Issue 14, 2023

https://doi.org/10.23939/sisn2023.14.373

УДК 004: 378

АРХІТЕКТУРА СИСТЕМ ОЗЕР ДАНИХ У ГАЛУЗІ ОСВІТИ: АНАЛІТИЧНИЙ ОГЛЯД

Андрій Пришляк[1], Микола Орлов[2]

Національний університет "Львівська політехніка", кафедра інформаційних систем та мереж,
вул. С. Бандери, 12, Львів, Україна,
[1] E-mail: andrii.a.pryshliak@lpnu.ua; ORCID: 0000-0003-1681-5178
[2] E-mail: orlov.nv.86@gmail.com; ORCID: 0009-0007-4201-1188

Здійснено аналітичний огляд розвитку концепту Data Lakes та його застосування у різних галузях як частини рішень концепту Big data. Розглянуто наявні стандартні архітектурні рішення для організації Data Lake. Також взято до уваги спеціалізовані напрями, що потребують відмінних чи додаткових аспектів для вирішення поставлених завдань, залежно від галузі використання Data Lake. Для правильної організації Data Lake застосовують різноманітні засоби опрацювання даних, зокрема розподілені системи зберігання даних, семантичні мережі та особливо метадані. Метадані відіграють величезну роль у розпізнаванні призначення даних та можливих зв'язків між ними та сутностями. Проаналізовано перспективи застосування Data Lake, зокрема в контексті розумного міста, дистанційної освіти та освітньої галузі загалом.

Ключові слова: великі дані; озеро даних; метадані; дистанційне навчання; онлайн-освіта; семантичні мережі; добування даних; розумне місто.

Вступ

Концепт Data Lake (озеро даних) належить до архітектури та підходу до зберігання та опрацювання великих обсягів даних. Він пропонує централізоване зберігання різноманітних даних у неструктурованому або напівструктурованому вигляді.

Основна ідея Data Lake полягає у тому, щоб зібрати всі можливі джерела даних, такі як бази даних, журнали, файлові системи, сенсори, соціальні медіа тощо, у єдиному місці без необхідності стандартизації або опрацювання даних заздалегідь. Замість цього дані зберігають в їх первинному форматі, а також всю супутню контекстну інформацію.

Data Lake використовує розподілені системи зберігання, такі як Hadoop Distributed File System (HDFS) або cloud-платформи, для забезпечення масштабованості та надійності зберігання даних, що дає змогу використовувати паралельне опрацювання та аналіз для розуміння даних та отримання відповідних результатів.

Мета статті є аналіз концепту "озеро даних" та архітектур побудови інформаційних систем з його використанням.

Завдання дослідження – опрацювання варіантів формування рівня метаданих та конструктивів формування сховищ даних, джерелом для яких слугують озера даних, розгляд сфер застосування архітектур озер даних для побудови інформаційних систем освітньої галузі.

Поняття озера даних

Озера даних – доволі недосліджена сфера для багатьох рішень аналізу даних, вони стають все популярнішими в дослідженнях. Їх часто пов'язують із випадками опрацювання великих даних, використовують, наприклад, як центральні системи керування даними дослідницьких установ або як основну сутність процесів машинного навчання. Основна ідея збереження даних у їх рідному форматі в озері даних полегшує широкий спектр процедур опрацювання та поліпшує повторне використання даних. Однак зберігання таких величезних обсягів неопрацьованих даних створює певні проблеми, починаючи від загального моделювання даних та індексування для стислих запитів до інтеграції відповідних і масштабованих обчислювальних можливостей [1].

Озеро розглядається насамперед як певний репозиторій для подальшого аналізу даних. Відповідно для отримання потрібних даних використовують алгоритми кластеризації, розпізнавання сутностей та зв'язків для кращого розуміння схематичності озера даних [2]. Модель роботи із сутностями подано на рис. 1.

Далі буде розглянуто спосіб подання озер за допомогою семантичних мереж із застосуванням багатошарових графів та навігаційного графа-моделі, що сприяють роботі із розпізнавання та зіставлення даних.

