COP 6726: Database Systems Implementation

Spring 2018

Weekly Assignment 1

01-23-2018

* Asm.js is not just faster at processing, it is faster to compute.
* Web assembly is a further extension of the idea of ASM.js and applies it to the idea chain of commands
* Data Representation on Disks
  + JSON: Inter computer communication
  + YAML: for DB
  + XML: For old people … java
* JSON YAML AND XML can be converted to each other by a lot of products nowadays.
* XML died out because it’s too verbose
* BSON: Binary JSON: In case you are not implementing string. It basically removes text, we have picture perfect representation of numbers.
* CSV: Large tables are shared primarily in this format
* How do I find the simplest packaging for my data? That’s the question we always try to answer.
* Column Store and Row store are two different ideas of storing data on disk
* Idea of defining a good Disk Block is very important.
* Disk Block truly ingrained in everything.
* Its 4KB (default) for disks … although it should be waaaaay more
* Core data structure for a DB is a tuple.
* Tuples are packed into Disk Blocks.
* Pack Disk Blocks into some fancier Meta Structure
* How do we read these disk blocks?
* First DB to successfully implement column store is C Store -> Sold Vertica
* Virtually all column store based DBs are bad for transactional DB.
* They are good for analytics.
* In column store you have to provide exceptional column processing.
* You have to deal with an entire Disk Block at a time even if you want to alter a single tuple/record.
* DD command in Linux is dangerous but lets us showcase how disk blocks affect performance.
* 4KB for SSDs is a terrible block size, 128 KB showed by professor is significantly better (order of magnitude 1 or more)

01-25-2018

* Code for project1 can be modernized, but it has mode components coming in soon, be careful if you intend to change lot of it.
* Column store can be hard on memory are you have to at least maintain 3 in memory at any given time, two from where you compare data and a 3rd one where you write a result.
* Columnstore are more compressible though, especially if they are simple and predictable.
* Certain operations are faster on column.
* Disk speeds
  + Spindle disk < - 200 MBps
  + SSDs?? Many promises but are fairly deceiving at times
* There are 16 TB SSDs from Samsung, already on the market.
* Platters spin at 5400, 7200,10000,15000 RPM.
* Best HDDs need 2ms to position the head, that’s very slow.
* Move from track to track .. 0.1 ms
* Random Read repositioning delay : 2-10 ms
* Typically data is stored in cylinder
* Sector size is ½ KB
* Speed of head is typically a physical limit.
* Nowadays sectors are of same size irrespective of position on platter.
* Disk block in HDD is 32KB
* HDDs like large files.