CSE 232A Fall 2019 Midterm Exam 2
Answers

Part 1. [11pts] Let us continue shopping! Given the following simplified relational database schema for shopping, indicate which of the following relational algebra queries is logically equivalent to the given query (only one option is correct) in each question except the last one.

Products (PID, PName, Vendor, Price)
Customers (CID, CName, Zipcode, Age)
Purchases (CID, PID, Timestamp, Quantity)

Products.PID and Customers.CID are primary keys in their respective tables. In Purchases, CID and PID are foreign keys referring to Customers.CID and Products.PID, respectively. We use aliases P for Products, C for Customers, and R for Purchases.

1. [2pts] \( \pi_{Vendor}(\sigma_{Price>50}(P)) \)

   (A) \( \sigma_{Price>50}(\pi_{Vendor}(P)) \)

   (B) \( \pi_{Vendor,Price}(\sigma_{Price>50}(P)) \)

   (C) \( \pi_{Vendor}(\gamma_{Price>50}(P)) \)

   (D) \( \pi_{Vendor}(\sigma_{Price>50}(\pi_{Vendor,Price}(P))) \)

   (E) \( \sigma_{Price>50}(\pi_{Vendor,Price}(P)) \)

   ANSWER: (D)

2. [3pts] \( \pi_{PID,CID}(P \times C) - \pi_{PID,CID}(R) \)
(A) \( \pi_{PID,CID}((P \times C) - R) \)

(B) \( \pi_{PID,CID}(P - R) \times \pi_{PID,CID}(C - R) \)

(C) \( (\pi_{PID}(P) \times \pi_{CID}(C)) - \pi_{PID,CID}(R) \)

(D) \( \pi_{PID,CID}(P \bowtie C \bowtie R) \)

(E) \( \pi_{PID,CID}(P \times C - P \bowtie C \bowtie R) \)

**ANSWER: (C)**

3. **[3pts]** \( \gamma_{\text{AVG}(Age)}(\pi_{CID,Age}(\sigma_{PID=123}(C \bowtie R))) \)

   (A) \( \gamma_{\text{AVG}(Age)}(\sigma_{PID=123}(R) \bowtie \pi_{CID,Age}(C)) \)

   (B) \( \gamma_{\text{AVG}(Age)}(\pi_{CID,Age}(\sigma_{PID=123}(R) \bowtie \pi_{CID,Age}(C))) \)

   (C) \( \gamma_{\text{AVG}(Age)}(C \bowtie \sigma_{PID=123}(R)) \)

   (D) \( \gamma_{\text{AVG}(Age)}(\sigma_{PID=123}(R) \bowtie \sigma_{Age}(C)) \)

   (E) \( \gamma_{CID,AVG(Age)}(\pi_{CID}(\sigma_{PID=123}(R) \bowtie C)) \)

   **ANSWER: (B)**

4. **[3pts]** \( \sigma_{CID=1, PID=3}(P \bowtie C \bowtie R) \). Indicate which of the following is not logically equivalent to this query (only one option is correct).
Part 2. [12pts] For the following questions, use these pieces of information about an instance of the same shopping database from above, as well as the machine environment.

Products (PID, PName, Vendor, Price)
Customers (CID, CName, Zipcode, Age)
Purchases (CID, PID, Timestamp, Quantity)

\[ N_{Products} = 10^6, \quad N_{Customers} = 10^7, \quad \text{and} \quad N_{Purchases} = 10^9 \] (respective number of pages of each table in a row-store format heap file).
The size of each attribute is 8 B, except PName, Vendor, and CName, which are 40 B each. Page size is 8 KB. Available buffer memory is 32 GB.
The fudge factor for hash tables is 1.4.

Assume there are no indexes and the buffer pool is empty to start with. Exclude the cost of writing the output of a given query in your I/O cost calculations.

For each given query, what is the lowest possible I/O cost of the following join query? Use only the physical operator implementations discussed in class.

5. [3pts] \textit{Products} \bowtie \textit{Purchases}

(A) 1001 \cdot 10^6 \quad (B) 1002 \cdot 10^6 \quad (C) 1004 \cdot 10^6 \quad (D) 2001 \cdot 10^6 \quad (E) 2002 \cdot 10^6

\textbf{ANSWER: (A).} Hash join with Products as inner table.
6. [3pts]  \( \text{Customers} \bowtie \text{Purchases} \)

(A) \( 101 \cdot 10^7 \)      (B) \( 103 \cdot 10^7 \)      (C) \( 303 \cdot 10^7 \)      (D) \( 305 \cdot 10^7 \)      (E) \( 505 \cdot 10^7 \)

ANSWER: (C). Hash join with Customers as inner table; requires one stage of partitioning.

