

Роман Юринець, Ігор Пірко*

¹ НЛТУ України, Львів, Україна, ORCID: 0009-0006-3589-580,
yurynets.roman@nltu.lviv.ua;

² НЛТУ України, Львів, Україна. ORCID: 0009-0008-2378-2929,
pirko@nltu.edu.ua;

Актуалізація даних у сховищах: інтеграція Change data capture (CDC) для керування видаленнями

Анотація. Описано принципи реалізації технології для відстеження та оброблення змін даних у базах даних, що дає змогу ідентифікувати та реагувати на додавання, оновлення чи видалення записів. Здійснено порівняння ефективності CDC з іншими технологіями трекінгу змін в даних. Описано проблеми і виклики при інтеграції CDC у великомасштабні сховища даних.

Ключові слова – Change Data Capture, база даних, інтеграція, сховище, керування, трекінг, технологія.

Change Data Capture (CDC) – це технологія для відстеження та оброблення змін даних у базах даних, що дає змогу ідентифікувати та реагувати на додавання, оновлення чи видалення записів [1]. CDC популярна у сценаріях, коли необхідно постійно синхронізувати дані між різними системами або оперативно аналізувати зміни у базах даних. CDC стає важливим інструментом для забезпечення цілісності даних, їхнього моніторингу та відновлення видалених даних у разі непередбачених ситуацій або помилок. В контексті актуалізації даних у великих сховищах, важливість CDC ще більше зростає. Інтеграція технології CDC в управління сховищами даних вимагає ретельного підходу та розуміння специфік взаємодії інформаційних систем, забезпечуючи, Отже, надійний контроль за всіма змінами в даних, враховуючи їх видалення.

Отримані результати. Основні переваги Change Data Capture:

1. цілісність та доступність даних;
2. реальне відстеження змін;
3. синхронізація між системами.

Якщо порівнювати ефективність Change Data Capture з іншими технологіями трекінгу змін в даних, то варто відзначити, що Polling порівняно з Change

Data Capture має низьку ефективність, затримку та високе споживання ресурсів. Transaction Log Mining має високу швидкодію, гнучкість але мінімізовану спеціалізованість [3]. Database Triggers складна у налаштуванні та підтримці, хоча масштабованість вражаюча [4]. Snapshot-based Comparison має низьку точність та вплив на навантаження системи, хоча володіє підвищеною швидкістю [5].

Change Data Capture, завдяки своїй здатності працювати в реальному часі та з мінімальним навантаженням на систему відстежувати зміни даних, є більш ефективним та зручним методом у порівнянні з іншими технологіями трекінгу. Polling та Snapshot-based Comparison можуть підійти для простих і менш критичних до часу рішень, але CDC є незамінним у системах, де потрібна висока швидкість оновлення.

Інтеграція Change Data Capture у великомасштабні сховища даних – це потужний підхід для забезпечення актуальності даних у реальному часі, але він супроводжується низкою проблем та викликів [2]. Ось основні з них:

1. обробка високого навантаження:

- великий обсяг даних;
- питання масштабування.

2. забезпечення цілісності та узгодженості даних:

- конфлікти змін;
- транзакційна узгодженість;
- проблеми дублювання.

3. затримки та узгодженість в режимі реального часу:

- затримки передачі даних;
- складнощі з обробкою подій в режимі реального часу.

4. управління схемами та змінами у структурі даних:

- еволюція схем;
- відстеження версій.

5. Проблеми безпеки та відповідності вимогам:

- захист конфіденційних даних;
- відповідність регуляторним вимогам.

6. складність налаштування та підтримки інфраструктури:

- інтеграція з існуючими технологіями;
- підтримка та моніторинг.

7. витрати на ресурси:

- використання обчислювальних ресурсів;
- витрати зберігання даних.

Висновок. Таким чином варто наголосити, що Інтеграція технології Change Data Capture (CDC) у великомасштабні системи сховищ даних сприяє актуалізації інформації та стабільному керуванню даними, зокрема керуванню процесом видалення. CDC ефективно відстежує зміни у режимі реального часу, що дає змогу оперативно ідентифікувати видалення та забезпечує можливість відновлення даних. Ця технологія незамінна у системах, де критично важливо зберігати цілісність та доступність інформації.

Проте інтеграція CDC вимагає комплексного проектування та постійної технічної підтримки, оскільки вона може збільшувати навантаження на систему. Важливо розробити механізми для зниження цього впливу, забезпечуючи ефективну обробку та моніторинг змін без значного зниження продуктивності. Впровадження CDC вимагає глибокого аналізу потенційних викликів та ретельного врахування специфіки ділових та технічних процесів для забезпечення стійкості інформаційних систем.

Список використаних літературних джерел

- [1] Lekkala, Hemadri. (2023). An Implementation of Change Data Capture (CDC). International Journal of Computer Trends and Technology. 71. 15-18. 10.14445/22312803/IJCTT-V71I2P103.
- [2] Hao, Liang & Jiang, Tao & Lin, Yantuan & Lu, Yitong. (2023). Methods for Solving the Change Data Capture Problem. 10.1007/978-3-031-20738-9_87.
- [3] He, P., Zhu, J., He, S., Li, J., & Lyu, M. (2018). Towards Automated Log Parsing for Large-Scale Log Data Analysis. IEEE Transactions on Dependable and Secure Computing, 15, 931-944. 10.1109/TDSC.2017.2762673.

- [4] Feng-ling, W. (2011). Intelligent Application of Trigger in Database. Coal Technology.
- [5] Fink, M., Meilicke, C., & Stuckenschmidt, H. (2020). Explaining Differences Between Unaligned Table Snapshots., 133-144. 10.5441/002/edbt.2020.13.

Data updates in repositories: Change data capture (CDC) integration for deletion management

Roman Yurynets, Igor Pirko

The paper describes the principles of implementing a technology for tracking and processing data changes in databases, which allows identifying and responding to the addition, updating, or deletion of records. The efficiency of CDC is compared with other data change tracking technologies. The problems and challenges of integrating CDC into large-scale data warehouses are described.