/data-never-sleeps-8.

8. Enormous Growth in Data is Coming — How to Prepare for It, and Prosper From It. URL: https://blog.seagate.com/business/enormous-growth-in-data-is-coming-how-to-prepare-for-it-and-prosper-from-it/.

9. Data types: Structured vs. Unstructured Data. URL: https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/.

10. Structured vs. Unstructured Data: A Complete Guide. URL: https://www.talend.com/resources/structured-vs-unstructured-data/.

11. Pokorny J. How to Store and Process Big Data: Are Today's Databases Sufficient? *13th IFIP International Conference on Computer Information Systems and Industrial Management (CISIM)*, Ho Chi Minh City, Vietnam, 5–7 November, 2014. Springer, Berlin, Heidelberg, 2014. P. 5–10. https://doi.org/10.1007/978-3-662-45237-0_2.

12. Nayak A., Poriya A., Poojary D. Type of NoSQL Databases and its Comparison with Relational Databases. *International Journal of Applied Information Systems (IJAIS)*. 2013. Vol. 5. No. 4. P. 16–19.

**Zakharchyn N.H., Zakharchyn N.R. GROWTH AND MANAGEMENT OF STRUCTURED AND UNSTRUCTURED DATA: GENERAL ASPECTS**

*In the context of the rapid growth of humanity in the era of information, big companies, institutions, and enterprises are facing the issue of the importance of the efficient handling of big data that is at their disposal. Proper use of this data is crucial as well as a complicated task due to the constant dynamics and variability of data, and its uncertainties. This article discusses the main problems that occur while working with structured and unstructured data within the context of its rapid global growth. The article provides and analyzes statistics of data volume growth in recent years, and some predictions and forecasts based on them, regarding information growth in the next years. Taking into account the research that was previously held, it is concluded that it is inevitable for data to continue to grow rapidly in the next years. Moreover, the article examines the terms of structured and unstructured data, their characteristics, advantages, and disadvantages of both types, and their practical use. This study also introduces some common methods and tools to work with big data, and pays attention to the main instances of their use. The significant moment of the study that is worth mentioning from the perspective of practical use is a difference in whether or not one or another tool is pertinent in the context of working with these types of big data. Tools designed for structured data usually cannot be applied to unstructured data. The volume of unstructured data which exists in the modern world highly exceeds the volume of structured data. Apart from that, unstructured data is capable of bringing great benefits to society, entrepreneurship, and businesses owing to its multiformity. Consequently, the article emphasizes that big data, especially unstructured data, requires seeking and finding new effective approaches that would enable the individuals of entrepreneurship to benefit more from the same amount of information.*

***Key words:*** *big data, databases, data growth, entrepreneur, Internet of Things.*