УДК 004.62:004.65:004.75

Шафорост В. В., Корнієнко К. К., Хмель А. Є., здобувачі 4 курсу спеціальності 122 Комп'ютерні науки,

Комаров В. Ф., канд. техн. наук, старший викладач кафедри інформаційних технологій

МОДЕЛЮВАННЯ ВЕЛИКИХ НАБОРІВ ДАНИХ

Донецький національний університет імені Василя Стуса, м. Вінниця

В епоху цифрових технологій та інтернету ми спостерігаємо значне збільшення обсягу та доступності різнобічної інформації. Вона надходить з багатьох джерел, як-от комп'ютери, мобільні пристрої, датчики чи соціальні медіа, які створюють структуровані, напівструктуровані або неструктуровані дані безперервно. Обсяг даних, із якими ми стикаємося, зріс до терабайтів або петабайтів і продовжує збільшуватися далі, а типи даних, які генеруються програмами, стають різноманітнішими, ніж раніше. Внаслідок цього традиційні реляційні бази даних стикаються з проблемами збору, зберігання, пошуку, обміну, аналізу та візуалізації даних. Багато IT-компаній намагаються розв'язувати означені задачі, використовуючи бази даних (БД) типу NoSQL, як-от Cassandra або HBase, та розподілені обчислювальні системи, як-от Hadoop [1]. Бази даних NoSQL зазвичай є нереляційними, розподіленими, горизонтально масштабованими та вільними від схем.

Моделювання даних – це процедура створення концептуального представлення об'єктів даних та їх асоціацій один з одним. Процес моделювання даних зазвичай включає кілька етапів, зокрема збір вимог, концептуальне проєктування, логічне проєктування, фізичне проєктування та впровадження. На кожному етапі розробники моделей даних співпрацюють з усіма учасниками процесу, щоб визначити вимоги до даних, визначити сутності та атрибути, встановити зв'язки між об'єктами даних і створити модель, яка точно відображає дані та може використовуватися програмою, розробниками, адміністраторами БД та іншими учасниками [2].

Традиційне моделювання даних зосереджене на вирішенні складних зв'язків між даними, що підтримують схему, однак воно не стосується нереляційних баз даних без схем. Старі способи моделювання даних більше не застосовуються – потрібна інша методологія керування великими даними.

Моделювання великих наборів даних містить дві термінології, а саме «Великі дані» (Big Data) та «Моделювання даних» (Data Modeling). Термін «великі дані» використовується для опису великих, складних і швидко збільшуваних

наборів даних, які мають численні, автономні та незалежні джерела[3]. Точний аналіз таких даних може мати значні переваги, оскільки він дає змогу розпізнавати закономірності та зв'язки в наборах даних [4].

Моделювання даних зазвичай виконується з допомогою кількох рівнів концептуалізації, зокрема: концептуального, логічного і фізичного.

Під час створення концептуального рівня визначається, що необхідно включити в конфігурацію моделі для опису та координації комерційних принципів. Він зосереджений насамперед на бізнес-записах, функціональності та комунікації. За його розробку насамперед відповідають архітектори даних і бізнес-стейкхолдери. Концептуальна модель даних використовується для визначення сфери застосування методу. Це інструмент для організації, визначення та візуалізації ідей компанії. Метою розробки обчислювальних моделей даних є розробка нових сутностей, зв'язків і атрибутів. Архітектори даних і зацікавлені сторони часто створюють обчислювальні моделі даних.

Логічна модель даних пояснює розташування структур даних і їх зв'язки. Вона допомагає включити додаткові дані в компоненти концептуальної моделі даних і закладає основу для побудови фізичної моделі. Ця модель дає нам змогу перевіряти та оновлювати інформацію про використання зв'язків, які були встановлені раніше. Логічна модель даних налаштовується і створюється окремо від системи керування базами даних (СКБД). Вона визначає дані, необхідні для проєкту, але взаємодіє з іншими логічними моделями даних відповідно до складності проєкту.

Фізична модель пояснює, як модель даних реалізується в базі даних, як використовувати систему керування базою даних для виконання моделі даних. Вона описує процес з погляду таблиць, операцій створення, читання, оновлення і вилучення (CRUD), індексів, розділення тощо. Фізичний рівень передбачає створення конкретних деталей про те, як дані будуть зберігатися, зокрема типи даних, індекси та іншу технічну інформацію.

У такий спосіб моделювання даних допомагає підвищити узгодженість іменування, правил та безпеки. Це покращує аналітику даних. Акцент робиться на необхідності наявності та організації даних незалежно від способу їх застосування [5].

Проте моделювання великих даних – це не лише про розуміння складних даних, але й про те, як ці дані можуть бути використані для підвищення ефективності та продуктивності роботи бази даних. Використовуючи різні моделі даних, ми можемо виявити закономірності, тренди та зв'язки, які можуть бути приховані в сировинних даних. Щоб використовувати великі дані повною мірою, ми повинні моделювати їх із допомогою різних типів моделей даних, як-от ієрархічна модель, мережева модель тощо.

