УДК 004.415.2.043

І. І. Бородій
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ВИБІР ТЕХНОЛОГІЇ ПРОЄКТУВАННЯ ПРОГРАМНОЇ СИСТЕМИ ФОРМУВАННЯ АГРЕГОВАНИХ НАДВЕЛИКИХ МАСИВІВ ДАНИХ

I. I. Borodii

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

Сучасні підприємства оперують надвеликими масивами даних (Big Data), які можуть вимірюватись петабайтами або навіть ексабайтами. Ці дані необхідні для відображення числових значень у агрегованому вигляді для виконання математичних операцій. Наприклад, для підприємства галузі сфери продажів важливими є такі агреговані дані як прибуток, кількість транзакцій або унікальних клієнтів в будь-якому часовому або іншому розрізі. Аналогічно, представлення даних в агрегованому вигляді допомагає компаніям спрогнозувати майбутній розвиток, приймати релевантні бізнес-рішення та контролювати фінансові показники.

Створення швидкої, надійної та стійкої до помилок програмної системи, що виконує збір, обробку та завантаження надвеликих об’ємів даних з подальшим їх агрегуванням є актуальною та важливою проблемою в сучасному світі. Перед розробкою програмної системи потрібно виконати етап проєктування, щоб обрати найбільш оптимізовані та релевантні технології для реалізації та розробити відповідну архітектуру системи.

Для створення програмної системи, що містить агреговані дані, використовують технології сховищ даних, які отримують дані з різноманітних джерел, що можуть зберігатися в абсолютно різних системах управління базами даних (СУБД). Міграція даних з однієї бази даних в іншу називається процесом вилучення, трансформації та завантаження даних ETL (Extract, Transform, Loading). ETL дозволяє інтегрувати дані з різних джерел в одному місці, щоб їх можна було обробляти, аналізувати та ділитися із зацікавленими сторонами [6]. Процес ETL повинен виконуватись періодично з метою забезпечення актуалізації даних.

Фахівці, що виконують задачі обробки та аналізу даних повинні мати доступ до усієї необхідної інформації та використовувати відповідні засоби для її відображення. Технології бізнес-інтелекту, сховищ даних та процесу ETL є спрямовані на вирішення цих проблем [1].

Процес збору, трансформації та завантаження надвеликих масивів даних в агреговане сховище даних може бути реалізований мовою програмування Python. Зокрема, Python є безкоштовною мовою програмування з читабельним синтаксисом, містить код з відкритим доступом, а також надає можливість працювати з різноманітною кількістю бібліотек. Однією з таких бібліотек є Pyspark, яка є рішенням для інтегрування Apache Spark в Python [3]. Spark дозволяє здійснювати реалізацію задач, які оперують надвеликими масивами даних і вимагають великої швидкості обробки даних [2].

Для побудови програмної системи агрегованих даних пропонується використати СУБД Apache Hive, яка оптимізована для аналізу надвеликих масивів даних. Можливості Apache Hive дозволяють аналізувати, перезаписувати та обробляти петабайти даних за допомогою унікальної структурованої мови запитів HiveQL, яка по суті є формою реалізації SQL. Важливо зазначити, що Hive - це нереляційна СУБД, що дозволяє організовувати дані в схожі таблиці на основі їх об’єму. Ці таблиці

складаються з окремих секцій (partitions), що дозволяє розділити таблиці на різні частини на основі інформації про дані. Ці секції можна розділити за допомогою процесу, який називається групування (bucketing). Цей процес розбиває дані з метою прискорення запитів до них [5]. Після виконання процесу ETL і завантаження даних в СУБД Hive, потрібно застосувати механізм представлення агрегованих даних. Для виконання HiveQL запитів релевантним рішенням буде використання системи запитів Trino. Це система розподілених SQL запитів, призначена для виконання запитів до надвеликих масивів даних, розподілених по одному чи кількох різнорідних джерелах даних [4]. Trino забезпечує обробку даних, що містять різноманітні формати файлів даних, використовуючи Hive. Використання Trino буде найефективнішим рішенням при проєктуванні програмної системи.

Для відображення агрегованих даних у читабельному та зручному для користувача вигляді, доцільно використовувати спеціалізовані засоби візуалізації даних. Одним з таких засобів є Microsoft Power BI, який запропоновано використовувати для проєктування програмної системи. Використовуючи Power BI, будуть розроблені SQL запити, які виконуватимуть агрегування даних з використанням системи запитів Trino.

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.

3. Мінухін С. В. Дослідження продуктивності кластера Apache Spark на платформі Azure для методів машинного навчання. / Мінухін С. В. // Збірник наукових праць Харківського національного університету Повітряних Сил. 2020. № 1(63). С. 81-88.

4. Trino 433 Documentation [Electronic resource] / trino – Access mode: https://trino.io/docs/current/overview.html

5. What Is Apache Hive? [Electronic resource] / Alex Williams – Access mode: https://builtin.com/data-science/apache-hive

6. What is ETL? (Extract, Transform, Load) Methodology & Use cases [Electronic resource] / Haziqa Sajid – Access mode: https://www.unite.ai/what-is-etl-methodology-and-use-cases