ФОРМУЛИРОВКА ЗАДАЧИ ОПТИМИЗАЦИИ ПРИЛОЖЕНИЙ, РАБОТАЮЩИХ С БОЛЬШИМИ ДАННЫМИ

Жук А. А.
Кафедра интеллектуальных информационных технологий, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: san91130324@gmail.com

Рассматривается проблемы постановки задачи оптимизации приложений, работающих с большими данным, специфика конвейеров обработки данных их типы, формулируются критерии оптимальности для задачи оптимизации приложений, работающих с большими данным

Введение

В связи с ростом количества приложений и технологий, работающих с большими данными, возникает потребность в решении задачи оптимизации данных приложений. Формулировка данной задачи и ее анализ является первым и важным шагом к построению эффективных приложений, работающих с большими данными

I. Конвейер обработки данных

В общем случае приложение, работающее с большими данными, можно представить как конвейер обработки данных, состоящий из блоков манипуляции (чтение, запись, трансформация) над данными. Блоки манипуляции могут выполнятся как строго последовательно (см. рис. 1) так и параллельно (см. рис. 2).

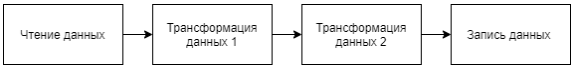

Рис. 1 – Пример конвейера обработки данных с последовательными блоками

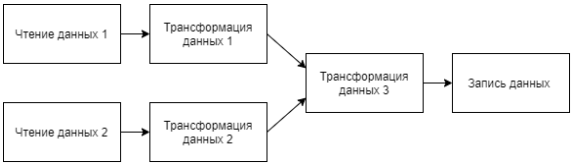

Рис. 2 – Пример конвейера обработки данных, имеющего блоки, которые могут выполнятся параллельно

Существует 2 типа конвейеров обработки данных:

- пакетные;
- потоковые.

Пакетные конвейеры обработки данных. Данный тип конвейеров обработки данных характеризуется тем, что работает с уже заранее готовыми наборами данных, хранящихся в определенных хранилищах данных.

Потоковые конвейеры обработки данных. Данный тип конвейеров обработки данных характеризуется тем, что работает с потоком данных в режиме реального времени.

II. Критерий оптимальности

В рамках задачи оптимизации приложений, работающих с большими данными можно выделить 2 критерия оптимальности: время обработки данных и затраченные ресурсы на обработку данных. Для разных типов конвейеров обработки данных эти критерии оптимальности будут формулироваться следующим образом.

Для пакетного конвейера обработки данных:

- время выполнения приложения;
- затраченные ресурсы на выполнение приложения.

Для потокового конвейера обработки данных:

- время обработки заранее определенного объема данных;
- затраченные ресурсы на обработку заранее определенного объема данных.

III. Формулировка задачи оптимизации для пакетного конвейера обработки данных

Из всего выше описанного можно обозначить, что существуют функции f_t и f_r, где f_t - функция зависимости времени выполнения конвейера обработки данных от входных параметров, вариантов реализации блоков этого конвейера и объема данных, которого необходимо обработать, а f_r - функция зависимости затраты ресурсов на выполнения конвейера обработки данных от входных параметров, вариантов реализации блоков этого конвейера и объема данных, которого необходимо обработать. Тогда можно обозначить и функции f_i и g_i, где f_i - функция зависимости времени выполнения i-го блока конвейера обработки данных от входных параметров, вариантов реализации блоков этого блока и объема данных, которого необходимо обработать, а g_i - функция зависимости затраты ресурсов на выполнения i-го блока конвейера обработки данных от входных параметров, вариантов реализации блоков этого блока и объема данных, которого необходимо обработать. Функция f_i и g_i будут равны:

$$f_i = f_i(\lambda_i, N_i, x_1, x_2, ..., x_k) \quad (1)$$

$$g_i = g_i(\lambda_i, N_i, x_1, x_2, ..., x_k) \quad (2)$$

λ_i - это объем данных, который необходимо обработать i-му блоку конвейера обработки данных, N_i - это вариант реализации i-го блока конвейера обработки данных $N_i \in \{N_{i1}, N_{i2}, ..., N_{iz}\}$, $x_1, x_2, ..., x_k$ - параметры конвейера обработки данных (количество узлов кластера хранения данных, количество оперативной памяти в кластере обработки данных и т.п.).

Тогда функция f_t принимает вид:

$$f_t = \phi(f_1(\lambda_1, N_1, x_1, x_2, ..., x_k), f_2(\lambda_2, N_2, x_1, x_2, ..., x_k), ..., f_n(\lambda_n, N_n, x_1, x_2, ..., x_k)) \quad (3)$$

ϕ - это функция суммирования значений функций f_i, n - это количество блоков обработки данных.

