Database Systems (CPH) – Fall 2024
Lecturer: Tiantian Liu
Exam Date: Jan 7, 2025

AALBORG UNIVERSITY
DENMARK

Preamble

Please read the following instructions carefully:

- This exam has seven questions and 13 pages. Make sure your copy contains all exercises and all questions. The following table summarizes the questions in this exam and their weights.

Question	**Points**
1. Relational Algebra	10
2. SQL	20
3. ER Modeling	16
4. Design Theory	14
5. Physical Design	12
6. Query Processing and Optimization	14
7. Transactions, Concurrency Control, and Recovery	14

- Read the text of each exercise carefully before solving it.
- If you get stuck on an exercise, proceed with another one and come back later.
- If your solution is based on an assumption that you have made, please add a comment for clarification.
- Aids are allowed during the exam, i.e., you can consult aids such as books, notes, etc.
- As a reference, we recommend the course material we endorsed (book, slides, exercises, etc.) as it complies with the notation we introduced in the course, which is the same one used in the exam.
- It is NOT allowed to seek help from or consult others (human beings, artificial intelligence, etc.) during the exam.
- Good luck!

AALBORG UNIVERSITY
DENMARK

1. Relational Algebra (10 points)

Consider the following relations capturing information about deliverymen, restaurants, cuisines, and delivery orders:

restaurant(*rid*, *name*, *street*, *number*, *postalCode*, *city*)
cuisine(*cid*, *name*, *type*, *rid* → *restaurant*)
deliveryman(*pid*, *CPR*, *name*, *age*)
dorder(*pid* → *deliveryman*, *cid* → *cuisine*, *orderTime*, *orderDate*, *tip*)

where the primary keys are underlined and the foreign keys are denoted by arrows.

Fill in the missing information that should go into the boxes to make the query compute the required information. Note that your solution should not output additional information.

To enter the operation symbols into the answer sheet, please use the following replacements:

Operation symbol	What to write in the answer sheet
ρ	rho
γ	gamma
Π	pi
σ	sigma
⋈	join
×	x

1.1 (4 points) Find all the deliverymen who are older than 30. For each such deliveryman, list his/her pid, name, and age.

$$\boxed{\text{Box 1}}_{pid,\ name,\ age}\left(\boxed{\text{Box 2}}_{age\ >\ 30}(deliveryman)\right)$$

$\Pi_{pid,\ name,\ age}(\sigma_{age\ >\ 30}(\text{deliveryman}))$

1.2 (6 points) Find all the deliverymen that have not delivered orders of any Italian cuisine (the cuisine type is "Italian"). For each such deliveryman, list his/her pid.

$$\Pi_{pid}(deliveryman) - \boxed{\text{Box 3}}_{pid}(dorder\boxed{\text{Box 4}}_{dorder.cid\ =\ cuisine.cid}(\boxed{\text{Box 5}}_{type\ =\ \text{"Italian"}}(cuisine)))$$

$\Pi_{pid}(\text{deliveryman}) - \Pi_{pid}(\text{dorder} \bowtie_{dorder.cid\ =\ cuisine.cid}(\sigma_{type\ =\ \text{"Italian"}}(\text{cuisine})))$

2. SQL (20 points)

2.1 Evaluate whether the following statements about SQL are true or false.
2.1.1 (2 points) SQL is case-sensitive for keywords like SELECT, FROM, and WHERE.
(a) True (b) False
2.1.2 (2 points) The UPDATE statement can modify the structure of a table.
(a) True (b) False

2.2 Consider again the relational schema of Exercise 1, which is:

restaurant(*rid*, *name*, *street*, *number*, *postalCode*, *city*)
cuisine(*cid*, *name*, *type*, *rid* → *restaurant*)
deliveryman(*pid*, *CPR*, *name*, *age*)
dorder(*pid* → *deliveryman*, *cid* → *cuisine*, *orderTime*, *orderDate*, *tip*)

where the primary keys are underlined and the foreign keys are denoted by arrows.

