CSE 232A Fall 2019 Final Exam
Answers

Part 1. [60pts] For each question below, select the right option (only one is correct).

1. What is main benefit of doing partial aggregation in the implementation of a group by aggregate in a parallel RDBMS?
   (A) Reduces communication costs    (B) Reduces computation costs
   (C) More accurate results          (D) All of ABC    (E) None of ABC
   ANSWER: (A)

2. Which of these ML systems is designed to work directly on RDBMS-resident data?
   (A) Scikit-learn    (B) MADlib    (C) XGBoost    (D) PyTorch    (E) R
   ANSWER: (B)

3. Which of the following tells us if a page in the buffer pool has been written to by a query?
   (A) PinCount    (B) DirtyBit    (C) Wait queue
   (D) Buffer replacement policy    (E) Number of buffer frames
   ANSWER: (B)

4. Which of these capabilities distinguishes an object-relational DBMS from a purely relational DBMS?
   (A) UDTs/UDFs    (B) PL/SQL    (C) Column stores    (D) Indexing    (E) Buffer pool
   ANSWER: (A)

5. It is not necessary that a star join is always also the following form of join.
   (A) Snowflake Join    (B) Equi Join
   (C) Theta Join        (D) Inner Join    (E) Natural Join
   ANSWER: (E)
6. Which of the following data systems were introduced for write-heavy workloads with a simple put-get API?
   (A) Key-value stores  (B) RDBMSs  (C) Hadoop  (D) Graph DBMSs  (E) Spark
   ANSWER: (A)

7. What was the novel technical capability of Spark relative to parallel RDBMSs when it was introduced?
   (A) Scales to a cluster  (B) Caches data in RAM  (C) Declarative queries/API
   (D) Optimized query execution  (E) Lineage-based fault tolerance
   ANSWER: (E)

8. What is a major reason to prefer SGD over GD for large-scale deep learning?
   (A) Faster per-epoch runtimes  (B) Easier to implement  (C) Can be parallelized
   (D) Scales to larger-than-RAM data  (E) Better accuracy for non-convex models
   ANSWER: (E)

9. Which of the following data systems introduced the first major SQL-on-Hadoop dialect?
   (A) Pig  (B) MongoDB  (C) Redis  (D) Hive  (E) SparkSQL
   ANSWER: (D)

10. Which of the following machine resources will likely face dramatically higher loads in the serverless and resource-disaggregated design of cloud computing relative to previous generations of cloud designs?
    (A) CPU  (B) RAM  (C) Disk  (D) Network  (E) None of these
    ANSWER: (D)

11. Which function in the User-Defined Aggregate API of an RDBMS accumulates the gradient vector in an in-RDBMS implementation of scalable SGD?
    (A) Initialize  (B) Transition  (C) Merge  (D) Finalize  (E) None of these
    ANSWER: (B)
12. Which of the following is usually not a reason why data integration is challenging in practice?

(A) Heterogeneity of data value representations
(B) Ambiguity of value semantics
(C) Manual data entry mistakes
(D) Evolution of source schemas
(E) High compute resource needs

**ANSWER:** (E)

13. Which of these is a major advantage of cloud computing as against operating clusters on-premise?

(A) Better resource manageability
(B) Pay-as-you-go resource pricing
(C) Better resource elasticity
(D) All of ABC
(E) None of ABC

**ANSWER:** (D)

14. Which relational operation requires union compatibility?

(A) \( \cup \)  
(B) \( \cap \)  
(C) \( - \)  
(D) All of ABC  
(E) None of ABC

**ANSWER:** (D)

15. What is the name of the Spark API that supports SQL queries?

(A) RDD  
(B) DataBags  
(C) DataFrames  
(D) PCollection  
(E) None of these

**ANSWER:** The best answer is DataFrames but unfortunately that is not necessarily the only one depending on how one views “Spark API.” For instance, SparkR also allows SQL queries but is at a level higher than DataFrames. Due to this ambiguity, this question is waived and everyone will be awarded these 2pts.