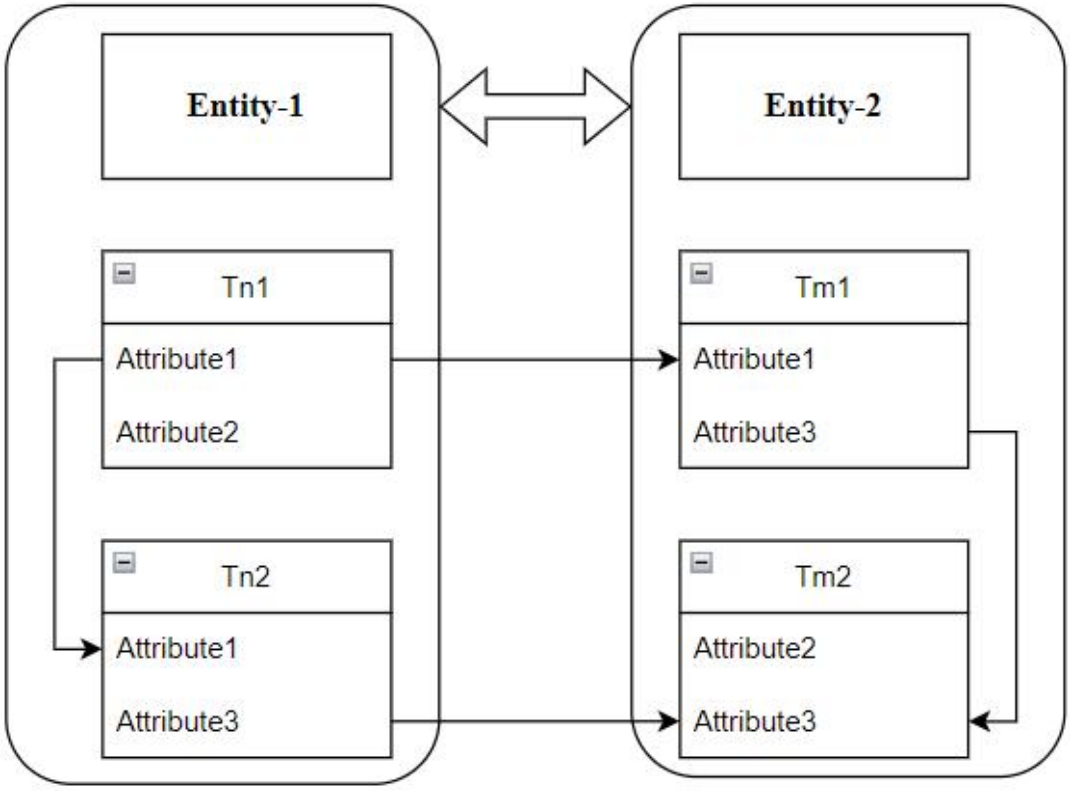

Рис. 1. Модель процесу розпізнавання сутностей

Архітектура озера даних описує структуру та компоненти системи, що вказує, як зберігати, упорядковувати та використовувати дані. Існує декілька варіацій архітектури озера даних [3]:

· Архітектура резервуару розподіляє прийняті дані за їхнім статусом і використанням. Спочатку дані зберігаються у резервуарі неопрацьованих даних, потім перетворюються та переміщуються до відповідних резервуарів за типами даних. Крім того, відповідні процеси застосовуються для підготовки даних для подальшого аналітичного опрацювання. Наприклад, аналогові дані, створені автоматизованим пристроєм, переміщуються до резервуара аналогових даних. Потім обсяг

аналогових даних регулюється до можливого розміру (зменшення даних). Ставок архівних даних зберігає невикористані дані.

· Архітектура зон, яка розділяє етапи опрацювання кожного набору даних на різні етапи. Наприклад, можуть бути окремі зони для завантаження даних і перевірки їх якості, зберігання неопрацьованих даних, зберігання очищених і підтверджених даних, виявлення та вивчення даних або використання даних для аналізу.

· Загальна архітектура містить чотири рівні: приймання даних, зберігання, опрацювання та доступ.

Запропоновано модель, яка забезпечує трирівневий функціонально-орієнтований зв'язок [25], схематично зображену на рис. 2.