В случае конвейера обработки данных с последовательными блоками функция f_t принимает вид:

$$f_t = \sum_{i=1}^{n} f_i(\lambda_i, N_i, x_1, x_2, ..., x_k) \quad (4)$$

Функция f_r принимает вид:

$$f_r = \omega(g_1(\lambda_1, N_1, x_1, x_2, ..., x_k), g_2(\lambda_2, N_2, x_1, x_2, ..., x_k), ..., g_n(\lambda_n, N_n, x_1, x_2, ..., x_k)) \quad (5)$$

ω - это функция суммирования значений функций g_i, n - это количество параметров конвейера обработки данных. В большинстве случаев функция ω представляет собой сумму значений функций g_i и принимает вид:

$$f_r = \sum_{i=1}^{n} g_i(\lambda_i, N_i, x_1, x_2, ..., x_k) \quad (6)$$

Поскольку значения функций f_t и f_r имеют обратную зависимость и не могут в общем случае в задаче оптимизации рассматриваться независимо друг от друга, поэтому вводится функция β - функция сведения значений функций f_t и f_r к некоторому новому критерию оптимальности конвейера обработки данных. Тогда функция приобретает вид:

$$\beta = \beta(f_t(\lambda_1, \lambda_2, ..., \lambda_n, N_1, N_2, ..., N_n, x_1, x_2, ..., x_k), f_r(\lambda_1, \lambda_2, ..., \lambda_n, N_1, N_2, ..., N_n, x_1, x_2, ..., x_k)) \quad (7)$$

В этом случае возможно сформулировать задачу безусловной оптимизации (формула 8)

$$\min[\beta(\lambda_1, \lambda_2, ..., \lambda_n, N_1, N_2, ..., N_n, x_1, x_2, ..., x_k)] \quad (8)$$

Задача конкретизации функции β в большинстве случаев лежит на эксперте предметной области, в рамках которой разрабатывается конвейер обработки данных, так как именно он обладает знанием о том сколько ресурсов можно потратить и за какое время надо обработать данные. В случае когда функция β не может быть четко определена, можно сформулировать задачу условной оптимизации (условие 9)

$$\begin{cases} f_t = f_t(\lambda_1, \lambda_2, ..., \lambda_n, \\ N_1, N_2, ..., N_n, x_1, x_2, ..., x_k) \le T \\ f_r = f_r(\lambda_1, \lambda_2, ..., \lambda_n, \\ N_1, N_2, ..., N_n, x_1, x_2, ..., x_k) \le R \end{cases} \quad (9)$$

T - это максимально допустимое значение времени выполнения конвейера обработки данных, R - это максимально допустимое значение затрат ресурсов на выполнение конвейера обработки данных. Данные параметры формулируются экспертом предметной области, в рамках которой разрабатывается конвейер, поэтому нет гарантий, что при заданных T и R существует такие параметры $N_1, N_2, ..., N_n, x_1, x_2, ..., x_k$ при которых условие 9 будет выполнятся. В случае если решения не существует, то надо пересмотреть значения T и R, если пересмотр значений T и R не привел к решению поставленной задачи, то можно сделать вывод о том что при текущем техническом оснащении невозможно удовлетворить потребность, которую призвано решить разрабатываемое приложение.

IV. Формулировка задачи оптимизации для потокового конвейера обработки данных

Аналогично, как и для задачи оптимизации пакетного конвейера обработки данных, формулируется задача для потокового конвейера обработки данные, единственным отличием для потоковых конвейеров является то, что функции f_t и f_r базируются не на времени и затратах на выполнения конвейера обработки данных, а времени и затратах на обработку некоторого заранее определенного объема данных.

V. Выводы

Из сформулированной задачи оптимизации конвейера обработки данных можно сделать вывод о том, что качество оптимизации конвейера обработки данных зависит от того насколько была проделанная работа по определению значений T и R, и по поиску всех возможных значений параметров $N_1, N_2, ..., N_n, x_1, x_2, ..., x_k$. Так же стоит отметить что приложение оптимальное для обработки объема данных λ_1 может быть не оптимальным для обработки данных объема λ_2, поэтому всегда стоит заранее обсуждать вопрос текущего объема данных и то до какого объема он может вырасти с экспертом предметной области.

1. Дэн Саймон. Алгоритмы эволюционной оптимизации. / Дэн Саймон // 2020. – 1002 с.