УДК 004.65
А. А. Станько, Ph.D; С. З. Кульчицький; Ю. В. Срібний
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ФОРМУВАННЯ КЕРОВАНИХ ОЗЕР ДАНИХ: МЕТОДОЛОГІЇ, ІНСТРУМЕНТИ ТА ПРАКТИЧНІ АСПЕКТИ

UDC 004.65
A. A. Stanko, Ph.D; S. Z. Kulchytskyi; Y. V. Sribnyi

FORMATION OF MANAGED DATA LAKES: METHODOLOGIES, TOOLS AND PRACTICAL ASPECTS

Конвеєри даних широко використовуються для збору даних з гетерогенних джерел, для виконання низки перетворень над ними та для передачі перетворених даних до системи призначення для аналізу даних. Такий конвеєр даних також можна назвати процесом ETL (вилучення, перетворення, завантаження), який часто використовується для надходження чистих і трансформованих даних до сховища даних [1]. Конвеєри даних все ще вважаються найсучаснішою технологією, хоча деякі недоліки добре відомі [2]. Однією з таких проблем є втрата інформації під час процесу ETL, що може статися, наприклад, коли найбільш релевантні атрибути агрегуються для того, щоб зробити формат придатним для сховища даних. Як наслідок, для подальшого аналізу доступна лише заздалегідь визначена підмножина атрибутів. Таким чином, семантика «схема-на-запис» цього підходу обмежує повторне використання даних, що зберігаються в традиційних системах управління даними, для аналітики за межами початкового обсягу. Для того, щоб запобігти цій втраті інформації, у 2010 році Діксоном було введено концепцію озер даних. Озеро даних, на відміну від сховища даних, зазвичай розробляється як центральне сховище для всіх наборів даних з усіх джерел даних у їхньому первинному форматі [4]. Оскільки для надходження даних в озеро даних не потрібно ніяких перетворень і не робиться ніяких припущень щодо подальшого аналізу, використовується підхід, заснований на зчитуванні схем, що забезпечує високу гнучкість і можливість багаторазового використання.

Хоча не існує загальноприйнятої концепції озера даних, узгоджена схема за Діксоном вимагає масштабованої системи зберігання гетерогенних даних, де вчені можуть досліджувати та аналізувати ці набори даних. Ці вимоги йдуть пліч-о-пліч з потребою в недорогих технологіях і спочатку призвели до сильної асоціації реалізацій озер даних з Apache Hadoop. Потім їх витіснили пропрієтарні хмарні рішення на базі Azure або AWS [4], які привнесли в концепцію озер даних перевагу розділення ресурсів зберігання і обчислювальних ресурсів.

Оскільки необроблені дані в озері даних, швидше за все, піддаються багатьом послідовним перетворенням, що призводить до появи декількох артефактів, які будуть поглинуті назад в озеро даних, збереження стислої інформації про походження є дуже складним завданням, але має вирішальне значення для збереження керованості озером даних. Оброблені дані, які щойно зберігаються у сховищі даних, згодом буде важко знайти і, ймовірно, неможливо зрозуміти та відтворити, що потенційно зробить їх марними.

Література

1. Dinesh, Lina, and K. Gayathri Devi. «An efficient hybrid optimization of ETL process in data warehouse of cloud architecture.» Journal of Cloud Computing 13.1 (2024): 12.
2. Munappy, A.R., Bosch, J., Olsson, H.H.: Data pipeline management in practice: Challenges and opportunities. In: Morisio, M., Torchiano, M., Jedlitschka, A. (eds). Product-Focused Software Process Improvement, pp. 168–184. Springer, Cham (2020)
3. Zhao, Xiaoyan, Conghui Zhang, and Shaopeng Guan. «A data lake-based security transmission and storage scheme for streaming big data.» Cluster Computing 27.4 (2024): 4741-4755.
4. Errami, S. A., Hajji, H., El Kadi, K. A., & Badir, H. (2023). Spatial big data architecture: from data warehouses and data lakes to the Lakehouse. Journal of Parallel and Distributed Computing, 176, 70-79.