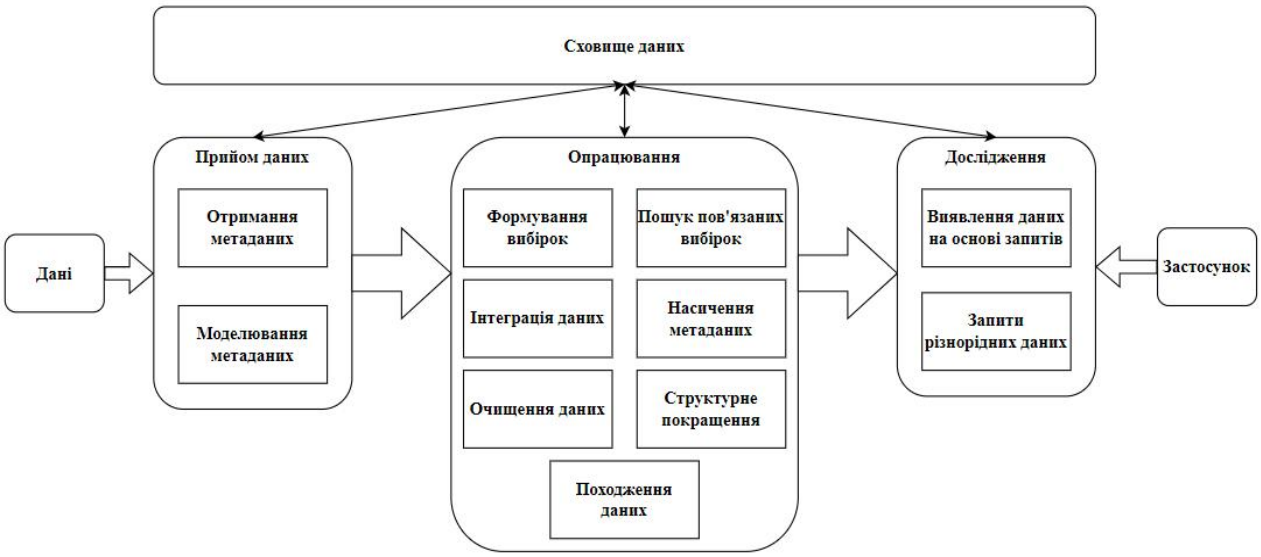

Рис. 2. Модель архітектури трирівневого функціонально-орієнтованого зв'язку

Оскільки озера даних постійно наповнюються, то у них можуть з'являтись і нові залежності між даними, які раніше не було можливості пов'язати. Також за рахунок великої кількості джерел з'являтимуться суперечливі або не зовсім однозначні дані щодо певного об'єкта, які не можна просто так об'єднати чи надати перевагу тільки частковим показникам. У такому випадку пропонують використовувати онтології та семантичні правила, для вирішення проблем неоднорідності й чіткого розуміння тих чи інших даних [4].

Метадані як невід'ємна частина озера даних

Озера даних зберігають дані та інформацію про процеси, які над цими даними виконувались. Оскільки дані можуть надходити із різних джерел, метадані відіграють надзвичайно важливу роль у характеристиці щодо різноманітності та неоднорідності наборів даних. Метадані передбачають роботу з даними та об'єктами різних рівнів деталізації. Деталізація істотно пов'язана із концепцією озер даних, найчастіше в аспектах розпізнавання даних різних сутностей [5].

Метадані самі по собі є інформацією щодо даних та процесів, які збирає озеро даних, і для них потрібні окремі механізми керування. За останні роки запропоновано декілька таких механізмів, які орієнтуються на категоризацію або список функцій керування метаданими, або ж їх поєднання. Але цих підходів чи їх комплексу не зажди достатньо, тому науковці запропонували нову модель, яка орієнтується на отримання метаданих на різних рівнях деталізації, щодо будь-якої категоризації, як для певних специфічних метаданих, так і для загальніших випадків [6].

Певні метадані за правильних юридичних аспектів дослідники можуть застосувати для специфічних цілей. Тобто сумісні набори таких даних можна зіставляти і компонувати, зважаючи на юридичну складову [7].

Для отримання даних із озера пропонують різні підходи із застосуванням семантичних типів даних [8]:

· На основі пошуку – із використанням зовнішньої інформації для "опису" семантичної інформації щодо наборів даних.

· На основі навчання – спрямовані на побудову моделі машинного навчання, яка зможе виводити семантичні типи стовпців.

Підготовка даних для аналізу, як правило, потребує затрат часу, особливо в озерах даних, де кількість типів даних завжди дуже велика. Для зменшення трудових затрат пропонують структуру, яка поєднуватиме модель метаданих і операції алгебраїчного перетворення: фільтрацію, форматування, агрегацію, обчислення, консолідацію, компонування та об'єднання [9].