Answer the following questions:

2.2.1 (4 points) Which of the following queries can find all restaurants in the city 'Copenhagen' with a postal code starting with '2'? For each such restaurant, list its *rid* and *name*. Choose **only one** option:

(a)
```
SELECT rid, name
FROM restaurant, cuisine
WHERE city = 'Copenhagen' AND postalCode = '2';
```

(b)
```
SELECT rid, name
FROM restaurant
WHERE city = "Copenhagen" AND postalCode LIKE "2_";
```

(c)
```
SELECT rid, name
FROM restaurant
WHERE city = 'Copenhagen' AND postalCode LIKE '2%';
```

(d)
```
SELECT rid, name
FROM restaurant
WHERE city = "Copenhagen" AND postalCode = "2%";
```

(e)
```
SELECT rid, name
FROM restaurant
WHERE city = 'Copenhagen' AND postalCode LIKE '2_';
```

(f) None of the above

2.2.2 (4 points) Which of the following queries can find the delivered cuisine types and the number of orders delivered for that type? Choose **only one** option:

(a)
```
SELECT c.type, COUNT(*)
FROM dorder o JOIN cuisine c ON o.cid = c.cid;
```

(b)
```
SELECT c.name, COUNT(c.cid)
FROM dorder o, cuisine c
GROUP BY c.name;
```

(c)
```
SELECT c.type, COUNT(c.cid)
FROM dorder o, cuisine c;
```

(d)
```
SELECT c.type, COUNT(*)
FROM dorder o JOIN cuisine c ON o.cid = c.cid
GROUP BY c.type;
```

(e)
```
SELECT c.type, COUNT(*)
FROM dorder o JOIN cuisine c ON o.cid = c.cid
GROUP BY c.name;
```

(f) Nove of the above

2.2.3 (4 points) Fill in the missing information that should go into the boxes of the following query: find all restaurants in the city 'Aarhus' that **do not** serve 'Seafood' (the cuisine name is 'Seafood'). For each such restaurant, list its *rid* and *name*.

```
SELECT r.rid, r.name
FROM restaurant r
WHERE [Box 1] AND [Box 2] NOT IN (
   SELECT [Box 3]
   FROM cuisine c
   WHERE [Box 4]
);
```

```
SELECT r.rid, r.name
FROM restaurant r
WHERE r.city = 'Aarhus' AND r.rid NOT IN (
   SELECT c.rid
   FROM cuisine c
   WHERE c.name = 'Seafood'
);
```

2.2.4 (4 points) Assume that the relation *dorder* contains 10 arbitrary rows, and the relation *cuisine* contains 3 arbitrary rows. Consider the following SQL query:

SELECT * FROM *dorder* NATURAL JOIN *cuisine*;

Which of the following statements is **correct**? Choose **one or more** options:

(a) The result of this query contains at most 3 rows.

(b) The result of this query contains at most 10 rows.

(c) The result of this query contains at most 30 rows.

(d) The result of this query contains at most 13 rows.

(e) The result of this query contains 9 attributes.

(f) The result of this query contains 8 attributes.

(g) The result of this query contains 13 attributes.

(h) Nove of the above

3. ER Modeling (16 points)

3.1 Evaluate whether the following statements about ER model concepts are true or false.
3.1.1 (2 points) The ER model is used during the conceptual design phase of database design.
(a) True (b) False
3.1.2 (2 points) A dashed ellipse represents a derived attribute in an ER diagram.
(a) True (b) False
3.1.3 (2 points) All attributes in an ER model must be single-valued.
(a) True (b) False

3.2 (4 points) Consider the relationship *treat* where
each *patient* must be treated by one doctor (total participation)
each *doctor* can treat many patients or no patient.
Which of the following ER diagrams is **correct**? Choose **only one** option:

(a)


(b)

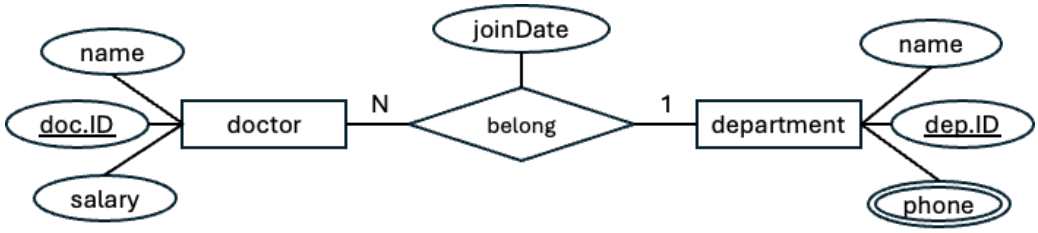

(c)


(d)


(e) None of the above

(b)

3.3 (6 points) Which combinations of relations should be created for the following ER diagram? Note that we aim for an improved representation for the relationship *belong*. Choose **one or more** options:



(a) *doctor*(*doc.ID*, *name*, *salary*)

(b) *doctor*(*doc.ID*, *name*, *salary*, *joinDate*)

