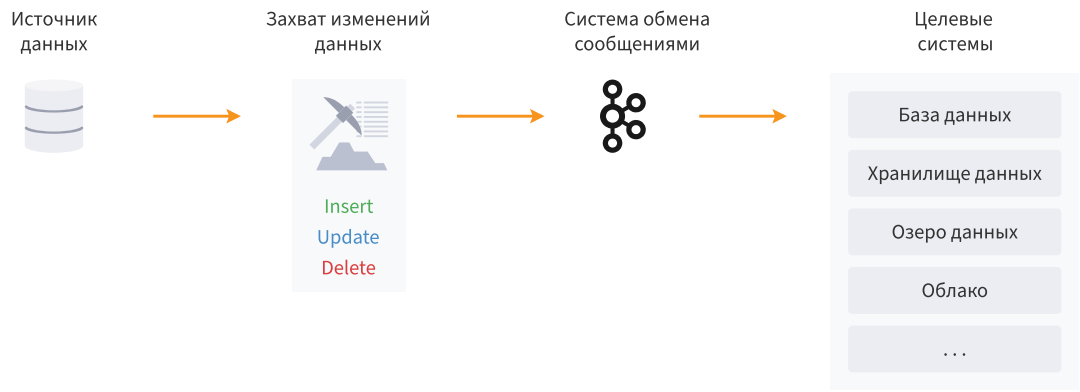
Захват изменений данных (Change Data Capture)

> Синонимы: Отслеживание изменений в данных, CDC
> Разделы: Бизнес-задачи

Захват изменений данных (CDC) — это процесс обнаружения и регистрации изменений информации в источниках данных и их репликация в целевые системы хранения в реальном времени. Он широко применяется в средах хранилищ данных (ХД) корпоративных информационных систем.



Данные в ХД загружаются из множества источников. Наиболее распространенный вид — OLTP-системы, где фиксируются транзакции, отражающие события в бизнесе. Поток транзакций в OLTP-систему имеет различную интенсивность в зависимости от особенностей бизнеса: где-то они идут непрерывным потоком, а где-то — раз в день или реже. Все изменения в источнике должны передаваться в ХД через ETL или ELT, но возможны и другие варианты. Для организации этого процесса можно использовать два режима:

- **оффлайн (пакетная)** — загрузка изменений в соответствии с регламентом. Например, перенос в ХД производится один раз в сутки, причем загружаются все транзакции, дата и время создания которых позже, чем у предыдущей загрузки.
- **онлайн (потоковая)** — изменения в источниках отражаются в ХД сразу после их появления в режиме реального времени.

Можно выделить два варианта организации процесса CDC:

- **«толкать» (push)** — захват и отправка изменений в ХД производится средствами источников данных.

- **«тянуть» (pull)** — ETL-процесс хранилища сам отслеживает изменения, выгружает их из источника и переносит в ХД. Он должен непрерывно опрашивать источники для получения изменений и принятия соответствующих действий.

Для реализации CDC используются следующие методы.

На основе временных меток. В таблицу вводится столбец, отражающий время последнего изменения (LAST_MODIFIED, LAST_UPDATED и т.д.). Целевая система будет запрашивать только записи, которые были обновлены с момента предыдущей загрузки.

ID	First Name	Last Name	Email	Last Modified
101	Ivan	Petrov	ivan.petrov@yandex.ru	2022-04-28T16:50:34

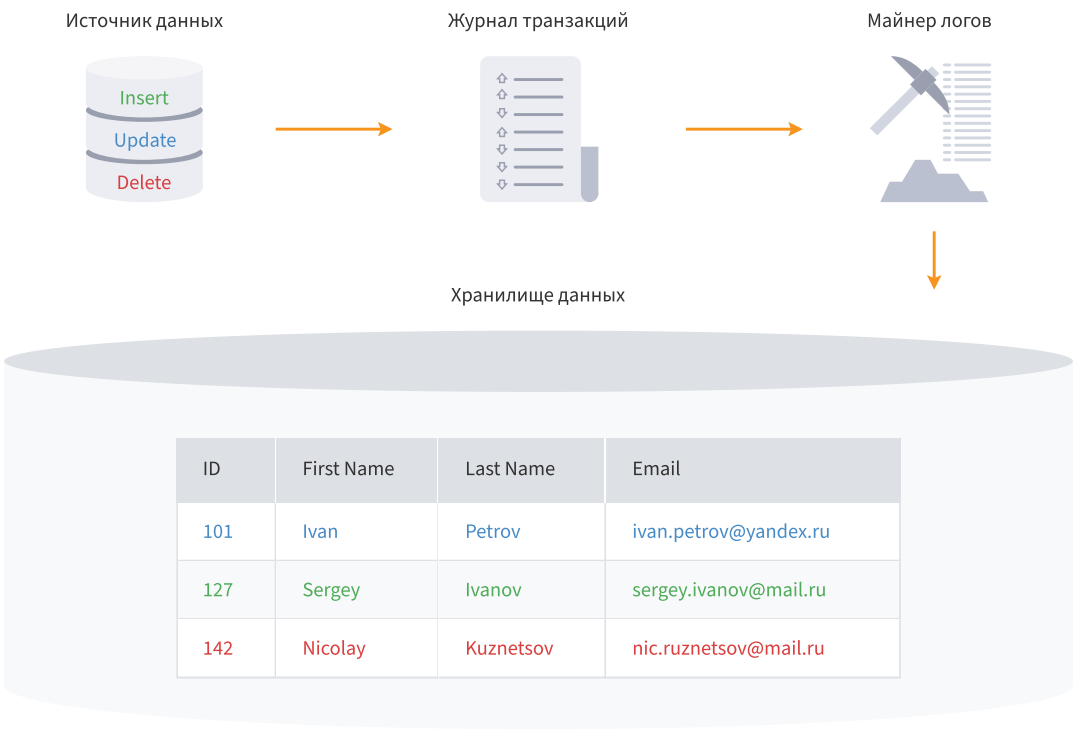
ID	First Name	Last Name	Email	Last Modified
101	Ivan	Petrov	iv.petrov@yandex.ru	2022-04-30T10:37:49

На рисунке видно, что изменение поля электронной почты Email сопровождается изменением поля временной метки обновления Last Modified.

На основе триггеров. Большинство баз данных поддерживают триггеры — хранимые процедуры, которые автоматически выполняются при возникновении в таблице определенного события (например, действий INSERT, UPDATE или DELETE). Для регистрации любых изменений данных необходим один триггер для каждой операции в таблице.

Изменения сохраняются в той же базе данных, но в отдельной таблице, которая обычно называется **теневой** или **таблицей событий**. Кроме того, могут быть использованы системы обмена сообщениями, которые позволят публиковать изменения данных в очередях, где на них «подписываются» соответствующие целевые системы.

На основе журналов. СУБД регистрируют все изменения — INSERT, UPDATE и DELETE — произошедшие в базе данных, и соответствующие им временные метки в файлах, называемых журналами транзакций.



Эти журналы используются для резервного копирования, но их также можно применять для распространения изменений в целевые системы. Изменения данных фиксируются в режиме реального времени.

Первоначально CDC получил распространение как альтернативное решение пакетной репликации для заполнения хранилищ данных. Однако в последние годы он стал популярным и для миграции в облако.