Робота з даними

Модель озер даних зосереджена на тому, що потрібно зберігати інформацію, а не на тому, які власне дані потрібні чи яке їх призначення. Тому можна доповнити озера семантичною інформацією, яка б формувала правила підбору даних. Для цього використовують семантичні мережі із застосуванням багатошарових графів [10].

Ще один варіант роботи з графами для кращого розуміння структури озера даних – створення навігаційного графа-моделі озера. Кожен вузол поданий набором атрибутів і ребра, що будуть зв'язками між вузлами [11]. Відповідно до результатів цього дослідження, користувачі отримували результати, які неможливо одержати через пошук за ключовими словами.

Серед інших можливостей роботи із даними – опрацювання великої кількості документації, різних видів та типів. Для аналізу і управління такими даними пропонують застосовувати Large Language Models(LLM), які попередньо навчаються на широкому спектрі даних [12].

Стосовно опрацювання наборів даних науковці пропонують виявлення особливих наборів зі спільною семантикою. Пошук таких вибірок здійснюється через зіставлення контексту та об'єднання таблиць, що можуть бути пов'язані [13].

Нині найпрактичнішою у плані архітектури озера даних є зонна. Але науковці шукають альтернативи та пропонують модифіковані варіанти, наприклад архітектура на основі FAIR Digital Objects(FDO) [14]. Такий підхід до архітектури ґрунтується на понятті гнучкості озера даних і на більш прямому зв'язку користувача і системи. Якщо користувач бажає опрацювати певні дані, він повинен спочатку налаштувати типи, новий тип метаданих, який міститиме пари ключ – значення відповідних відображень. Далі озеро приймає ці дані й "оцифровує" відповідно до заданого маніфесту – набору семантичних правил.

Використання озер даних пов'язують із багатьма складнощами, основні з них [15]: складність (проблема перетворення даних), вартість обчислень (обчислення даних великої розмірності доволі вартісне), різноманітність (пов'язана із нечіткістю управління даними), інтеграція у пам'яті (виведення даних, із кількох джерел, без їх об'єднання), відсутність чітких рішень (недостатній рівень контролю даних та великих даних), проблема узагальнення (аналіз стосується здебільшого специфічної предметної області), масштабованість (постійний приплив нових даних) та варіативність (неструктурованість та неоднорідність даних).

Під час роботи з озерами даних може виникати проблема із джерелами даних, власне із їх достовірністю або змінами домену (особливо якщо це стосується постійно повторюваної вхідної інформації). На такий випадок пропонують функції автоматичної неконтрольованої перевірки джерел [16].

Результати досліджень із використанням озер даних

У мережі інтернет наявні джерела, які стосуються досліджень, пов'язаних зі створенням структур для запису даних та метаданих у випадку опрацювання векторних і растрових просторових даних, зокрема даних просторового розташування та їх відповідного відображення [17].

Ще однією сферою застосування озер даних є гуманітарне спрямування, зокрема археологічні проєкти [18]. Основна проблема археологічних даних – їх різноманітність, зокрема текстові документи, зображення, дані давачів, документи, звіти про розкопки, створюються різними програмними чи апаратними засобами, які не завжди взаємосумісні. Схожий також фреймворк PEXESO, орієнтований на роботу із виявлення сумісних наборів даних [19]. Дослідники звертають увагу на проблеми, які можуть виникати за великих розмірностей, а також у разі використання різних форматів даних та помилки у наповненні текстових даних.

Платформа Jupyter Notebook використовує озера даних для покращення обміну даними між науковцями через компонування пов'язаних наборів даних [20]. Це стосується навчальних та дослідницьких даних, серед особливо важливих – функції зменшення розмірностей та роботи з метаданими. Серед найчастіше використовуваних типів файлів зберігання даних – тип Parquet, але його недолік – неможливість працювати із геопросторовими даними. Порівняно недавно створено розширення Spatial Parquet, яке цілком стабільно працює зі стандартним Parquet, та створює додаткові функції для поліпшення аналізу [21].

Для покращення роботи користувача науковці досліджують роботу із запитами до озер даних природною мовою. Дослідники пропонують проєктне рішення SympHony, яка має розкладати запити природною мовою на підзапити, оцінювати їх певними наборами даних і видавати результат об'єднанням цих підзапитів [22].