7. [3pts]  \( \pi_{\text{CID}}(\text{Customers} \bowtie \text{Purchases}) \)

(A) \( 101 \cdot 10^7 \)      (B) \( 103 \cdot 10^7 \)      (C) \( 303 \cdot 10^7 \)      (D) \( 305 \cdot 10^7 \)      (E) \( 505 \cdot 10^7 \)

ANSWER: (A). Push the project down through the join to both base tables. Non-deduplicated CID column in Customers is \( 8/64 \) of its size, which is 10GB. So, hash table on it easily fits in memory. Overall, we do a hash join with inner table being the CID column of Customer and the outer table being Purchases; this finishes with one pass over each table.

8. [3pts]  \( \text{Products} \bowtie \text{Customers} \bowtie \text{Purchases} \)

(A) \( 1011 \cdot 10^6 \)      (B) \( 1033 \cdot 10^6 \)      (C) \( 3011 \cdot 10^6 \)      (D) \( 3013 \cdot 10^6 \)      (E) \( 3031 \cdot 10^6 \)

ANSWER: (E). Right-deep tree of hash joins; Products as inner table on top and Customers as inner table at bottom.

Part 3. [5pts] For each statement below, indicate if it is (A) True or (B) False.

9. It is not possible to have an index-only execution plan for a Group By Aggregate.
   False

10. It is possible to use multiple matching indexes when executing a selection query.
    True

11. The data model of SQL++ is a superset of the data model of SQL.
    True

12. In Local-as-View, the mediated schema is a view over all source schemas.
    False
13. Sort-based project is always a blocking operator.  
   **False**

**Part 4. [22pts]** For each question below, select the right option (only one is correct).

14. Which of the following forms of Information Extraction (IE) requires extracting the schema (attribute names) as well?
   (A) Closed-world IE  
   (B) Closed IE  
   (C) Open IE  
   (D) None of these  
   (E) All of these  
   **ANSWER: (C)**

15. Which of the following SQL aggregates is *not* an algebraic aggregate?
   (A) SUM  
   (B) MAX  
   (C) VARIANCE  
   (D) STDEV (Standard Deviation)  
   (E) MODE  
   **ANSWER: (E)**

16. Suppose a given query takes 50 minutes to finish with a single-node RDBMS. After moving the database to a parallel RDBMS with 10 worker nodes, the same query now takes $x$ minutes. For which value of $x$ below is the speedup super-linear?
   (A) 4  
   (B) 5  
   (C) 6  
   (D) None of these  
   (E) All of these  
   **ANSWER: (A)**

17. For which of the following is a new physical operator in a parallel RDBMS?
   (A) Hash Join  
   (B) Hash-based Project  
   (C) Intersection  
   (D) Sort-Merge Join  
   (E) Split  
   **ANSWER: (E)**

18. Which join order is typically amenable to pipelining with INLJ?
   (A) Left-deep tree  
   (B) Right-deep tree  
   (C) Bushy tree  
   (D) None of these  
   (E) All of these  
   **ANSWER: The question was supposed to be on BNLJ but I had INLJ by mistake. It is not possible to say which order is amenable for pipelining for INLJ without**
further information about what indexes are available. So, everyone will get these 2 points.

19. Which of the following paradigms is most commonly used by parallel RDBMSs?
   (A) Shared-disk         (B) Shared-memory       (C) Shared-nothing
   (D) Shared-everything   (E) All of these are equally common.
   ANSWER: (C)

20. Which of the following data integration-related problems does HoloClean address?
   (A) Information Extraction       (B) Schema Alignment
   (C) Entity Matching             (D) Data Cleaning
   (E) All of these
   ANSWER: (D)

21. Which equi-join implementation’s actual I/O cost is heavily dependent on the buffer replacement policy?
   (A) Block-Nested Loop Join      (B) Index Nested Loop Join
   (C) Hash Join                   (D) Sort-Merge Join
   (E) None of these
   ANSWER: (B)

22. Which data partitioning strategy is the most common in parallel RDBMSs?
   (A) Round Robin                (B) Hash-based       (C) Range-based
   (D) Random                     (E) All these are equally common
   ANSWER: (B)

23. Which of the following SQL aggregates is not amenable to incremental maintenance of a materialized view of the aggregate with just logical query rewrites.
   (A) COUNT          (B) SUM           (C) AVG           (D) MIN       (E) MAX
   ANSWER: (C)

24. Which equi-join implementation’s I/O cost may cost go up if the distribution of the join attribute(s) is skewed?
(A) Block-Nested Loop Join  
(B) Index Nested Loop Join  
(C) Hash Join  
(D) Sort-Merge Join  
(E) None of these

ANSWER: (C)