CS3743 Hwk 2 Physical Performance (30 pts) – due 2/8

Consider a MovieInterest table that contains the following attributes:

custNr customer number, 8 characters, uniquely defines a customer. When combined with movieCd, it is the unique key for MovieInterest.

movieCd 12 character code (e.g., COMDUMB, COMROMANCE, SCIFISPACE). When combined with custNr, this is the unique key for this table.

locState customer's location state, 2 characters, 50 distinct values

locZip customer's location zip+4, 9 characters, most duplicates are only because of the same customer

lastActivityDt 4 byte date representing the last activity for this row

custNm customer name, max of 40 characters, about 0.01% are duplicates with other customers. Names average 20 characters.

streetAddress customer's street address, max of 150 characters. These average 40 bytes.

Assumptions:

1. There are ten million rows in the MovieInterest Table.
2. Data and index blocks are 4096 bytes.
3. An index entry for a B+Tree contains: the key (size depends on the contents) and a Row Id which is 4 bytes.
4. There are two variable-sized attributes (custNm and streetAddress). Each will use a one byte length. Near the beginning of the row, there will be an additional 2 byte offset to access the streetAddress that follows the custNm.
5. Assume there are 50,000 customers in TX. 2000 of those are interested in SCIFISPACE.
6. Each customer will have an average of 5 rows in the table since customers have an average number of movie codes of 5.

Part 1: Space (show your work for partial credit)

Assumptions: All my assumptions for Part 1 are on the third page.

1. How big is one data row? You must consider:
   * size of each fixed-size attribute
   * For each variable-sized attribute, consider the average size plus one for the length. Also, each attribute after the first variable-sized attribute must have a two byte offset.
   * include a two byte row size per row

Average row size = 101 bytes

1. How many rows fit in a data block? You must consider:
   * average size of a row (from part 1A)
   * 20 byte overhead in the data block. This isn't per row.
   * two bytes for each entry in the line array per row.

Rows per data block = 39 rows per block

1. How many data blocks are needed?

Number of data blocks = 256,411 data blocks

1. For a unique index (called I1) on the combination of custNr and movieCd:
   * what is the size of an index entry? 24 bytes
   * how many entries fit in an index block if we assume 20 bytes of overhead per index block? 169 entries/block
   * how many leaf level index blocks are needed? 59,172 leaf level index blocks
   * how many index entries are needed for the next higher level? 351 blocks on the next level
2. For index I1, how many total **index blocks** are needed including the leaf level?

Leaf level = 59,172 (from part 1D, 3rd ans)  
Next Level = 351 (from part 1D, 4th ans)

Next Level = 3

Root Level = 1

========

59,527 total index blocks for I1

Part 2: Reads (show your work for partial credit)

Additional assumptions:

1. We have another index(called **I2**) on the combination of stateCd , movieCd, and rowId:
   * The index nodes for this index have **226** entries per index block.
   * There are **three** index levels.
2. Do not worry if your answer to this is different in part 1. These values are arbitrary:

* Assume an average of **40** data rows per data block.
* Assume **four** index levels in the unique index, **I1**, on custNr and movieCd.

1. Assume there are 50,000 customers in TX. 2000 of those are interested in SCIFISPACE.

Questions:

For A-E below, we want to find all customers in **TX who are interested in SCIFISPACE** using the index **I2**. State your assumptions.

**Assumption:** Here we know the state and movie code so we have to read all the possible entries of people who live in Texas that happen to like SCIFISPACE, 2,000 Texans interested \* 5 rows per customer = 10,000 entries to read

1. We have three levels in index I2 with a total of 10,000,000 index entries.
   * How many entries in index I2 will satisfy this request based on the assumptions?   
     Number of entries in I2 matching the criteria = 2,000
2. How many leaf level index blocks are there satisfying this request? (approximate)  
   Leaf level index blocks = 10,000,000 index entries / 226 entries per index block = 44,248 leaf level index blocks
3. How many index blocks (in total) are read to satisfy this request?
   * It isn't necessary to re-read index blocks when processing this query.

Leaf level reads = 44,248 (from part 2B)

Level 2 reads = 44,248 index entries / 226 entries per index block = 196

Level 1 reads (root level) = (195 index entries / 226 entries per index block) = 1

Total = 44,248 + 196 + 1 = 44,445

1. Assuming **I1** is the clustering index, how many total reads are necessary to satisfy the query?

Index reads = 3 index reads (from Part 2C)

Data reads = 10,000 entries / 40 Entries per Block = 250 data reads(state any valid assumptions)  
Total = 250 + 3 = 253

1. Assuming **I2** is the clustering index, how many total reads are necessary to satisfy the query?

Index reads = 3 index reads

Data Reads = 10,000 entries / 226 entries per index block = 45 data reads  
Total = 45 + 3 = 48

For F and G, how many reads are necessary to find a particular row by the combination of custNr and movieCd using index I1?

**Assumption:** Here we have no way of knowing what state they live in so we have to read all the entries for data reads. I assumed if this were the case we would need to read all the SCIFISPACE entries, but we don’t know the total number of people who are interested in SCIFISPACE out of the 50 distinct state codes.

1. Assuming **I1** is the clustering index, how many total reads are necessary to find a particular row using **I1**?

Index reads = 4 index reads

Data Reads = 10,000,000 entries/40 entries per block = 250,000 data reads  
Total = 250,000 + 4 = 250,004

1. Assuming **I2** is the clustering index, how many total reads are necessary to find a particular row using **I1**?

Index reads = 4 index reads

Data Reads = 10,000,000 entries / 226 entries per block = 44,248 data reads

Total = 44,248 + 4 = 44,252