Цікаві дослідження, які пропонують застосування озер даних у галузі страхування. Хоч таке рішення супроводжується додатковими ризиками, його все одно вважають дуже перспективним. Для страхових компаній можливість отримувати інформацію про клієнтів буде безцінною [23]. Аналогічно специфічним є застосування цього концепту в медичних структурах, оскільки медичні дані часто оберігаються лікарською таємницею і їх можна надавати для оприлюднення чи аналізу тільки за певними юридичними аспектами [24, 26].

Можливості та перспективи озер даних у галузі освіти

Озера даних відкривають широкі можливості в контексті аналізу освітніх даних, зокрема у питанні формування освітніх траєкторій та загальної картини навчального процесу. За наявною інформацією, озера даних пропонують просувати через призму "розумного міста". Науковці запропонували проєкт системи на основі хмарних обчислень, який орієнтований на роботу із найрізноманітнішими даними, з метою розроблення і створення освітнього контенту [27]. Цей сервіс містить декілька основних елементів: хмарну платформу, багатоплатформність контенту, авторські інструменти для вчителів, перегляд вмісту, механізм висновків для підбору навчальних інструментів та засоби захисту для контенту. Загальний вигляд роботи такої системи зображено на рис. 3.

Інформація про здобувача освіти – загальна інформацію про особу та предметні здобутки протягом часу навчання. Характеристика здобувача освіти стосується зацікавлень та взаємодії особи з предметами і способами отримання та засвоєння інформації.

Можливості застосування озер даних доволі різноманітні у сфері освіти, їх подекуди уже використовують дослідники. В контексті освіти найважливіший момент – збирання даних, особливо

щодо здобувачів освіти, створення їх особистої “карти” потреб та перспектив. Варто також наголосити, що ці дані потребують постійного моніторингу та оновлення, коригування.

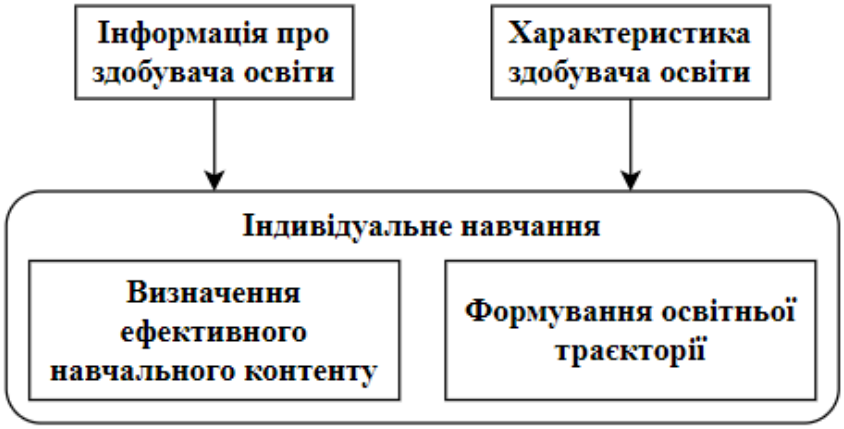

Рис.3. Схема роботи системи

Із наведеної вище інформації також маємо розуміти, що без концепту “розумного міста” ідея застосування озер даних не дасть необхідних результатів та потребуватиме більше часу та ресурсів і, звичайно, навряд буде достатньо результативною. “Розумне місто” надає набагато більше інформації, ніж робота зі звичайними системами дистанційного навчання, оскільки ми орієнтуємось не тільки на суху статистику досягнень здобувача освіти, але також можемо ураховувати додаткові фактори та обставини, які моніторяться, аналізувати інформаційний простір як усього міста, так і в певних потрібних межах.

Але найважливіше, що варто пропонувати і що вже було взято до уваги, – це наявність інструментів для створення навчальних матеріалів, а також, на нашу думку, потрібно звернути увагу на створення можливостей щодо надання відгуків на ці матеріали і відповідно тих, хто ці матеріали створює. Наявність якісного контенту сприятиме формуванню персональних освітніх траєкторій та їх подальшому розвитку.

Висновки

У статті розглянуто останні дослідження щодо концепту озер даних. Проаналізовано основні проблеми, з якими стикаються дослідники під час роботи. Озера даних перебувають на доволі складному щаблі еволюції, оскільки перспективи застосування дуже широкі, а дослідження в умовах постійного нагромадження інформації відповідно непрості й часто матеріально вартісні.

