УДК 004.6

Зелінський О.О., студент 3 курсу спеціальності 122 «Комп'ютерні науки»
Ніколюк П.К., професор кафедри Інформаційних технологій

ОСОБЛИВОСТІ РОЗРОБКИ БАЗ ДАННИХ ТА BIG DATA

Донецький національний університет імені Василя Стуса, м. Вінниця

Зростаючі обсяги інформації потребують апаратних і програмних засобів, здатних ефективно і швидко обробляти великі обсяги інформації, зі значним зниженням вартості збору, обробки, зберігання і передачі інформації. Ці процеси стали підґрунтям у новому та перспективному напряму розвитку інформаційних послуг – Big Data. Великі дані (Big Data) – це позначення структурованих и неструктурованих даних величезних обсягів і значного розмаїття, що піддаються ефективній обробці програмних інструментів, які горизонтально масштабуються та з'явились у кінці 2000-х років, і альтернативних традиційних систем управління базами даних і рішенням класу рішень Business Intelligence.[1]

Дані змінюють наш світ і спосіб життя з небувалою швидкістю. Великі дані - це нова наука аналізу та прогнозування поведінки людини та машини шляхом обробки дуже величезної кількості супутніх даних. Великі дані відносяться до швидкого зростання обсягу структурованих, напівструктурованих та неструктурованих даних. За оцінками, в 2018 році буде генеруватися 50 000 Gb даних в секунду. Швидкість, з якою дані генерували потребу, потрібно ефективно зберігати та обробляти. Великі дані породжуються з різних джерел і надходять у різних форматах. Big Data певним чином означає лише "всі дані". Великі дані можна описати через проблеми управління даними, які - завдяки збільшенню обсягу, швидкості та різноманітності даних - не можуть бути вирішені традиційними базами даних. Великі дані надходять від датчиків, пристроїв, відео / аудіо, мереж, файлів журналів, транзакційних програм, Інтернету та соціальних медіа –значна частина їх генерується в режимі реального часу та в дуже великих масштабах.[2]

БД - це сукупність пов'язаних даних. Існує два типи баз даних - система управління базами даних відношень, а інша - нереляційна система управління базами даних. Нереляційну базу даних також називають NoSQL. Ми зберігаємо різні типи даних у різних базах даних. Ми зберігаємо структуровані дані у реляційних базах даних. Існують різні типи реляційних баз даних, такі як SQL, Oracle, SQL Server, DB2, Teradata. Ми зберігаємо напівструктуровані або неструктуровані дані у нереляційних базах даних. Бази даних ми вибираємо на основі типів даних. Якщо ми зберігаємо і здатні обробляти дуже великий обсяг даних у базах даних, ми, безумовно, можемо зберігати та обробляти великі дані за допомогою реляційних або нереляційних баз даних. Ні, Big Data не замінить

бази даних. У тій чи іншій формі ми будемо використовувати бази даних SQL для зберігання та обробки великих даних. У зв'язку з цим Big Data повністю відокремлений від БД. [2]

Різниця між великими датами та базами даних:

- Big Data - термін, застосовуваний до наборів даних, розмір чи тип яких перевищує можливості традиційних реляційних баз даних. Традиційна база даних не здатна захоплювати, керувати та обробляти великий обсяг даних з низькою затримкою. Хоча База даних - це сукупність інформації, яка організована таким чином, щоб її можна було легко захоплювати, отримувати доступ, керувати і оновлювати.

- Big Data стосується технологій та ініціатив, які включають занадто різноманітні дані, тобто різновиди, швидкозмінні або масивні для навичок, звичайних технологій та інфраструктури, щоб ефективно вирішуватись. У той час як система управління базами даних (СУБД) витягує інформацію з бази даних у відповідь на запити, але це в обмежених умовах.

- Великі дані можуть бути будь-якими різновидами даних, тоді як БД можна визначити за допомогою якоїсь схеми.