16. Which of these is a major difference in the data access pattern of an SGD epoch as against an SQL \( \text{SUM} \)?

(A) Sequential dependency of mini-batch updates  
(B) SGD is not algebraic  
(C) Random shuffling may be needed  
(D) All of ABC  
(E) None of ABC

**ANSWER:** (D)
17. Which of these is not a step in typical entity matching workflows?
   (A) Pairwise check           (B) Blocking           (C) Clustering
   (D) Wrapper induction       (E) All of these
   ANSWER: (D)

18. Which of these forms of information extraction (IE) require extracting the attributes as well?
   (A) Closed-world   (B) Closed   (C) Open   (D) None of ABC   (E) All of ABC
   ANSWER: (C)

19. Which classical RDBMS idea does Krypton reimagine in a novel way for neural computational graphs in deep learning systems?
   (A) SQL           (B) Indexing           (C) Query compilation
   (D) Sorting       (E) Incremental view maintenance
   ANSWER: (E)

20. What is the key novel form of parallelism exploited by BitWeaving for accelerating SQL queries?
   (A) Multi-core parallelism   (B) Intra-cycle parallelism   (C) Hardware accelerator
   (D) Multi-node parallelism   (E) Symmetric multi-processing
   ANSWER: (B)

21. Which of the following index alternatives causes the size of a data entry to depend on the arity of the relation being indexed?
   (A) AltRecord   (B) AltRID   (C) AltRIDList   (D) None of ABC   (E) All of ABC
   ANSWER: (A)

22. Which relational operator’s implementation in a parallel RDBMS is most similar to that of MapReduce?
   (A) σ           (B) △           (C) γ           (D) ρ           (E) ×
   ANSWER: (C)
23. Which of the following cloud-native RDBMS follows the shared-nothing parallelism paradigm?
   (A) Redshift  (B) Snowflake  (C) Athena  (D) All of ABC  (E) None of ABC
   **ANSWER: (A)**

24. Which of these is a major distinguishing systems capability of ParameterServer compared to model averaging-based distributed SGD?
   (A) Synchronous updates  (B) Asynchronous updates  (C) Works for GLMs
   (D) Supports any mini-batch size  (E) Works on a multi-node cluster
   **ANSWER: (B)**

25. Which of the following storage hardware is byte addressable?
   (A) Magnetic hard disk  (B) Flash / SSD  (C) Tape
   (D) Non-volatile RAM  (E) All of these
   **ANSWER: (D)**

26. Which of these is a major capability of deep learning systems?
   (A) High-level APIs for differentiable programming  (B) Autodiff
   (C) Optimized compilation of neural programs  (D) All of ABC  (E) None of ABC
   **ANSWER: (D)**

27. What was the major motivating workload that led to the creation of MapReduce?
   (A) SQL analytics  (B) Matrix arithmetic  (C) Inverted indexing for Web search
   (D) Machine learning-based analytics  (E) Streaming video transcoding
   **ANSWER: (C)**

28. SQL++ was designed to enable SQL-compatible querying of this semistructured data format.
   (A) XML  (B) HTML  (C) YAML  (D) JSON  (E) ORC
   **ANSWER: (D)**
29. Which family of ML models has recently emerged as the state of the art for many knowledge base construction tasks?

(A) Generalized linear models  (B) Deep learning  (C) Bayesian networks
(D) Decision tree-based models  (E) Support vector machines

ANSWER: (B)

30. Consider the relation schema Customers(CustomerID, Name, Age, Zipcode, SSN) with primary key CustomerID and alternate key SSN. Which of the following indexes is not a unique index?

(A) Clustered B+ tree index on (SSN, Name)
(B) Clustered B+tree index on (Age, CustomerID)
(C) Unclustered B+tree index on (Name, Age, SSN)
(D) Unclustered B+tree index on (Name, Age)
(E) Clustered B+tree index on (CustomerID, SSN)

ANSWER: (D)

Part 2. [20pts] Holiday shopping is here! You are given the following simplified relational database schema for shopping. We use aliases P for Products, G for Category, C for Customers, and R for Purchases. Answer the following questions about queries over this database. Only one option is correct for each question.

P: Products (PID, PName, Vendor, Price)
G: Categories (PID, Category)
C: Customers (CID, CName, Zipcode, Age)
R: Purchases (CID, PID, Date, Quantity)

Products.PID and Customers.CID are primary keys in their respective tables. In Categories, PID is a foreign key referring to Products.PID. In Purchases, CID and PID are foreign keys referring to Customers.CID and Products.PID, respectively. All relations are assumed to be sets of tuples (no bag semantics).

