УДК 004.4
А. Буковська
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ПАРАЛЕЛЬНЕ ТА РОЗПОДІЛЕНЕ ГЕНЕРУВАННЯ POWERSET З ВИКОРИСТАННЯМ ПЛАТФОРМИ ОБРОБКИ ВЕЛИКИХ ДАНИХ

UDC 004.6
A. Bukovska

PARALLEL AND DISTRIBUTED POWERSET GENERATION USING A BIG DATA PLATFORM

Новий напрямок досліджень – Великі дані. Прикладами гілки великих даних є структуровані і неструктуровані дані, медіа або випадкові процеси, оскільки вони практично не можуть бути оброблені традиційним способом. На зміну традиційним монолітним системам приходять нові асинхронні та паралельні рішення. Ці нові рішення забезпечують можливість роботи з великими даними [1].

Інформаційна технологія Big Data – це сукупність методів і засобів обробки різних типів структурованих (бази даних) і неструктурованих (текст, потік) динамічних великих обсягів даних для їх аналізу та використання для підтримки прийняття рішень.

Кластеризація є одним із способів зменшити часову складність обробки великих даних. Слід враховувати два варіанти масштабування, тобто горизонтальне та вертикальне масштабування. Горизонтальне масштабування ділить набір даних і розподіляє дані між кількома серверами або кластерами.

Основними технологіями обробки великих даних є: NoSQL; MapReduce; Apache Hadoop; Apache Spark.

Проблему збільшення обсягу інформації неможливо вирішити за допомогою класичних реляційних архітектур. Однією з проблем класичної реляційної бази даних є проблема роботи з масивними даними та проектами з високим навантаженням. Перший цільовий підхід полягає в тому, щоб розширити базу даних, якщо SQL досить гнучкий, а не переміщувати її, де б вона не виконувала свої завдання. Також реляційний підхід не підтримує обидва типи масштабування (вертикальне та горизонтальне).

Існують класичні підходи та парадигми розвитку засобів обробки даних. Однією з них є парадигма MapReduce [2]. Ця модель розподіленої обробки даних запропонована Google для обробки значного обсягу даних на обчислювальних кластерах. MapReduce забезпечує організацію даних у вигляді списків, які проходять 3 етапи обробки: етап карти (map stage), етап перемішування (shuffle stage), зменшення сцени (reduce stage).

Алгоритм Powerset є одним із алгоритмів, які використовуються для вилучення ознак у інтелектуальному аналізі даних [3], а також для генерування та вибору унікальних функцій [4].

Формування наборів потужностей за допомогою обробки MapReduce:

1. Partitioning S into subsets S[m].
2. for i ¼ 0tosizeof mdo.
3. if(0< S_m < m)S[i] = d//d is number of elements in subset.
4. END if.
5. END for.
6. Mapping S[m] to Mapper Machines.
7. Mapper (Input S[i], output P(S_i)).
8. E[S] = all possible combination of basic element for S[m].
9. END mapper.
10. Reducers receive sub-powerset.
11. Reducer (Input E[S_j], output P(S)).

12. Combine results from all tasks to find powerset using Union operation.
13. END Reducer.

Алгоритм генерує набір потужностей шляхом поділу заданого набору на підмножини та відображення підмножин серед різних картографів у кластер. Машини картографів обчислюють набір потужностей для кожної підмножини. Потім редуктор об'єднує піднабори ступенів, використовуючи операцію об'єднання вмісту набору (u), щоб створити остаточний набір ступенів набору S. Програма powerset (рис. 1) застосовується до кожної платформи.

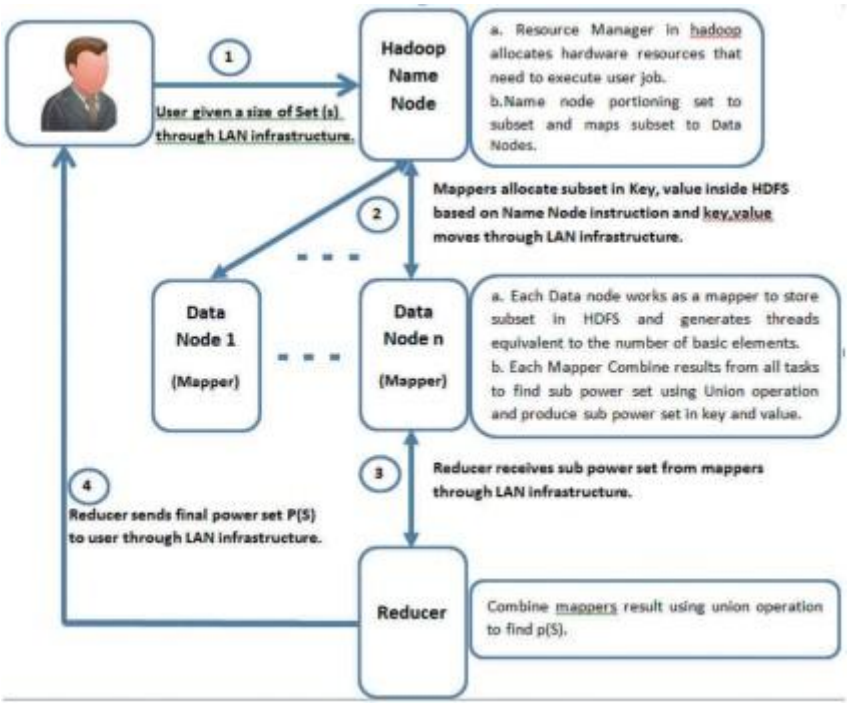

Рисунок 1. Створення Powerset за допомогою MapReduce

Це проста програма, яка створює набір даних розміру, заданого користувачем. Потім він розділяє набір на підмножини, які обробляються в паралельних і розподілених середовищах.

Література

1. Janssen, M., van der Voort, H., & Wahyudi, A. (2017). «Factors influencing big data decision-making quality». Journal of Business Research, 70: 338–345.
2. De Mauro, A., Greco, M., & Grimaldi, M. (2016). «A formal definition of Big Data based on its essential features». Library Review, 65 (3): 122–135.
3. IEEE International Conference on Cluster Computing (CLUSTER), 433–42, Taipei, Taiwan, IEEE. Esfandiari, M., R. Babavalian, et al. 2014.
4. Spolaôr, E.A.Cherman, and et al. 2013. ReliefF for multi-label feature selection. Brazilian Conference on Intelligent Systems, 6–11.