Наприклад, ієрархічна модель використовує деревоподібну структуру для організації даних. Кожен запис у такій моделі має один корінь, який з'єднує всі дані. Цей корінь розширюється, як гілка, з'єднуючи батьківські вузли з відповідними дочірніми вузлами, причому кожен дочірній вузол має лише однин батьківський вузол. У цій моделі дані організовано в реляційну систему з кореляцією один-до-багатьох, що дає змогу легко знаходити і маніпулювати певними типами даних. Якщо йдеться про однорідні документи, вони впорядковані особливим способом – це фізичний порядок, у якому зберігається інформація. Метод можна застосувати до різних взаємозв'язків моделі в реальному часі.

Ієрархічна модель може бути обмеженою, коли потрібно представити більш складні відносини між даними, які не можуть бути чітко виражені в ієрархічній структурі. Незважаючи на це, ієрархічна модель даних залишається важливим інструментом для організації та обробки великих обсягів даних.

Мережева модель даних є гнучкою альтернативою ієрархічній моделі. Вона дає змогу кожному дочірньому вузлу мати кілька батьківських вузлів, у такий спосіб створюючи багатофакторну мережеву структуру, яка допомагає показати більш складні зв'язки між об'єктами. Наприклад, у базі даних компанії співробітники можуть мати різні ролі в різних проєктах, а мережева модель даних дасть змогу кожному співробітнику бути пов'язаним із різними проєктами в різних ролях.

Окрім зазначених моделей треба також згадати й інші [2]: модель ER (Entity-Relationship), реляційну, об'єктно-орієнтовану та об'єктно-реляційну моделі.

Незважаючи на численні переваги, моделювання даних у контексті великих наборів даних має обмеження та проблеми. Моделі даних можуть бути негнучкими, що ускладнює адаптацію до мінливих вимог або структур даних. Моделі даних можуть бути складними та складними для розуміння, через що може бути важко вносити дані чи ефективно взаємодіяти з ними. Моделювання даних може бути трудомістким процесом, особливо для великих або складних наборів даних.

Отже, складність реляційної бази даних обмежує масштабованість зберігання даних, проте дає змогу легко запитувати дані через механізм SQL. БД типу NoSQL мають протилежні властивості – необмежену масштабованість із більш обмеженими можливостями запитів. А великі дані потребують одночасно легкої масштабованості та можливості легко надсилати запити на дані.

У майбутньому для великих наборів даних будуть з'являтись і поширюватись нові гібридні системи, що поєднуватимуть атрибути обох підходів, втім розробка моделей та моделювання даних залишаться трудомістким, але критично важливим процесом у розробці програмного забезпечення або систем баз даних.

Список використаних джерел

1. Data Modeling for Big Data. URL: https://citeseerx.ist.psu.edu/ document? repid=rep1&type=pdf&doi=6779310266a5f15a6e85fec3cbb3f37170a40c16#page=77

2. What is Data Modelling? Overview, Basic Concepts, and Types in Detail. URL: https://www.simplilearn.com/what-is-data-modeling-article

3. Gil D., Song I.-Y. Modeling and Management of Big Data: Challenges and opportunities. *Future Generation Computer Systems, Elsevier BV*. 2016. № 63. P. 96–99. DOI: 10.1016/j.future.2015.07.019.

4. Ribeiro A., Silva A., da Silva A. R. Data Modeling and Data Analytics: A Survey from a Big Data Perspective. *Journal of Software Engineering and Applications, Scientific Research Publishing, Inc.* 2015. № 08. P. 617–634. DOI: 10.4236/jsea.2015.812058.

5. Big Data Modeling. URL: https://hkrtrainings.com/big-data-modeling

УДК 004.021:004.67

Ватаманеску С. К., здобувач 2 курсу спеціальності 122 Комп'ютерні науки,

Потапова Н. А., канд. екон. наук, доцент, доцент кафедри інформаційних технологій

ОПТИМІЗАЦІЯ АЛГОРИТМІВ ДЛЯ ВЕЛИКИХ ОБСЯГІВ ДАНИХ

Донецький національний університет імені Василя Стуса, м. Вінниця

У сучасному інформаційному суспільстві обробка та аналіз великих обсягів даних стали важливим завданням у багатьох галузях, як-от бізнес, наука та медицина. Збільшення обсягів даних виникає через використання технологій, як-от Інтернет речей та соціальні мережі. Однак це також призводить до викликів для алгоритмів обробки, що потребують постійної оптимізації.

Дослідження оптимізації алгоритмів для великих обсягів даних вимагає огляду принципів роботи з такими обсягами інформації. Застосування паралельного програмування та розподілених обчислень разом з ефективним використанням ресурсів визначає успішну обробку даних у реальному часі.

В опрацьованій літературі виділено ключові принципи, як-от паралельне програмування та розподілені обчислення для оптимальної обробки великих обсягів даних. Розподілені системи зберігання даних, як-от Apache Hadoop та Apache Spark, а також техніки індексації та кешування, використовуються для покращення швидкодії обробки даних.

Дослідницька спільнота активно досліджує різні підходи до оптимізації алгоритмів. Розв'язки включають використання розподіленої системи зберігання