Городецька В. В., магістрантка
Левківський В. В., ст. викладач
кафедри комп'ютерних наук
Вакалюк Т. А., д.пед.н., проф.,
професор кафедри інженерії програмного забезпечення
Державний університет «Житомирська політехніка»

ЩО ТАКЕ ETL І ДЛЯ ЧОГО ЦЕ ПОТРІБНО

Проблема, через яку в принципі народилася необхідність використовувати рішення ETL, полягає у потребах бізнесу в отриманні достовірної звітності, серед великої кількості не сортованої інформації, який відбувається у даних ETL-системах.

ETL (Extract, Transform, Load) – сукупність процесів управління сховищами даних, які включають в себе: вилучення даних із зовнішніх джерел; перетворення та очищення даних згідно з бізнес-потребами; завантаження обробленої інформації у корпоративне сховище даних.

Поняття **ETL** виникло в результаті появи безлічі корпоративних інформаційних систем, які необхідно інтегрувати один з одним з метою уніфікації та аналізу даних, що зберігаються в них. Реляційна модель представлення даних, придатна потреб транзакційних систем, виявилася неефективною для комплексної обробки та аналізу інформації. Пошук уніфікованого рішення призвів до розвитку сховищ та вітрин даних – самостійних систем зберігання консолідованої інформації у вигляді вимірів та показників, що вважається оптимальним для формування аналітичних запитів [1].

Прикладне призначення **ETL** полягає в тому, що потрібно організувати таку структуру даних за допомогою інтеграції різних інформаційних систем. Враховуючи, що BI (business intelligence) –технології позиціонують себе, як «концепції та методи для покращення прийняття бізнес-рішень з використанням систем на основі бізнес-даних» [1] можна зробити висновок про пряму належність ETL до технологічного стеку. Незалежно від особливостей побудови та функціонування ETL-система має забезпечувати виконання трьох основних етапів процесу ETL-процесу: вилучення даних з одного або кількох джерел та підготовка їх до перетворення (завантаження у проміжну область, перевірка даних на відповідність специфікаціям та можливість подальшого завантаження у сховище даних); трансформація даних – перетворення форматів та кодування, агрегація та очищення; завантаження перетворених даних, включаючи інформацію про структуру їх представлення (метадані) у необхідну систему зберігання або вітрину даних.

На практиці **ETL** виступає як проміжний шар між OLTP- і OLAP-системами. OLTP – це транзакційні системи для обробки безперервного потоку невеликих за розміром транзакцій у режимі реального часу: ERP-, MES-, банківські та біржові програми. Вони автоматизують структуровані завдання, що повторюються, обробки даних, наприклад, введення замовлень і банківські транзакції, у великій кількості за короткі проміжки часу [1]. Для таких запитів призначені OLAP-системи. OLAP– це інтерактивне аналітичне опрацювання, підготовка сумарної (агрегованої) інформації на основі великих масивів даних, структурованих за багатовимірним принципом [2].

Таким чином, основні функції ETL-системи можна подати у вигляді послідовності операцій з передачі даних з OLTP до OLAP:

1. Завантаження в ETL "сирих" даних (Raw Data) довільної якості для подальшої обробки, при цьому виконується звір суми рядків, що прийшли: якщо в системі-джерелі більше рядків, ніж у Raw Data, то завантаження пройшло з помилкою;
2. Валідація даних, коли дані послідовно перевіряються на коректність та повноту, складається звіт про помилки для виправлення;
3. Налаштування відповідності маппінгованих даних з цільовою моделлю, коли до валідованої таблиці прилаштовуються стовпці за кількістю довідників цільової моделі, а потім у кожному прибудованому осередку кожного рядка проставляються відповідність значень цільових довідників (1:1, *:1, 1:* або *: *);
4. Агрегація даних, необхідна через різницю деталізації даних в OLTP та OLAP-системах. OLAP є повністю денормалізованою таблицею фактів і навколишні таблиці довідників за схемою зірочка або сніжинка. У цьому максимальна деталізація сум OLAP дорівнює кількості перестановок (агрегацій) всіх елементів всіх довідників. OLTP-система може містити кілька сум для одного набору елементів довідників. Щоб простежити, з яких рядків OLTP сформувалася сума в осередку OLAP-системи, необхідний маппінг OLTP-деталізації, а потім «склейка» даних в окремій таблиці для завантаження в OLAP;
5. Вивантаження в цільову систему з використанням конектора та інтерфейсних інструментів.

Отже, облік різних аспектів ETL-процесів з прицілом - на майбутнє дозволить ретельно спланувати необхідні роботи, уникнути збільшення загального часу реалізації та вартості проекту, а також забезпечити BI-систему надійними та актуальними даними для аналізу.

Список використаних джерел

1. ETL: що таке, навіщо і для кого. https://chernobrovov.ru/articles/etl-chto-takoe-zachem-i-dlya-kogo.html.
2. Що таке ETL. https://ru.wikipedia.org/wiki/ETL.