Огляд показав дуже багато спільного у підходах до роботи з озерами даних, а саме метадані, робота із сутностями, семантичні правила, використання графів і, звичайно, набори даних. До найважливішого зарахуємо метадані, оскільки вони надають інформацію щодо походження даних, можливих зв’язків між ними чи предметними областями, що становить великий пласт інформації.

Щодо перспектив в освіті, то є певні орієнтири і напрацювання науковців, які охоплюють як роботу із системами дистанційного навчання, так і з іншими електронними системами у дуже широкому розумінні, особливо в контексті “розумного міста”. Наявні підходи надають користувачам можливості навчатись та навчати. Також важливим аспектом є зберігання інформації та моделей навчання, матеріалів та здобутків кожного здобувача освіти. Всі ці дані, безперечно, зі

збереженням конфіденційності, повинні використовуватись для подальшого аналізу і, звичайно, поліпшення освітнього процесу.

Список літератури

1. Wieder, P., & Nolte, H. (2022). Toward data lakes as central building blocks for data management and analysis. *Frontiers in big Data*, *5*.

2. Alhammad, N., Bogatu, A., & Paton, N. W. (2022). Towards Schema Inference for Data Lakes. *arXiv preprint arXiv:2206.03881*.

3. Hai, R., Miller, R., Jarke, M., & Quix, C. J. (2020). *Data Integration and Metadata Management in Data Lakes* (Doctoral dissertation, Ph. D. Dissertation. RWTH Aachen University. https://doi. org/10.18154/RWTH-2020-08233).

4. Piantella, D. (2022). A Research on Data Lakes and their Integration Challenges. In *The 30th Italian Symposium on Advanced Database Systems*.

5. Chen, Z. (2022). Observations and Expectations on Recent Developments of Data Lakes. *Procedia Computer Science*, *214*, 405–411.

6. Eichler, R., Giebler, C., Gröger, C., Schwarz, H., & Mitschang, B. (2021). Modeling metadata in data lakes – a generic model. *Data & Knowledge Engineering*, *136*, 101931.

7. Thorogood, A. (2020). Policy-aware data lakes: a flexible approach to achieve legal interoperability for global research collaborations. *Journal of Law and the Biosciences*, *7*(1), lsaa065.

8. Langenecker, S., Sturm, C., Schalles, C., & Binnig, C. (2021). Towards learned metadata extraction for data lakes. *BTW 2021*.

9. Megdiche, I., Ravat, F., & Zhao, Y. (2021). Metadata management on data processing in data lakes. In *SOFSEM 2021: Theory and Practice of Computer Science: 47th International Conference on Current Trends in Theory and Practice of Computer Science, SOFSEM 2021, Bolzano-Bozen, Italy, January 25–29, 2021, Proceedings, 47*, 553–562. Springer International Publishing.

10. Cayeux, E., Damski, C., Macpherson, J., Laing, M., Annaiyappa, P., Harbidge, P., ... & Carney, J. (2022). Connecting Multilayer Semantic Networks to Data Lakes: The Representation of Data Uncertainty and Quality. *SPE Drilling & Completion*, 1–16.

11. Nargesian, F., Pu, K. Q., Zhu, E., Ghadiri Bashardoost, B., & Miller, R. J. (2020, June). Organizing data lakes for navigation. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 1939–1950.

12. Arora, S., Yang, B., Eyuboglu, S., Narayan, A., Hojel, A., Trummer, I., & Ré, C. (2023). Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes. *arXiv preprint arXiv:2304.09433*.

13. Fan, G., Wang, J., Li, Y., Zhang, D., & Miller, R. (2022). Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning. *arXiv preprint arXiv:2210.01922*.

14. Nolte, H., & Wieder, P. (2022). Realising data-centric scientific workflows with provenance-capturing on data lakes. *Data Intelligence*, *4*(2), 426–438.

15. Couto, J. C., & Ruiz, D. D. (2022, June). An overview about data integration in data lakes. In *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–7. IEEE.

16. Song, J., & He, Y. (2021, June). Auto-Validate: Unsupervised Data Validation Using Data-Domain Patterns Inferred from Data Lakes. In *Proceedings of the 2021 International Conference on Management of Data*, 1678–1691.

17. Villarroya, S., Viqueira, J. R., Cotos, J. M., & Taboada, J. A. (2022). Enabling efficient distributed spatial join on large scale vector-raster data lakes. *IEEE Access*, *10*, 29406–29418.

