СЕКЦІЯ 3. КОМП'ЮТЕРНІ СИСТЕМИ ТА МЕРЕЖІ

УДК 004.415.2.043
Іван Бородій, Галина Осухівська, канд. техн. наук, доц.
Тернопільський національний технічний університет імені Івана Пулюя

ПРОЄКТУВАННЯ ПРОГРАМНОЇ СИСТЕМИ ФОРМУВАННЯ АГРЕГОВАНИХ НАДВЕЛИКИХ МАСИВІВ ДАНИХ

Ivan Borodii, Halyna Osukhivska, Ph.D., Assoc. Prof.

DESIGN OF A SOFTWARE SYSTEM FOR THE FORMATION OF AGGREGATED BIG DATA ARRAYS

В теперішній час підприємства, в процесі своєї діяльності, виконують обробку величезних масивів даних, які досягають рівня петабайтів чи ексабайтів. Ці обсяги даних необхідні для представлення числових значень у агрегованому вигляді для виконання математичних операцій. Наприклад, для підприємств у галузі продажів агрегованими даними є прибуток, кількість унікальних клієнтів або транзакцій, які можуть використовуватись з метою прийняття стратегічних бізнес-рішень, прогнозу майбутніх тенденцій, контролю фінансових показників тощо. Таким чином, розробка програмних систем із застосуванням агрегованих даних є актуальною та важливою задачею в контексті сучасного бізнес-середовища.

Джерелами даних для проєктованої програмної системи формування агрегованих надвеликих масивів даних є чотири системи управління базами даних (СУБД): MySQL, PostgreSQL, MongoDB та СУБД типу «ключ-значення» Redis. Процес збору, трансформації та завантаження даних (ETL) реалізовано за допомогою мови програмування Python з використанням бібліотеки Pyspark, яка дозволяє інтегрувати в Python інструмент обробки даних Apache Spark, оптимізований для роботи з великими об'ємами даних [2]. Для постійної актуалізації даних в системі, виконання процесу ETL здійснюється періодично. Сховище даних, куди дані будуть завантажені процесом ETL, реалізоване за допомогою СУБД Apache Hive, в якій розроблено результуючу структуру даних та виконується агрегування даних.

Сховище даних у системі спроєктовано на основі моделі «сутність – зв'язок», яка передбачає строго структуроване сховище, що дозволяє спростити реалізацію ETL [1]. Для представлення агрегованих даних використано систему запитів Trino, яка призначена для обробки надвеликих масивів даних та виконання SQL запитів над ними. Візуалізація агрегованих даних у читабельному форматі виконана за допомогою засобу Microsoft Power BI. За допомогою Trino, Power BI отримує дані зі сховища даних Hive та дозволяє генерувати релевантні звіти. Запропонована програмна система дозволяє забезпечити формування агрегованих надвеликих масивів даних для потреб сучасного бізнесу.

Література

1. Бородій, І. І., Парамуд, Я. С., Сав'як, В. В. Принципи побудови програмної системи формування агрегованих даних. / Бородій, І. І., Парамуд, Я. С., Сав'як, В. В // Вісник Національного університету Львівська політехніка. Комп'ютерні системи та мережі. 2018. (905). С. 25-32.

2. Горішня К. О. Аналіз високоефективних кластерів для обробки великих даних / К. О. Горішня // Радіоелектроніка та молодь у XXI столітті : матеріали 27-го Міжнар. молодіж. форуму, 10–12 травня 2023 р. – Харків : ХНУРЕ, 2023. – Т. 5. – С. 29–30.