Gurjus Singh

MSDS 432 Foundations of Data Engineering

July 12th, 2020

Module 3 – Reading Comprehension

1. **What is a B-Tree and how does it work? Explain and give an example.**

A B-Tree is used for indexing data in a database and are known as the standard way to implement an index. They store data in a key-value pair and are usually sorted by key. The database by key is stored as blocks in the B-Tree structure. The typical block size is 4 KB. The root node of the B-Tree is for one page, this will lead to traversing through the children till finding the specific key for the reference to a location of data in the database.

2. **What is Column Oriented Storage and how is it different than an RDBMS? Give an example.**

A column-oriented storage is based on storing all the values from each column together. This is different from a typical database because usually the RDBMS stores all of the values from each row together. When you want to get the values of a specific row you can find the 1st element of each column which corresponds to the first row.
For example, in Figure 3-10 it shows data of each column:

Columnar storage layout:

date_key file contents:	140102, 140102, 140102, 140102, 140103, 140103, 140103, 140103
product_sk file contents:	69, 69, 69, 74, 31, 31, 31, 31
store_sk file contents:	4, 5, 5, 3, 2, 3, 3, 8
promotion_sk file contents:	NULL, 19, NULL, 23, NULL, NULL, 21, NULL
customer_sk file contents:	NULL, NULL, 191, 202, NULL, NULL, 123, 233
quantity file contents:	1, 3, 1, 5, 1, 3, 1, 1
net_price file contents:	13.99, 14.99, 14.99, 0.99, 2.49, 14.99, 49.99, 0.99
discount_price file contents:	13.99, 9.99, 14.99, 0.89, 2.49, 9.99, 39.99, 0.99

Figure 3-10. Storing relational data by column, rather than by row.

For example, the 1st row would have 140102, 69, 4, NULL, NULL, 1, 13.99, 13.99 as values.

3. What is an SSTable and how does it work? What are some of the advantages over log segments with hash indexes?

A SSTable is a sequence of key-value pairs that are sorted by key. The advantages it has over log segments is that it is more efficient and simpler to merge segments. This uses an algorithm called merge sort, to sort an array. If multiple segments contain the same key, keep the most recent segment. The other advantage is that you do not need to keep the location of all the keys. You can keep the locations that are at a relative location to that key. Also, the segments in the database can also be grouped into a block to allow faster reads with SSTables.

4. What is tail recursion and how is it different than regular recursion?

The difference between regular recursion and tail recursion is that in tail recursion the recursive call to the function occurs at the end, no other statements are after the recursive call. In regular recursion, there is a base case, and the simple recursive call is done in the return statement with the expression to do recursion on such as in the factorial example from the textbook. A thing about using tail-recursion is use save space on the stack which is another difference. This allows optimization of the compiler.