31. [5pts] Which of the following relational algebra queries is a correct translation of the following natural language query?
Get all the distinct products that customer with ID 123 bought on the same date as when this same customer (ID 123) bought eggnog.

\[(A) \pi_{PID}(\sigma_{CID=123}(R) \bowtie \sigma_{\text{Name LIKE 'eggnog'}}(P))\]

\[(B) \pi_{PID}(\sigma_{CID=123}(R) \bowtie \sigma_{\text{Name LIKE 'eggnog'}}(P) \bowtie R)\]

\[(C) \pi_{PID}(\pi_{\text{Date}}(\sigma_{CID=123}(R)) \bowtie \sigma_{\text{Name LIKE 'eggnog'}}(P) \bowtie R)\]

\[(D) \pi_{PID}(\sigma_{CID=123}(R) \bowtie \pi_{\text{Date}}(\sigma_{\text{Name LIKE 'eggnog'}}(P) \bowtie \sigma_{CID=123}(R)))\]

\[(E) \pi_{PID}(\rho_{PID2,\text{Date2}}(\pi_{\text{PID,Date}}(R)) \bowtie \text{Date2=Date} \sigma_{CID=123}(R) \bowtie \sigma_{\text{Name LIKE 'eggnog'}}(P))\]

ANSWER: (D)

32. [5pts] Which of the following relational algebra queries is not a correct translation of the following SQL query?

```sql
SELECT *
FROM Products P, Categories G, Customers C, Purchases R
WHERE C.CID = R.CID AND P.PID = R.PID AND G.PID = P.PID AND Age <= 30 AND Date LIKE '%2019%' AND Category = 'Bakery';
```

\[(A) \sigma_{\text{Age}\leq30}(C) \bowtie \sigma_{\text{Date LIKE '2019'}}(R) \bowtie \sigma_{\text{Category='Bakery'}}(G) \bowtie P\]

\[(B) \sigma_{\text{Date LIKE '2019'}}(\sigma_{\text{Age}\leq30}(C) \bowtie R) \bowtie \sigma_{\text{Category='Bakery'}}(G \times P)\]

\[(C) \sigma_{\text{Age}\leq30 \land \text{Date LIKE '2019'}}(\sigma_{\text{Category='Bakery'}}(G) \bowtie R \bowtie (P \times C))\]

\[(D) \sigma_{\text{Category='Bakery'}}(G \bowtie P) \bowtie \sigma_{\text{Age}\leq30 \land \text{Date LIKE '2019'}}(C \bowtie R)\]

\[(E) \sigma_{\text{Date LIKE '2019'}}(R \bowtie \sigma_{\text{Age}\leq30}(C)) \bowtie \sigma_{\text{Category='Bakery'}}(P \bowtie G)\]

ANSWER: (B)
33. [5pts] Which of the following relational algebra queries is not equivalent to the following relational algebra query? Ignore the output column names.

\[ \gamma_{Date, \text{SUM}(\text{Price*Quantity})} (P \bowtie R) \]

(A) \[ \gamma_{Date, \text{SUM}(\text{Price*Quantity})} (R \bowtie \pi_{\text{PID,Price}}(P)) \]

(B) \[ \gamma_{Date, \text{SUM}(\text{Price*Total})} (\pi_{\text{PID,MAX(Price) AS Price}}(P) \bowtie \gamma_{Date,\text{PID,\text{SUM(Quantity) AS Total}}}(R)) \]

(C) \[ \gamma_{Date, \text{SUM}(\text{Price*Quantity})} (\pi_{\text{PID,Price,Vendor}}(P) \bowtie R) \]

(D) \[ \gamma_{Date, \text{SUM}(\text{Price*Total})} (\pi_{\text{PID,Price}}(P) \bowtie \gamma_{Date,\text{PID,\text{SUM(Quantity) AS Total}}}(R)) \]

(E) \[ \gamma_{Date, \text{SUM}(\text{Price*Total})} (P \bowtie \pi_{\text{PID,Total}(\gamma_{Date,\text{PID,\text{SUM(Quantity) AS Total}}}(R))) \]

**ANSWER: (E)**

34. [5pts] Which of the following relational algebra or SQL queries is a correct translation of the following natural language query?

