**CSDS 234 Structured and Unstructured Data  
Department of Computer & Data Sciences, Fall 2024**

**Assignment 3**

1. **[Query cost] [25]** Given two tables r and s, where r contains **nr** tuples, s contains **ns** tuples, and **ns>nr>0**. Give the maximum and minimum possible number of tuples for the result produced by the following relational algebra expressions. For each case, also state the assumptions of the schema of r (denoted as **R**) and the schema of s (denoted as **S**) needed to make the expression and results meaningful.
2. r ∪ s
   1. Assumptions:
      1. The schemas must be the same (identical domains): R = S
   2. Maximum Number of Tuples:
      1. nr + ns
   3. Minimum number of Tuples:
      1. nr
3. r ∩ s
   1. Assumptions:
      1. Both schemas (r and s) must be identical for the intersection
   2. Maximum number of Tuples:
      1. nr
   3. Minimum number of Tuples:
      1. 0
4. r – s
   1. Assumptions:
      1. Both schemas must be identical
   2. Maximum number of Tuples:
      1. nr
   3. Minimum number of Tuples:
      1. 0
5. r × s
   1. Assumptions:
      1. There are no schema restrictions for the product
   2. Maximum number of Tuples:
      1. nr x ns
   3. Minimum number of Tuples:
      1. nr x ns
6. σA=2(r)
   1. Assumptions:
      1. The attribute must exist in the schema of r
   2. Maximum number of Tuples:
      1. nr
   3. Minimum number of Tuples:
      1. 0
7. **[Storage] [10]** You are given a hard disk with in total 6 platters. Each platter is double-sided (i.e., has two surfaces that can store data). Each surface has 2K tracks. Each track has 50 sectors. Each sector can store 512 bytes data. Answer the following questions.   
   1. What is the capacity of a single track (the size of the data it can store, in bytes)? A single platter? The entire hard disk?
      * Capacity of a single track:
        + 50 \* 512 = 25600 bytes
      * Capacity of a single platter:
        + Total capacity for the surface = 2000 \* 25600 = 51200000 bytes
        + Capacity for the platter = 2 \* 51200000 = 102400000 bytes
      * Capacity of the hard disk:
        + Capacity of the hard disk = 6 \* 102400000 = 614400000 bytes
   2. Identify the valid single block sizes from the following numbers and explain why the rest are not valid: *280 bytes, 1540 bytes, 2048 bytes, 26112 bytes.*
      * The sector size is 512 bytes. Only the blocks that are integers of the sector size can be valid points
      * The only valid block sizes are 2048 (2048/512 = 4) and 26112 (26112/512 = 51)
      * The invalid block sizes are 280 (280/512 = 0.527) and 1540 (1540/512 = 3.0078)
8. **[B+ tree Index] [15]** Consider the following B+ tree (partly shown). Leaf nodes are linked (not shown in the figure). The nodes are annotated with a name (e.g., “I1” for intermediate node 1 and “L2” for a leaf node 2). Answering the following questions.

Diagram

Description automatically generated

1. Give the names of all the tree nodes that must be fetched to answer the query: “find all records with search key greater than 51.”
   1. I1 (Root Node), I2, I3 (Intermediate Nodes), L4, L5, L6, L7, L8 (Leaf Nodes)
2. Give the names of all the tree nodes that must be fetched to answer the query: “find all records with search key at least 38 and no larger than 90.”
   1. I1 (Root Node), I2, I3 (Intermediate Nodes), L2, L3, L4, L5, L6 (Leaf Nodes)
3. Subtree B is not shown. Nonetheless, what can you infer about this subtree's contents and shape? Put down as much as you can.
   1. Subtree B contains the keys that between 10 (inclusive) and 20.
   2. All of the leaf nodes will be at the same depth
   3. The nodes will be in sorted order
   4. It will have approximately the same amount of leaf nodes as the shown part of the tree
4. **[Query Processing -Join] [25]** Let the schema of a relation r as R(A,B,C), and a relation s has schema S(C,D,E). Relation table r has 40K tuples, relation s has 60K tuples. The block factor of r is 25. The block factor of s is 30. Let the average block transfer time is tT. Assume you have a memory that contains M data blocks, but M< 40K/25. Assume there is no index. Estimate the worst-case costs, in terms of the total number of data blocks to be transferred, for the following query r  s using each of the following join strategies:

R has 40k tuples (the number of blocks is 1600)

S has 60K tuples (the number of blocks is 2000)

R’s block factor is 25 (1 block holds 25 tuples)

S’s block factor is 30 (1 block holds 30 tuples)

The block transfer time is tT

Available memory: M blocks (M < 40K/25 = 1600 blocks)

No index’s

* + - Nested Loop Join
      * For each r, scan all s
      * Worst case scenario: Scan all blocks of r
      * Total Cost = 1600 + (1600\*2000) = 1600 + 3200000 = 3201600 block transfers
    - Block Nested Loop Join
      * M blocks of r into memory and join with all blocks of s
      * Worst case scenario: Read all blocks of s for every M blocks of r
      * Total Cost = 1600 + ((1600/M) \* 2000)
    - Hash join
      * Partition both r and s, scan all blocks once during partition. Join matching partitions
        + Cost for scanning partitions = 1600 + 2000 = 3600 block transfers
        + Cost for joining partitions = 1600 + 2000 = 3600 block transfers
      * Total Cost = 3600 + 3600 = 7200 block transfers

1. **[XML] [15] XML/DTD] [25]** Consider the following XML for Solar System.

<solar\_system>

<star>

<name>Sun</name>

<spectral\_type>G2</spectral\_type>

<age unit="billions years">5</age>

</star>

<planet type="telluric">

<name>Earth</name>

<distance unit="km">149600000</distance>

<mass unit="kg">5.98e24</mass>

<diameter unit="km">12756</diameter>

<satellite number="1"/>

</planet>

<planet type = “gaseous”>

<name>Saturn</name>

<distance unit =“AU”>5.2</distance>

<mass unit= “Earth mass”>95</mass>

<diameter unit = “Earth Diameter”>9.4</diameter>

<rings>Yes</rings>

<satellite number = “18”/>

</planet>

<planet type = “gaseous”>

<name>Uranus</name>

<distance unit =“AU”>19.2</distance>

<mass unit= “Earth mass”>14.5</mass>

<diameter unit = “Earth Diameter”>4</diameter>

<rings>Yes</rings>

<satellite number = “15”/>

</planet>

</solar\_system>

Add a few lines to the above XML code to include all the facts below about Saturn and Uranus. Your XML needs to be well-formed.

*Both Saturn and Uranus are planets in Solar system. Both Saturn and Uranus have rings. Saturn and Uranus both have type “gaseous”. Saturn is a planet of distance 5.2 (in unit “UA” – astronomical unit), with an Earth mass 95 (unit = “Earth mass”), and a diameter 9.4 Earth diameter. It has 18 satellites. Uranus is a planet of distance 19.2 UA, 14.5 Earth mass, and a diameter of 4 Earth diameter. It has in total 15 satellites.*