

РОЗВ'ЯЗАННЯ ЗАДАЧІ ПАРАЛЕЛЬНИХ ОБЧИСЛЕНЬ ЗА ДОПОМОГОЮ АЛГОРИТМУ MAPREDUCE

Джадан Р.В.

Науковий керівник – доц. каф. КН, к.т.н. Хазін М.Б.

Впродовж останнього десятиріччя інформаційні технології зробили великий стрибок у своєму розвитку. Разом із цим зростають потреби людей у розв'язку все більш складних задач. Наприклад, розв'язання систем диференціальних рівнянь, розв'язання графів, обробка файлів статистики та їм подібних завдань. На допомогу приходять паралельні обчислення. З кожним роком об'єми даних, які необхідно обробити зростають. Очевидно, що для обробки великих об'ємів інформації одного надпотужного комп'ютера не достатньо. Саме тому для розв'язку таких задач застосовують методи паралельних обчислень. В зв'язку з вищевказаним розв'язок задачі побудови систем паралельних обчислень є актуальними і має важливе практичне значення.

В даній роботі розв'язано задачу обробки терабайтів лог-інформації та статистики із групуванням її по певним критеріям, з використанням алгоритму MapReduce.

MapReduce – це програмна модель (система паралельних обчислень). Концепція роботи алгоритму наступна: є вхідний потік даних довільного вигляду, який програмна модель ділить на частини і поступово подає їх на вхід деякій функції, яка називається `map()`. Вона обробляє ці дані і на виході видає декілька пар, кожна пара складається із ключа і значення (`key/value`). Наступним кроком є групування пар з однаковими ключами. Згруповані пари відправляються на вхід до ще однієї функції, ця функція називається `reduce()`. Вона приймає на вході ключ і якусь множину значень, що відповідають цьому ключу. На виході функція видає той самий ключ і деяке значення, що є певною агрегацією всіх вхідних значень. Те, що функція `reduce()` видає на виході, і є результатом роботи всієї програмної моделі.

Таким чином за допомогою алгоритму MapReduce було побудовано систему паралельних обчислень, що забезпечує наступні характеристики: автоматичний розподіл даних, стійкість до помилок, розподіл навантаження, оптимізацію мережевого та дискового обміну, високу продуктивність, моніторинг процесу.