Zagadka - MP (6)

Wojciech Szlosek

April 2020

1 Sortowanie dużego pliku

Zauważmy, że plik z zadania ma "tylko" 1gb. Jest to i dużo, i bardzo mało (patrząc na dzisiejsze sprzętowe i pamięciowe możliwości). Przyjmijmy jednak, że ilość ta znacząco obciąża nasze możliwości.
Chcąc posortoważ tak duży plik warto zainteresować się "sortowaniem zewnętrznym" (external sorting). Zewnętrzne sortowanie to klasa algorytmów sortowania, które mogą obsługiwać ogromne ilości danych. Zewnętrzne sortowanie jest wymagane, gdy sortowane dane nie mieszczą się w głównej pamięci urządzenia komputerowego. Jednym z przykładów zewnętrznego sortowania jest zewnętrzny algorytm sortowania scalającego, który jest algorytmem scalania metodą K-way. Sortuje fragmenty, które mieszczą się w pamięci RAM, a następnie je scala. By skrócić nieco rozważania przedstawię ogólny pomysł na takie sortowanie z treści zadania (przyjmijmy dla uproszczenia że 1 gb = 1000 MB, a do dyspozycji mamy jedynie 200 MB pamięci RAM).
Nasz plik zawiera jedynie litery i cyfry - przyjmijmy strategię, że chcemy sortować zgodnie z liczbą w kodzie ASCII (najpierw cyfry, potem duże, na końcu małe litery - leksykograficznie).

KROK 1
Odczytajmy 200 MB danych i posortujmy je (np. QuickSortem). Dane zapiszmy na dysku.

KROK 2
Powtarzaj krok 1, aż wszystkie dane zostaną posortowane (mamy $\frac{1000}{200} = 5$ fragmentów).

KROK 3
Przeczytaj pierwsze 40 MB każdej posortowanej porcji do buforów wejściowych w pamięci głównej i przydziel pozostałe 40 MB na bufor wyjściowy.

KROK 4
Wykonaj scalenie w 5 kierunkach i zapisz wynik w buforze wyjściowym. Kiedy bufor wyjściowy się zapełni, zapisz go do końcowego posortowanego pliku i

opróżnij. Ilekroć jeden z 5 buforów wejściowych zostanie opróżniony, wypełnij go następnymi danymi, aż tych nie będzie już dostępnych.

2 Sposoby na optymalizację (kompresję) pliku:

Sposobów na kompresję pliku z polecenia może być wiele. Doszedłem do wniosku, że skoro liczba znaków jest względnie niewielka (litery + 9 cyfr), a tychże jest dużo (skoro plik ma 1 GB), więc dobrym pomysłem jest optymalizacja długości poprzez zauważenie, że uporządkowane dane składają się z wielu duplikatów będących obok siebie.
Dla przykładu: "...YZZZZZaaaaaabbb..." możemy zamienić na: "...{5}Z{6}a..." gdzie liczba w klamrach oznacza ilość wystąpień (obok siebie) znaku po niej występującego.
Wymaga to (np. co najwyżej liniowego) zliczania wystąpień danego znaku, ale ostatecznie - warto. Do tego kompresja np. przez pakowanie ZIP i w ten sposób otrzymujemy zoptymalizowany rezultat naszej pracy.