Get the details of all products in categories that Nestle does not have products in.
(A) \((G \bowtie P) - (G \bowtie \sigma_{\text{Vendor LIKE } '%Nestle'\%(P)})\)

(B) SELECT P2.* FROM Products P1, Products P2, Categories G1, Categories G2
WHERE P1.PID = G1.PID AND P2.PID = G2.PID AND

(C) \(P \bowtie G \bowtie (\pi_{\text{Category}}(G) - \pi_{\text{Category}}(G \bowtie \sigma_{\text{Vendor LIKE } '%Nestle'\%(P)}))\)

(D) SELECT * FROM Products P, Categories G
WHERE P.PID = G.PID AND G.Category IN (SELECT Category FROM Products P2, Categories G2
WHERE P2.PID = G2.PID AND NOT (Vendor LIKE '%Nestle%'));

(E) \(P \bowtie (\pi_{\text{Category}}(G) - \pi_{\text{Category}}(G \bowtie \sigma_{\text{Vendor LIKE } '%Nestle'\%(P)}))\)

**ANSWER:** (C)

**Part 3. [20pts]** You are now given the following pieces of information about an instance of the same shopping database, as well as the machine environment. Recall that we use aliases P for Products, G for Category, C for Customers, and R for Purchases. All relations are assumed to be sets of tuples (no bag semantics).

- **P:** Products (PID, PName, Vendor, Price)
- **G:** Categories (PID, Category)
- **C:** Customers (CID, CName, Zipcode, Age)
- **R:** Purchases (CID, PID, Date, Quantity)

You are given that \(N_P = 10^6, N_G = 2 \cdot 10^6, N_C = 10^7,\) and \(N_R = 10^9\) (respectively number of pages of each table in a row-store packed format heap file).

The size of each attribute is 8 bytes, except for PName, Vendor, Category, and CName, which are 40 bytes each.

Page size is 8 KB. Available buffer memory is 64 GB. The fudge factor for hash tables is 1.4. There are no indexes and the buffer pool is initially empty.

There are no significant skews in the distributions of any foreign keys. You are also given following about the sets of foreign keys in the instance:

\[\pi_{\text{PID}}(P) = \pi_{\text{PID}}(G) = \pi_{\text{PID}}(R)\]  
and  
\[\pi_{\text{CID}}(C) = \pi_{\text{CID}}(R)\]
For each given operation or query, what is the lowest possible rough I/O cost (in pages) using only the physical implementations and techniques discussed in class? You must include (i.e., do not exclude) the I/O cost of writing the final output to a file on disk for each question.

35. [3pts] SELECT * FROM Purchases ORDER BY Date;

(A) $2 \cdot 10^9$  
(B) $4 \cdot 10^9$  
(C) $6 \cdot 10^9$  
(D) $8 \cdot 10^9$  
(E) $10^{10}$

ANSWER: (B). Run EMS. We have $B = 64GB/8KB \approx 8 \cdot 10^6$ and $N = 10^9$. Since we want lowest I/O cost, we must use replacement sort but not double buffering or blocked I/O. So, $F = B - 1$ and $N' = \lceil \frac{N}{2B} \rceil = \lceil \frac{1000}{16} \rceil$. We do not need to calculate this fraction—the log base is so high that the ceiling is just 1 (i.e., EMS needs only 1 merge pass). So, total I/O cost is $2N(1 + 1)$.


(A) $3 \cdot 10^6$  
(B) $4 \cdot 10^6$  
(C) $5 \cdot 10^6$  
(D) $6 \cdot 10^6$  
(E) $7 \cdot 10^6$

ANSWER: (E). Just a join of two tables of sizes 8GB and 16GB. Both fit comfortably in RAM. So, we need only one read of each table and one write of output. The join attribute in this key-foreign key join has the same set of values in both tables. So, the output cardinality is the same as that of G, since PID is not a primary key in G. Output size is roughly double of Categories, since PID and Price are both 8B, while Category and Vendor are both 40B.

