Gaurav Singh | UFID: 36398850

COP 6726: Database Systems Implementation
Spring 2018
Weekly Assignment 4

02-13-2018

- External Sort
- Would sorting tuples help?
- It would help with sub-full scan. Only helps figure out early stopping
- It wouldn't help with selection normally. So, sorting won't help with selection normally.
- Eliminating duplicates would become easier as you could eliminate them in step of sorting itself, don't have to pay extra cycles to find duplicates too.
- You typically sort in lexicographic order, ie. Column wise.
- Now, that doesn't mean columns have to be in a particular order… you can pick any order to read the columns like you can decide to read columns out of order, this ties back into the CPU trying to run instructions out of order, this idea again leverages the same fundamentals to ensure optimum usage of resources.
- Now removing duplicates locally can be efficient but removing duplicates that are really far apart would be hard to do as they may not reside in memory at the same time.
- There in comes in the idea sorting clusters. You try to keep similar values in one place.
- Don't worry too much which sorting method exactly are we going to use. Cause sorting is a sport, and there are so many ways to sort. Each may do something a little different to push just a little bit better performance in a particular computer.
- Imagine you want to find minimum across a trillion records. You don't want to sort through the entire line up one records at a time. Like, you want to scan all the records at once and just write the answer to the file rather than scan one record, write minimum to the file. Cause asking so many RW would cost more time.

- $for\ \ i\ \in [b{:}n-1]$
  $min\ =\ (min > A[i])?\, A[i]{:}\, min$

- This would scan in a single go and write to output directly.

02/15/2018

- Vast majority of joins are Foreign Key Joins and thus avoid cross product.
- In these joins, you basically follow the chain of foreign keys to the data.
- It also means, in virtually most of the result sets of following foreign keys, would be a most a single record.
- Thus you can probably make better join search by keeping dimension tables always in memory

- Now, sometimes you are forced to actually compute the Cross Product.
- It makes no sense but if the user requests it, then you have to produce it.
- Now there are ways in which figuring out cross products would be more manageable
- Nested Loop Joins helps figure out these cross products, in here you typically run two nested for loops, now it matters which element you put in the other loop and which one you put in the inner one.
- Its extremely hard to pump data into a multicore system at 4Ghz