АНАЛІЗ ТА ІНТЕГРАЦІЯ ДАНИХ В СУЧАСНИХ РОЗПОДІЛЕНИХ WEB-СИСТЕМАХ

Стьопін В.І., Горбатенко Б.В.
Науковий керівник – к.т.н., проф. Іванов В.Г.
Харківський національний університет радіоелектроніки
(61166, Харків, просп. Науки,14, каф. Системотехніки)
e-mail:{vladyslav.stopin, bohdan.horbatenko}@nure.ua,
телефон (095) 540-75-52

A method of data processing and integration is proposed. Nowadays, the majority of systems are distributed in some way and have both structured and raw data. Since cloud computing and storages become more and more popular, we have become to the task of data integration. In business, large databases can be created from databases of many other smaller companies. In a result, it turns out the system with real time updates and valid information. To achieve it, data should be transferred through secure and fast channels and be encrypted in an open-standard format such as XML or JSON. In Web surrounding many other factors are included and described in this article.

З появою великомасштабних та розподілених систем, стало можлива обробка великих об’ємів даних, що поставило нову задачу обробки вхідної інформації. Зважаючи на швидкий зріст кількості інформації, та зменшенню часу її актуальності, виникла проблема інтеграції даних. У web-системах, зазвичай, дані передаються у форматі JSON, завдяки його зручному та швидкому кодуванню. Основною задачею інтеграції є зведення неоднорідності зовнішніх джерел даних до єдиного формату.

Труднощі трапляються при створені архітектури майбутньої системи, тому що потрібно врахувати формат даних, що зберігаються та обробляються. Зведення даних до уніфікованого формату може призвести до втрати частини корисної інформації.

У роботі пропонується метод обробки інформації та її запису до бази даних включає створення окремих, незалежних сервісів, які розроблені спеціально під формат даних джерела, з якого вони завантажуються [1]. Головною функцією сервісів є отримання та трансформація даних у формат, в якому вони зберігаються в базі даних. Основними конфліктами при зберіганні даних є їх неоднорідність та структура.

Основна мета розробленого методу – це швидка інтеграція даних з великої кількості зовнішніх ресурсів, аналіз та зведення їх до формату, в якому вони зберігаються у розподіленій web-системі. Це дозволить підтримувати актуальність даних в системі.

Важливим етапом є проектування бази даних, яка буде зберігати інформацію у необхідному системі форматі, тому що її архітектура впливає на швидкість запису та читання інформації, а також на можливість розширення системи в майбутньому. Створюючи формат даних, необхідно

проаналізувати існуючі системи, з яких буде отримуватися інформація. Формат даних повинен бути таким, що містить як основну інформацію, так і всі тонкощі кожної системи. Таким чином, при інтеграції даних, деякі поля бази даних необхідно буде заповняти самостійно. Функціонування методу не вимагає використання розподіленої системи, але у середовищі з одним потоком даних він виявляється не ефективним.

Для кожного окремого джерела інформації створюється окремий сервіс, який, незалежно від інших компонентів, обробляє вхідну JSON-строку та приводить масив даних до єдиного формату. Після отримання даних, вони формуються у масив. Швидкість роботи з такою структурою у найгіршому випадку складає *O(1)* на доступ та *O(n)* на вставку [2]. При обробці масивів з 10 000 елементів, робота над аналізом даних займе не більше 0,01 секунди. Слабкою стороною методу є запис до бази даних, так як пропускна здатність каналу обмежена кількістю операцій в секунду часу. Важливою умовою у web-середовищі також є захист інформації та швидкість передачі даних (від сервера до клієнта). Основним критерієм для виконання цих умов є використання протоколу HTTPS/2. Для підвищення швидкості обробки даних можна додати реплікації бази даних та кешування пар ключ-значення для моментальних запитів до бази. Цей процес наглядно показано на Рисунку 1.

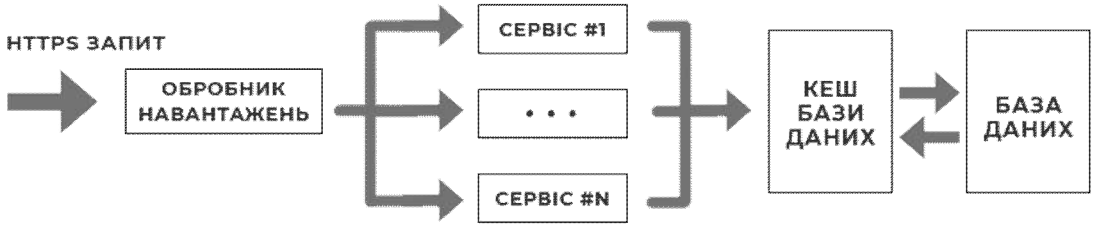

Рисунок 1 – Процес інтеграції даних у розподіленій web-системі

Таким чином, буде найшвидше вирішуватись задача інтеграції, а система зберігатиме умови розподілених систем, а саме доступність, продуктивність, надійність, масштабованість, керованість та вартість. Для підвищення керованості, систему можна запускати у контейнерах Docker, в які помістити один чи декілька сервісів, розмістивши їх у окремих вузлах.

Література

1. Kleppmann M. Designing Data-Intensive Applications : навч. посіб. Бостон : O'Reilly Media, 2017. 569 с.
2. Cormen T., Introduction to Algorithms : навч. посіб. 3-е вид., Кембридж : MIT Press, 2009. 1312 с.