

ПОРІВНЯННЯ КЕРОВАНИХ РІШЕНЬ ТА DIY-ПІДХОДУ У СТРАТЕГІЇ ІНТЕГРАЦІЇ ДАНИХ DATA LAKEHOUSE

Ірина Гуркліс, Тетяна Петрушина

COMPARISON OF MANAGED SOLUTIONS AND DIY APPROACH IN DATA LAKEHOUSE AS A DATA INTEGRATION STRATEGY

Iryna Hurklis, Tetiana Petrushyna

Одеський національний університет ім. І.І. Мечникова, вул. Всеволода Змієнка, 2,
Одеса, 65000, Україна

Abstract. *The paper provides a comprehensive comparative analysis of Data Lakehouse implementation strategies, focusing on managed cloud solutions versus a Do-It-Yourself (DIY) approach using open-source technologies. Managed platforms offer quick deployment and ease of use but may lead to vendor lock-in and high long-term costs. In contrast, DIY solutions offer greater control, flexibility, and cost predictability but require more technical expertise and maintenance effort. The analysis highlights that managed solutions are well-suited for organizations seeking rapid implementation with minimal internal overhead. Meanwhile, DIY architectures allow for fine-grained customization and technological independence, particularly valuable in complex business contexts. However, DIY setups demand significant investment in skilled personnel and integration efforts. The study outlines the trade-offs between these approaches and emphasizes that the optimal choice depends on organizational needs, budget, and team capabilities.*

У сучасному цифровому середовищі, організації стикаються з постійним зростанням обсягів та різноманітності інформації. Структуровані, напівструктуровані та неструктуровані дані надходять із різних джерел у режимі реального часу, що створює виклики для традиційних архітектур зберігання й обробки даних, таких як класичні сховища або озера даних. У цих умовах виникає потреба в уніфікованих підходах, які забезпечують ефективну, масштабовану та гнучку роботу з усіма типами даних для підтримки бізнес-аналітики, машинного навчання та штучного інтелекту.

Архітектура Data Lakehouse є відповіддю на вказані виклики, поєднуючи в собі переваги сховищ даних і озер даних у єдиній гібридній моделі. Цей підхід набуває широкої підтримки як серед комерційних платформ, так і у спільнотах відкритого програмного забезпечення. За прогнозами аналітичних агентств, таких як Gartner, до 2027 року більшість організацій впроваджуватимуть саме Lakehouse-архітектуру, що свідчить про її стратегічну значущість і сталість у розвитку управління даними. Саме тому дослідження підходів до реалізації Lakehouse – керованих (managed) і самостійно побудованих (DIY) – є своєчасним і практично важливим.

Керовані рішення Data Lakehouse пропонуються провідними хмарними провайдерами та спеціалізованими компаніями (Databricks, Snowflake, Google BigQuery, Amazon Redshift), які надають повністю інтегровані платформи, розроблені для спрощення розгортання та управління архітектурою Data Lakehouse. Завдяки попередньо інтегрованим компонентам, організації часто можуть розгорнути Data Lakehouse та розпочати аналіз даних набагато швидше. Керовані рішення часто позиціонуються як економічно ефективні, але довгострокові операційні витрати, включаючи плату за ліцензії, обчислювальні ресурси та зберігання, можуть бути значними, особливо при великих обсягах даних. Одним з суттєвих обмежень керованих рішень є ризик прив'язки до вендора, оскільки тісна інтеграція з екосистемою

конкретного постачальника може ускладнити міграцію. Керовані платформи також можуть не забезпечувати такого рівня глибокої кастомізації, як DIY-підхід, обмежуючи організації функціями, які пропонує вендор. Крім того, організації залежать від плану розвитку продукту вендора, через що нові функції або виправлення можуть з'являтися не одразу.

DIY-підхід передбачає створення Data Lakehouse з нуля, використовуючи різні технології з відкритим вихідним кодом (Apache Iceberg, Delta Lake, Apache Hudi). Однією з важливих переваг є відсутність прив'язки до вендора, адже відкриті стандарти та формати файлів, такі як Apache Parquet, Iceberg та ORC, дозволяють організаціям працювати з різними інструментами та постачальниками. Це забезпечує довгострокову незалежність і передбачуваність витрат, а також дає змогу уникнути дорогих та складних міграцій у майбутньому. DIY-підхід надає значний рівень гнучкості та кастомізації, оскільки програмне забезпечення з відкритим вихідним кодом дозволяє отримувати доступ до коду, змінювати його та розширювати відповідно до бізнес-вимог.

Початкові витрати на DIY можуть бути вищими через необхідність навчання персоналу та налаштування, проте довгострокова вартість володіння може бути значно нижчою, завдяки усуненню ліцензійних витрат та оптимізації ресурсів через вибір більш ефективних і дешевих компонентів з відкритим вихідним кодом. Контроль над базовою інфраструктурою дозволяє організаціям оптимізувати використання ресурсів і передбачати розподіл витрат на рівні окремих процесів, що є важливою перевагою DIY. Проте створення Data Lakehouse з відкритих компонентів може бути складним: потрібно ретельно налаштовувати систему, слідкувати за сумісністю версій інструментів, а оновлення одного компоненту може вимагати змін в інших. Неналежне управління гнучкістю призводить до поганої якості, неузгоджених та недостатньо задокументованих даних. Крім того, пошук фахівців з необхідними знаннями розподілених систем та конкретних технологій залишається значною проблемою.

Порівняння керованих рішень Data Lakehouse та DIY-підходу виявляє чіткі компроміси, що впливають на вибір стратегії інтеграції даних. Керовані рішення пропонують зручність, швидкість розгортання та надійну підтримку, тоді як DIY-підхід забезпечує контроль, гнучкість, довгострокову економічну ефективність та сприяє внутрішнім інноваціям. DIY-підхід з його особливостями буде доречним для організацій зі складними або постійно змінними вимогами, де стандартні керовані рішення можуть бути недостатніми. Рекомендується провести ретельну оцінку внутрішніх можливостей та бути готовими інвестувати в розвиток команди, розробити стратегії управління даними та забезпечення якості, щоб уникнути ризику перетворення озера даних на "болото даних".

Майбутні дослідження можуть зосередитися на детальному аналізі вдосконалених механізмів управління та автоматизації на базі III для DIY Data Lakehouse, що може ще більше спростити управління та підвищити ефективність цих архітектур.

Бібліографія

1. The Potential of the Data Lakehouse // Raconteur. – 2024. – [Електронний ресурс]. – Режим доступу: <https://www.raconteur.net/technology/the-potential-of-the-data-lakehouse> (дата звернення: 15.05.2025).
2. Vercellone-Smith P., Basha E., Zeigler D., Grier M. From Traditional Data Warehouses to Lakehouse Architectures: Tackling Modern Data Challenges. – [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/388991438_From_Traditional_Data_Warehouses_to_Lakehouse_Architectures_Tackling_Modern_Data_Challenges (дата звернення: 16.05.2025).
3. Hassan I. Storage structures in the era of big data: from data warehouse to lakehouse // *Journal of Theoretical and Applied Information Technology*. – 2024. – Т. 102. – № 6. – [Електронний ресурс]. – Режим доступу: <https://www.jatit.org/volumes/Vol102No6/16Vol102No6.pdf> (дата звернення: 17.05.2025).