УДК 004.75

Федорович І. А., аспірант,
Осухівська Г. М., канд. техн. наук, доцент, доцент
кафедри комп'ютерних систем та мереж

ЗАСТОСУВАННЯ SPARK-RAPIDS ДЛЯ ЗДІЙСНЕННЯ GPU-ОБЧИСЛЕНЬ НА РУШІЇ APACHE SPARK

Тернопільський національний технічний університет імені Івана Пулюя, м. Тернопіль

Apache Spark – це рушій для розподіленого обчислення великих даних, який має підтримку декількох мов програмування (Python, Java, Scala) [1] та може бути запущений як на одному вузлі обробки (з паралелізацією на рівні потоків), так і в кластерному режимі (з розподілом задачі на різні вузли в режимі «драйвер – виконавці») [2]. Завдяки цим властивостям Apache Spark широко застосовується для побудови високопродуктивних систем аналізу великих даних, зокрема в задачах машинного навчання, обробки природної мови, потокової аналітики та бізнес-інтелекту.

Попри значну масштабованість Apache Spark, більшість обчислень у стандартних конфігураціях виконується на центральних процесорах (CPU). Це створює суттєві обмеження щодо продуктивності під час оброблення великих обсягів даних, особливо у випадках, де необхідно виконувати складні обчислення над великими масивами інформації. Зі зростанням обсягів даних та вимог до швидкодії класичні CPU-орієнтовані підходи недостатні для використання у застосунках, орієнтованих на інтенсивну обробку даних.

Графічні процесори (GPU) відзначаються високим рівнем паралелізму та здатністю виконувати тисячі однотипних операцій одночасно, що робить їх особливо ефективними для задач обробки великих масивів даних. Проте безпосередня інтеграція GPU в екосистему Apache Spark є нетривіальною задачею через відмінності в архітектурі обчислень, моделі пам'яті та необхідність оптимізації обміну даними між CPU та GPU. Саме ці фактори тривалий час стримували масове впровадження GPU у застосунках Spark.

Для розв'язання зазначених проблем компанією NVIDIA було розроблено бібліотеку NVIDIA RAPIDS Accelerator for Apache Spark (Spark-RAPIDS), яка надає можливість виконувати частину обчислювальних операторів Apache Spark безпосередньо на GPU за допомогою використання бібліотек RAPIDS [3]. Spark-RAPIDS реалізований у вигляді плагіна до Apache Spark, який автоматично підміняє деякі CPU-оператори на їх GPU-аналоги без необхідності суттєвої модифікації прикладного коду користувача. Це дає змогу розробникам використовувати переваги GPU-обчислень, зберігаючи знайомий інтерфейс роботи зі Spark.

Основною ідеєю Spark-RAPIDS є переведення обчислень у колонковий (columnar) формат із використанням GPU-пам'яті замість оперативної пам'яті процесора. Бібліотека забезпечує апаратно-прискорене виконання операцій фільтрації, агрегування, сортування, об'єднання таблиць (join), а також перетворення даних. Саме ці операції становлять основу більшості ETL-процедур та підготовки даних для подальшого аналізу й машинного навчання.

Застосування Spark-RAPIDS є особливо доцільним у задачах аналізу великих текстових масивів, потокової обробки інформації, обробки лог-файлів, новинних стрічок, даних із соціальних мереж та систем моніторингу. У таких застосуваннях значна частина часу витрачається саме на попередню обробку та агрегацію даних, тому перенесення цих етапів на GPU дає змогу досягти суттєвого скорочення загального часу виконання обчислювальних задач.

Перевагою Spark-RAPIDS є також те, що він допомагає масштабувати GPU-обчислення в межах кластерної інфраструктури, використовуючи кілька графічних прискорювачів на різних вузлах. Це відкриває нові можливості для побудови високопродуктивних розподілених систем аналізу даних, у яких GPU використовується не лише для машинного навчання, але й для класичних операцій обробки даних.

Водночас використання Spark-RAPIDS має і певні обмеження. Зокрема, не всі оператори Apache Spark наразі підтримуються у GPU-варіанті, що призводить до автоматичного повернення виконання окремих етапів на CPU [4]. До того ж ефективність GPU-прискорення значною мірою залежить від апаратної конфігурації, обсягу відеопам’яті та накладних витрат на передавання даних між центральною та графічною пам’яттю. У деяких випадках, особливо для невеликих обсягів даних, використання GPU може не дати суттєвого виграшу у продуктивності.

Отже, бібліотека Spark-RAPIDS є перспективним інструментом для інтеграції GPU-обчислень у середовище Apache Spark та може суттєво підвищити ефективність розподіленої обробки великих даних. Її використання є доцільним у задачах, де переважають масові однорідні обчислення та операції агрегації, зокрема в задачах потокової аналітики та обробки текстових даних. Подальші дослідження у цьому напрямі можуть бути спрямовані на експериментальну оцінку продуктивності Spark-RAPIDS у прикладних задачах аналізу природної мови та розроблення гібридних підходів, що поєднують можливості Spark-RAPIDS і користувацьких GPU-обчислювальних модулів.

Список використаних джерел

1. Comparative analysis of large data processing in Apache Spark using Java, Python and Scala / I. Borodii, I. Fedorovych, H. Osukhivska, D. Velychko, R. Butsii. *Proceedings of the 3rd International Workshop on Computer Information Technologies in Industry 4.0 (CITI 2025)* (11–12 June 2025, Ternopil, Ukraine). P. 189–198. URL: https://ceur-ws.org/Vol-4057/paper13.pdf

2. Apache Spark™ – Unified Engine for large-scale data analytics. URL: https://spark.apache.org/ (дата звернення: 27.11.2025).

3. Overview – Spark RAPIDS User Guide. URL: https://docs.nvidia.com/spark-rapids/user-guide/latest/overview.html (дата звернення: 27.11.2025).

4. Supported Operators spark-rapids. URL: https://nvidia.github.io/spark-rapids/docs/supported_ops.html (дата звернення: 27.11.2025).