External Sorting

How would you sort a large file (20GB+)?

External Sorting

Let’s simplify this problem to just characters…

Imagine we have 5 Arrays of 5 Characters each:

['c', 'e', 'a', 'd', 'b'], ['a', 'b', 'e', 'd', 'c'], ['d', 'c', 'e', 'b', 'a'], ['e', 'c', 'b', 'd', 'a'], ['d', 'c', 'a', 'b', 'e']

And we want to make it look like this:

['a', 'a', 'a', 'a', 'a', 'b','b', 'b', 'b', 'b', 'c', 'c','c', 'c', 'c', 'd', 'd', 'd','d', 'd', 'e', 'e', 'e', 'e','e']

However our “RAM” Array can only hold 5 characters at a time:

[empty, empty, empty, empty, empty]

How in the world do we sort all 25 characters, if we can only sort 5 characters at a time?

External Sorting

Initial:

['c', 'e', 'a', 'd', 'b'], ['a', 'b', 'e', 'd', 'c'], ['d', 'c', 'e', 'b', 'a'], ['e', 'c', 'b', 'd', 'a'], ['d', 'c', 'a', 'b', 'e']

Final:

['a', 'a', 'a', 'a', 'a', 'b','b', 'b', 'b', 'b', 'c', 'c','c', 'c', 'c', 'd', 'd', 'd','d', 'd', 'e', 'e', 'e', 'e','e']

Step 1: Sort each individual Array

['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e']

External Sorting

Initial:

['c', 'e', 'a', 'd', 'b'], ['a', 'b', 'e', 'd', 'c'], ['d', 'c', 'e', 'b', 'a'], ['e', 'c', 'b', 'd', 'a'], ['d', 'c', 'a', 'b', 'e']

Final:

['a', 'a', 'a', 'a', 'a', 'b','b', 'b', 'b', 'b', 'c', 'c','c', 'c', 'c', 'd', 'd', 'd','d', 'd', 'e', 'e', 'e', 'e','e']

Step 1: Sort each individual Array

['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e']

Step 2: 5 Way Merge

['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e'], ['a', 'b', 'c', 'd', 'e']

[empty, empty, empty, empty, empty]

['a', 'a', 'a', 'a', 'a'] → Write to Disk

['a', 'a', 'a', 'a', 'a', 'b','b', 'b', 'b', 'b', 'c', 'c','c', 'c', 'c', 'd', 'd', 'd','d', 'd', 'e', 'e', 'e', 'e','e']

External Sorting (EG: Ram is 5 Gb)

20 GB Unsorted File → (5 GB F1), (5 GB F2), (5 GB F3), (5 GB F4)

RAM → Sort

(5 GB F1_Sorted), (5 GB F2_Sorted), (5 GB F3_Sorted), (5 GB F4_Sorted)

Ram → K Way Merge

(5 GB Output_1), (5 GB Output_2), (5 GB Output_3), (5 GB Output_4)

20 GB **Sorted** File

What we just did here is an Algorithm called External Sort.

External Sorting (EG: Ram is 5 Gb)

With External Sort, we managed to sort a 20 GB file, even though we just had 5 GB of RAM!

Again we broke the file into smaller parts, sorted those parts, and then did a K Way Merge. This is a divide and conquer approach.

External Sort was extremely useful in past, when RAM was hard to come by, and programmers still had to sort large amounts of data.