37. [4pts] $C \bowtie R$

(A) $101 \cdot 10^7$  
(B) $303 \cdot 10^7$  
(C) $412 \cdot 10^7$  
(D) $578 \cdot 10^7$  
(E) $603 \cdot 10^7$

ANSWER: Unfortunately, the correct answer was not listed among the options! So, this question is waived and everyone will be awarded these 4pts. The hash table on C does not fit in RAM; so hash join construction cost is $3(N_C + N_R)$. But note that the RAM size is such that BNLJ will have a construction cost of only $N_C + \lceil \frac{F_{NC}}{B-1} \rceil N_R \approx N_C + \lceil \frac{14}{8} \rceil N_R = N_C + 2N_R$. The join attribute in this key-foreign key join has the same set of values in both tables. So, the output cardinality is the same as that of $R$. The records expand due to extra attributes from C, adding size $40 + 8 + 8 = 56$ bytes to $R$ records of size 32 bytes. So, output size is $88/32 = 2.75$ times that of $R$. Overall, the answer is $N_C + 4.75N_R = 476 \cdot 10^7$. 

10
38. [4pts] \[
\text{SELECT SUM(P.Price} \ast R\text{.Quantity)} \\
\text{FROM Products P, Customers C, Purchases R} \\
\text{WHERE R.PID = P.PID AND R.CID = C.CID AND C.CID = 123;}
\]

(A) 1011 \cdot 10^6  \hspace{1cm} (B) 1033 \cdot 10^6  \hspace{1cm} (C) 3013 \cdot 10^6  \hspace{1cm} (D) 3031 \cdot 10^6  \hspace{1cm} (E) 3033 \cdot 10^6

**ANSWER: (A).** Right deep tree of hash joins, with P as inner table on top join and the output of the selection on C as the inner table for bottom join. Since CID is the primary key of C, the selection output just 1 tuple. That and the hash table on P both fit comfortably in RAM. So, the whole plan is fully pipelined and needs only 1 read of each table.

Note: There is an out-of-scope rewrite possible here, as pointed out by some students. :) Since R.CID = C.CID = 123 and no other attribute from C is being used, we can simply drop C from the query! This can technically reduce the I/O cost to 1001 \cdot 10^6. But this is only possible due to the out-of-band information given (\(\pi_{C}(C) = \pi_{C}(R)\)). A general RDBMS will not be able to do such a rewrite. This out-of-band information is only needed in the context of these questions to enable the output size to be calculated correctly. And the question specifically requires “using only the physical implementations and techniques discussed in class.” So, overall the correct answer is still 1011 \cdot 10^6.

39. [3pts] For this question alone, you are given that there is a clustered AltRID B+ tree index on R with Index Key (CID, PID). Assume that an RID pointer and the count statistic are each of size 8 bytes.

**\(\gamma_{CID}, count(distinct \ PID) (R)\)**

(A) 502.5 \cdot 10^6  \hspace{1cm} (B) 752.5 \cdot 10^6  \hspace{1cm} (C) 1002.5 \cdot 10^6  \hspace{1cm} (D) 1252.5 \cdot 10^6  \hspace{1cm} (E) 1502.5 \cdot 10^6

**ANSWER: (B).** Index-only aggregate with the aggregation list being a prefix of IndexKey; PID is also sorted for each CID at the leaf level. So, a single sequential scan of leaf level suffices. The rough size of the leaf level is 24/32 times that of R, which is 750 \cdot 10^6. The output has the same cardinality as Customers and is roughly of size 16/64 times that of C, which is 2.5 \cdot 10^6.

40. [3pts] For this question alone, you are given that there is a clustered AltRID B+ tree index on R with Index Key (\(C ID, PID\)) and a clustered AltRID B+ tree index on C with Index Key CID. Assume that an RID pointer is of size 8 bytes. Given this extra information, what is the new lowest I/O cost of \(C \bowtie R\)?
ANSWER: (C). Both tables are already available sorted on the join attribute CID. So, a sort-merge join computes the output in just one read of both tables. Output size is the same as before, 2.75 times $R$. 

(A) $101 \cdot 10^7$  (B) $303 \cdot 10^7$  (C) $376 \cdot 10^7$  (D) $578 \cdot 10^7$  (E) $679 \cdot 10^7$