- Великі дані важко зберігати та обробляти, тоді як Бази даних, як SQL, дані можна легко зберігати та обробляти. [2]

Big Data настільки популярний через такі характеристики:

- Об'єм: Обсяг, мабуть, найвідоміша характеристика великих даних. Як відомо, що майже 90% сучасних даних було створено за останні пару років. Обсяг відіграє головну роль при розгляді даних Big Data.
- Різноманітність: Коли ми говоримо про Big Data, нам потрібно враховувати дані в усіх форматах, як обробка структурованих, напівструктурованих та неструктурованих даних. Ми фіксуємо всі різновиди даних, будь то PDF, зображення, клік веб-сайту, зображення та відео. Ці змішані різновиди даних дуже важко зберігати та аналізувати.
- Швидкість: швидкість - швидкість або швидкість, з якою дані генеруються, клацаються, оновлюються, виробляються та отримуються доступ. Facebook генерує 500 Тб даних на день. YouTube завантажує 400 годин відео в хвилину. Google щодня переводить мільярди пошуків.
- Змінність: Невідповідність, показана даними часом, іноді сповільнить процес. Це декілька розмірів даних через безліч джерел даних.
- Вірогідність: Це стосується точності ваших даних. Наскільки точні ваші дані та наскільки значущі для аналізу на їх основі?[2]

Технології і тенденції роботи з Big Data:

Початково у сукупність підходів і технологій включались засоби масово-паралельної обробки невизначено-структурованих даних, такі як СУБД NoSQL, алгоритми MapReduce і засоби проекту Hadoop. У подальшому до технологій великих даних почали відносити й інші рішення, що забезпечують схожі за

характеристиками можливості обробки надвеликих масивів даних, а також деякі апаратні засоби.

MapReduce — модель розподілених обчислювань у комп'ютерних кластерах, представлена компанією Google. Згідно з цією моделлю, додаток розділяється на значну кількість однакових елементарних завдань, що виконуються на вузлах кластера і потім, природнім шляхом зводяться у кінцевий результат.

NoSQL (від англ. Not Only SQL, не лише SQL) — загальний термін для різних нереляційних баз даних і сховищ, не означає якусь конкретну технологію чи продукт. Звичайні реляційні бази даних добре підходять для досить швидких і однотипних запитів, а на складних і гнучко побудованих запитах, характерних для великих даних, навантаження перевищує розумні межі і використання СУБД стає неефективним.

Hadoop — набор утилітів, бібліотек і фреймворків, що вільно розповсюджується, для розробки і виконання розподілених програм, які працюють на кластерах із сотень і тисяч вузлів. Вважається однією з основоположних технологій більшості даних.

R — мова програмування для статистичної обробки даних і роботи з графікою. Широко використовується для аналізу даних і фактично став стандартом для статистичних програм.

Апаратні рішення. Корпорації Teradata, EMC та ін. др. пропонують апаратно-програмні комплекси, призначені для обробки великих даних. Ці комплекси поставляються як готові до установки телекомунікаційні шафи, що містять кластер серверів і керівне програмне забезпечення для масово-паралельної обробки. Сюди іноді відносять апаратні рішення для аналітичної обробки в оперативній пам'яті, зокрема, апаратно-програмні комплекси Hana компанії SAP і комплекс Exalytics компанії Oracle, незважаючи на те, що така обробка початково не є масово-паралельною, а об'єми оперативної пам'яті одного вузлаобмежуються кількома терабайтами.

Консалтингова компанія McKinsey, окрім технологій NoSQL, MapReduce, Hadoop, R, які розглядає більшість аналітиків, включає у контекст придатності для обробки великих даних також технології Business Intelligence і реляційні системи управління базами даних з підтримкою мови SQL.[1]

Список літератури

1. Технології і концепції Industry 4.0 [Електронний ресурс]. (*https://www.it.ua/knowledge-base/technology-innovation/big-data-bolshie-dannye*).

2. Є великі дані - це база даних? [Електронний ресурс]. (*https://uk.education-wiki.com/4905136-is-big-data-a-database*).