18. Darmont, J., Favre, C., Loudcher, S., & Noûs, C. (2020, October). Data lakes for digital humanities. In *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*, 1–4.

19. Dong, Y., Takeoka, K., Xiao, C., & Oyamada, M. (2021, April). Efficient joinable table discovery in data lakes: A high-dimensional similarity-based approach. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 456–467. IEEE.

20. Zhang, Y., & Ives, Z. G. (2020, June). Finding related tables in data lakes for interactive data science. In *Proceedings of the 2020 ACM SIGMOD International Conference on*

21. Saeedan, M., & Eldawy, A. (2022). Spatial parquet: A column file format for geospatial data lakes [extended version]. *arXiv preprint arXiv:2209.02158*.

22. Chen, Z., Gu, Z., Cao, L., Fan, J., Madden, S., & Tang, N. (2023). Symphony: Towards natural language query answering over multi-modal data lakes. In *Conference on Innovative Data Systems Research, CIDR* (pp. 8-151).

23. Molnár, B., Pisoni, G., & Tarcsi, Á. (2020). Data Lakes for Insurance Industry: Exploring Challenges and Opportunities for Customer Behaviour Analytics, Risk Assessment, and Industry Adoption. *ICETE (3)*, 127–134.

24. Eder, J., & Shekhovtsov, V. A. (2021). Data quality for federated medical data lakes. *International Journal of Web Information Systems*, *17*(5), 407–426.

25. Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data Lakes: A Survey of Functions and Systems. *IEEE Transactions on Knowledge and Data Engineering*.

26. Manco, C., Dolci, T., Azzalini, F., Barbierato, E., Gribaudo, M., & Tanca, L. (2023). HEALER: A Data Lake Architecture for Healthcare.

27. Suresh, P., Keerthika, P., Sathiyamoorthi, V., Logeswaran, K., Sentamilselvan, K., Sangeetha, M., & Sagana, C. (2021). Cloud-based big data analysis tools and techniques towards sustainable smart city services. In *Decision support systems and industrial IoT in smart grid, factories, and cities*, 63–90. IGI Global.

References

1. Wieder, P., & Nolte, H. (2022). Toward data lakes as central building blocks for data management and analysis. *Frontiers in big Data*, *5*.

2. Alhammad, N., Bogatu, A., & Paton, N. W. (2022). Towards Schema Inference for Data Lakes. *arXiv preprint arXiv:2206.03881*.

3. Hai, R., Miller, R., Jarke, M., & Quix, C. J. (2020). *Data Integration and Metadata Management in Data Lakes* (Doctoral dissertation, Ph. D. Dissertation. RWTH Aachen University. https://doi. org/10.18154/RWTH-2020-08233).

4. Piantella, D. (2022). A Research on Data Lakes and their Integration Challenges. In *The 30th Italian Symposium on Advanced Database Systems*.

5. Chen, Z. (2022). Observations and Expectations on Recent Developments of Data Lakes. *Procedia Computer Science*, *214*, 405–411.

6. Eichler, R., Giebler, C., Gröger, C., Schwarz, H., & Mitschang, B. (2021). Modeling metadata in data lakes – a generic model. *Data & Knowledge Engineering*, *136*, 101931.

7. Thorogood, A. (2020). Policy-aware data lakes: a flexible approach to achieve legal interoperability for global research collaborations. *Journal of Law and the Biosciences*, *7*(1), lsaa065.

8. Langenecker, S., Sturm, C., Schalles, C., & Binnig, C. (2021). Towards learned metadata extraction for data lakes. *BTW 2021*.

9. Megdiche, I., Ravat, F., & Zhao, Y. (2021). Metadata management on data processing in data lakes. In *SOFSEM 2021: Theory and Practice of Computer Science: 47th International Conference on Current Trends in Theory and Practice of Computer Science, SOFSEM 2021, Bolzano-Bozen, Italy, January 25–29, 2021, Proceedings 47*, 553–562. Springer International Publishing.

10. Cayeux, E., Damski, C., Macpherson, J., Laing, M., Annaiyappa, P., Harbidge, P., ... & Carney, J. (2022). Connecting Multilayer Semantic Networks to Data Lakes: The Representation of Data Uncertainty and Quality. *SPE Drilling & Completion*, 1–16.

11. Nargesian, F., Pu, K. Q., Zhu, E., Ghadiri Bashardoost, B., & Miller, R. J. (2020, June). Organizing data lakes for navigation. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 1939–1950.

