Я. Марильченко

ВИКОРИСТАННЯ НЕРЕЛЯЦІЙНИХ (NOSQL) БАЗ ДАНИХ У НАЙБІЛЬШИХ КОМПАНІЯХ СВІТУ

Ya. Marylchenko

USING OF NOSQL DATABASES IN THE WORLD'S LARGEST COMPANIES

Швидке збільшення обсягу інформації зумовило пошук нових підходів до вирішення проблеми з її збереження. Вважається, що нереляційні бази даних (NoSQL) найбільш придатні для збереження слабоструктурованих даних. NoSQL охоплює широкий спектр технологій баз даних, розроблених згідно з вимогами нових систем та їх потреб.

Хоча реляційні бази даних є найпоширенішим додатком для роботи з великими даними, вони не пристосовані для обробки експоненціального зростання даних у реальному часі. Наприклад, зростання обсягів інформації в Інтернеті є викликом для реляційних баз даних. Щодня у світі створюється 2,5 квінтильйона байт даних, причому 90% з них є неструктурованими. За оцінками, до 2020 р. було створено понад 40 зеттабайт даних.

Для допомоги подолання викликів цього неструктурованого зростання, багато розробників переходять на бази даних "NoSQL" або "Not Only SQL". Системи баз даних NoSQL — це розподілені, нереляційні бази даних, які також використовують не-SQL мову і механізми в роботі з даними. NoSQL бази даних можна знайти в таких компаніях, як Amazon, Google, Netflix і Facebook, які залежать від великих обсягів даних, що не підходять для реляційних баз даних. Ці бази даних можуть ефективно працювати з поточними неструктурованими даними, такими як соціальні мережі, електронна пошта та документи. NoSQL має просту мову запитів із високою масштабованістю та надійністю.

У якості допоміжних NoSQL бази даних є надзвичайно швидкими для виконання одного типу задач (наприклад, робота з графом) та абсолютно непридатними для вирішення інших (соціальній мережі не потрібна чітка узгодженість, водночас як для банківського додатку вона є невід'ємною).

Так, у системі WindyGrid для моніторингу міста Чикаго використовується нереляційна база даних MongoDB. У ній зберігають великий масив неструктурованих даних: платформа надає сенс мільйонам записів даних, що збираються щодня з 15 найважливіших відділів Чикаго, зокрема з поліції, транспорту та пожежної служби, щодо дорожніх робіт, затримки вивезення сміття, інформація про дзвінки до 911 (надзвичайні ситуації) та 311 (скарги про шум), публічні твіти про деталі роботи міста, розташування автобусів на їхньому маршруті, кольори світлофорів у реальному часі та багато іншого.

Відображення реклами або пропозицій на поточній вебсторінці — це рішення, яке зумовлює прямий дохід. Щоб визначити, на яку групу користувачів орієнтуватися, на якій вебсторінці показувати рекламу, платформи збирають поведінкові та демографічні характеристики користувачів. База даних NoSQL дозволяє рекламним компаніям відстежувати дані користувачів, а також розміщувати рекламу дуже швидко і збільшує ймовірність кліків. AOL, Mediamind та PayPal — це компанії, які використовують NoSQL для таргетингу реклами.

PayPal, світовий лідер у галузі онлайн-платежів, спершу використовував Couchbase (база даних NoSQL) для своєї медійної рекламної мережі Media Network

та для побудови міжканальної аналітики користувачів для вирішення задачі профілювання, сегментації, підтвердження справжності особи та ін. До 2014 р. компанія керувала понад 1 мільярдом документів та 10 терабайтами даних за допомогою Couchbase. PayPal розширив використання Couchbase в аналітиці для отримання інформації про користувачів, обробляючи мільйони оновлень на хвилину завдяки технологіям Kafka та Hadoop.

Couchbase поверх багаторівневої AWS (Amazon Web Services) архітектури використовується VoIP додатком Viber для оновлень профілів користувачів у реальному часі. Також Couchbase використовується низкою відомих брендів: eBay (лістинги товарів у реальному часів), LinkedIn (кешування інформації), Tesco (каталог та менеджмент інвентарю), Cisco (VSRM — технологія для шифрування відеопотоку (для платних телеканалів)).

Нині до Інтернету під'єднані мільярди пристроїв, зокрема смартфони, планшети, побутова техніка, системи, встановлені в лікарнях, автомобілях і на складах. На таких пристроях генеруються і продовжують генеруватися великі обсяги різноманітних даних.

Реляційні бази даних не здатні зберігати такі дані. NoSQL дозволяє організаціям розширити одночасний доступ до даних із мільярдів під'єднаних пристроїв і систем, зберігати величезні обсяги даних і забезпечувати необхідну продуктивність.

Neo4j є прикладом графової бази даних. Щоб отримати дані з такої бази даних, необхідна мова, відмінна від SQL, яка б дозволила легко оперувати графом. Тому за визначенням графові БД є нереляційними та використовують NoSQL. eBay, наприклад, використовує її для надання покупцям рекомендацій. NASA використовує Neo4j для опрацювання «вивчених уроків» (даних із попередніх місій) та для так званих Великих Даних (Big Data). За словами головного архітектора NASA, завдяки отриманим даним було зекономлено 2 роки роботи та близько мільйона доларів податків. Щодо Big Data, то NASA збирає як структуровані, так і неструктуровані дані. Neo4j було застосовано для дослідження взаємозв'язку між відгуками астронавтів після повернення з космічної станції. Це дозволило NASA опрацювати масив даних за останні 15 років та зробити висновки щодо того, на які системи та підсистеми вплинули експерименти, проведені на борту.

Monsanto — агрокультурна компанія, що займається розробкою ГМО (придбана компанією Bayer), використовує Neo4j для обробки інформації про геномні послідовності рослин, щоб покращити селекцію. Для вирішення генетичних проблем дослідники ставляться до даних як до сімейного дерева, у якому набагато більше предків. Через відповідність даних графовій моделі аналіз займає лічені секунди. Neo4j використовують також такі бренди, як Caterpillar, Airbnb, Lockheed Martin, Walmart, Comcast.

NoSQL (нереляційні) бази даних є потужним інструментом обробки великої кількості слабоструктурованих або неструктурованих даних, для яких на першому місці є швидкість обробки, а не повнота даних. Вибір цього інструменту залежить від конкретного задачі. Хоча NoSQL — це сфера, яка розширюється і кидає виклик багатьом припущенням, зробленим компаніями щодо підтримки застарілих систем, це надійний рух, який вирішує реальні проблеми, що виникають у зв'язку з великими даними.