COP 6726: Database Systems Implementation

Spring 2018

Weekly Assignment 6

27-02-2018:

Group By

* Its basically an operation run on the entire column at a time.
* You need to keep the process open until the whole column is processed.

Duplicate elimination

* Check if it is the hash, if not pass it on.
* Hash:
  + Build a hash
  + Look item in hash
  + Scan the hash (enumeration but random)
* 1st Build Hash -> Full Block Inversion
* R ^ S : Scan(Hr),Lookup(Hs) (can scan anyone and look the other one)
* Now consider we are building hashes as we are working
* Its valuable to provide data structure with trickle-ish data
* That ways we wouldn’t overwhelm ourselves with data and can parallelize everything.
* You need to design hash, so you can scan in parallel or do something else entirely.
* There are no set rules, you can break any of the above rule.
* Join algorithm is a generalized intersection algorithm.
* For blocking algorithm, you can get away with hashing just one of the tables.
* We can produce a join like duplicate elimination algorithm.
* What if the even the smaller relation doesn’t fit in memory?
  + Approach 1: mimics sorting, external hash algo
  + Approach 2: Adventurer algo.
* Sorting removes a lot of complicated.
* Sorting sells this idea too far.

01-03-2018

* Generalized Linear Aggregation
* To the system, all GLAs look the same. They are just some magic boxes with produce results.
* The magic of these black boxes is not only can they run specific operations on the data, they can parallelize most of the work.
* Lets talk about GLAs Abstract Data Type.
* Abstract separate interface from implementation.
* GLA as an AST
* GLAs must have some guarantee to make sure that they are generalized.
  + Associativity
  + Communicativity
* OrderBy : Think about
  + What the current state is?
  + How do we compare?
  + How do we merge?