12. Arora, S., Yang, B., Eyuboglu, S., Narayan, A., Hojel, A., Trummer, I., & Ré, C. (2023). Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes. *arXiv preprint arXiv:2304.09433*.

13. Fan, G., Wang, J., Li, Y., Zhang, D., & Miller, R. (2022). Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning. *arXiv preprint arXiv:2210.01922*.

14. Nolte, H., & Wieder, P. (2022). Realising data-centric scientific workflows with provenance-capturing on data lakes. *Data Intelligence*, *4*(2), 426–438.

15. Couto, J. C., & Ruiz, D. D. (2022, June). An overview about data integration in data lakes. In *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–7.

16. Song, J., & He, Y. (2021, June). Auto-Validate: Unsupervised Data Validation Using Data-Domain Patterns Inferred from Data Lakes. In *Proceedings of the 2021 International Conference on Management of Data*, 1678–1691.

17. Villarroya, S., Viqueira, J. R., Cotos, J. M., & Taboada, J. A. (2022). Enabling efficient distributed spatial join on large scale vector-raster data lakes. *IEEE Access*, *10*, 29406–29418.

18. Darmont, J., Favre, C., Loudcher, S., & Noûs, C. (2020, October). Data lakes for digital humanities. In *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*, 1–4.

19. Dong, Y., Takeoka, K., Xiao, C., & Oyamada, M. (2021, April). Efficient joinable table discovery in data lakes: A high-dimensional similarity-based approach. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 456–467. IEEE.

20. Zhang, Y., & Ives, Z. G. (2020, June). Finding related tables in data lakes for interactive data science. In *Proceedings of the 2020 ACM SIGMOD International Conference.*

21. Saeedan, M., & Eldawy, A. (2022). Spatial parquet: A column file format for geospatial data lakes [extended version]. *arXiv preprint arXiv:2209.02158*.

22. Chen, Z., Gu, Z., Cao, L., Fan, J., Madden, S., & Tang, N. (2023). Symphony: Towards natural language query answering over multi-modal data lakes. In *Conference on Innovative Data Systems Research, CIDR*, 8–151.

23. Molnár, B., Pisoni, G., & Tarcsi, Á. (2020). Data Lakes for Insurance Industry: Exploring Challenges and Opportunities for Customer Behaviour Analytics, Risk Assessment, and Industry Adoption. *ICETE (3)*, 127–134.

24. Eder, J., & Shekhovtsov, V. A. (2021). Data quality for federated medical data lakes. *International Journal of Web Information Systems*, *17*(5), 407–426.

25. Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data Lakes: A Survey of Functions and Systems. *IEEE Transactions on Knowledge and Data Engineering*.

26. Manco, C., Dolci, T., Azzalini, F., Barbierato, E., Gribaudo, M., & Tanca, L. (2023). HEALER: A Data Lake Architecture for Healthcare.

27. Suresh, P., Keerthika, P., Sathiyamoorthi, V., Logeswaran, K., Sentamilselvan, K., Sangeetha, M., & Sagana, C. (2021). Cloud-based big data analysis tools and techniques towards sustainable smart city services. In *Decision support systems and industrial IoT in smart grid, factories, and cities*, 63–90. IGI Global.

ANALYTICAL REVIEW OF DATA LAKES AND PERSPECTIVES OF APPLICATION IN THE FIELD OF EDUCATION

Andrii Pryshliak[1], Mykola Orlov

[1] Lviv Polytechnic National University, Department of Information Systems and Networks
12, S. Bandery str., Lviv, Ukraine
[1] E-mail: andrii.a.pryshliak@lpnu.ua; ORCID: 0000-0003-1681-5178

An analytical review of the development of Data Lakes and its application in various industries, as part of Big data concept solutions, was conducted. The available standard architectural solutions for the Data Lake organization are considered. Also, specialized areas that require different or additional aspects to solve the tasks, depending on the field of Data Lake use, are taken into account. For the proper organization of Data Lake, various data processing tools are used, including distributed data storage systems, semantic networks, and especially metadata. Metadata plays a huge role in recognizing the purpose of data and possible relationships between it and entities. An overview of the prospects for the use of Data Lake, in particular as context of Smart City, distance education and the education industry in general, was conducted.

Key words: Big Data; Data Lake; metadata; distance learning; online education; semantic networks; data mining; Smart City.