(c) *doctor*(*doc.ID*, *name*, *salary*, *joinDate*, *dep.ID* → *department*)

(d) *doctor*(*doc.ID*, *name*, *salary*, *joinDate*, *dep.ID* → *department*)

(e) *belong*(*doc.ID* → *doctor*, *dep.ID* → *department*, *joinDate*)

(f) *belong*(*doc.ID* → *doctor*, *dep.ID* → *department*, *joinDate*)

(g) *department*(*dep.ID*, *name*, *phone*)

(h) *department*(*dep.ID*, *name*, *phone*, *joinDate*)

(i) *department*(*dep.ID*, *name*, *phone*, *joinDate*, *doc.ID* → *doctor*)

(j) *department*(*dep.ID*, *name*, *phone*, *joinDate*, *doc.ID* → *doctor*)

(k) *department*(*dep.ID*, *name*)

(l) *phone*(*dep.ID* → *department*, *phoneNumber*)

(m) *phone*(*dep.ID* → *department*, *phoneNumber*)

AALBORG UNIVERSITY
DENMARK

4. Design Theory (14 points)

4.1 Evaluate whether the following statements about the design theory are true or false.
4.1.1 (2 points) A functional dependency $A \rightarrow B$ implies that $B \rightarrow A$ must also hold.
(a) True (b) False
4.1.2 (2 points) A candidate key must be a superkey.
(a) True (b) False

4.2 Consider a relation schema $R = (A, B, C, D, E, F)$ with the following functional dependencies (FDs):

$$AB \rightarrow C$$
$$B \rightarrow A$$
$$ABC \rightarrow D$$
$$D \rightarrow EF$$

Answer the following questions:

4.2.1 (4 points) Which of the following attributes (sets) can be the **superkey** of R? Choose **one or more** options:

(a) A
(b) B
(c) C
(d) D
(e) E
(f) F
(g) AB
(h) BC
(i) AC
(j) ABC
(k) ABCD
(l) ABCDE
(m) ABCDEF
(n) None of the above

4.2.2 (3 points) Which of the following statements is **correct**? Choose **only one** option:
(a) The relation R is in BCNF and 3NF.
(b) The relation R is in BCNF but not in 3NF.
(c) The relation R is in 3NF but not in BCNF.
(d) The relation R is not in BCNF or 3NF.
(e) None of the above

4.2.3 (3 points) Which of the following decompositions of the relational schema $R = (A, B, C, D, E, F)$ is lossless? Choose **one or more** options:
(a) $R_1 = (A, B, C)$ and $R_2 = (A, D, E, F)$
(b) $R_1 = (A, B, C, D)$ and $R_2 = (A, B, E, F)$
(c) $R_1 = (A, C, D)$ and $R_2 = (B, C, E, F)$
(d) $R_1 = (B, D, E)$ and $R_2 = (A, B, C, F)$
(e) None of the above

5. Physical Design (12 points)

5.1 Evaluate whether the following statements about B+ Tree are true or false.
5.1.1 (2 points) Searching in a B+ Tree always requires scanning the entire tree.
(a) True (b) False
5.1.2 (2 points) Keys in a B+ Tree are stored in sorted order, which makes range queries efficient.
(a) True (b) False

5.2 Consider the following B+ Tree with degree $d = 4$.

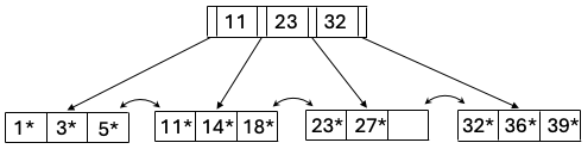


Figure 5: B+ Tree with $d = 4$

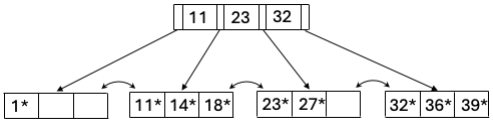
Please make the following assumptions:

- The left pointer of a key k in a non-leaf node leads towards keys less than k, while the right pointer leads towards keys greater than or equal to k.
- A leaf node underflows when the ***number of keys*** goes below $\lceil \frac{d-1}{2} \rceil$.
- An internal node underflows when the ***number of pointers*** goes below $\lceil \frac{d}{2} \rceil$.

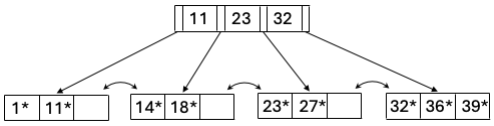
Answer the following questions:

5.2.1 (4 points) Delete 5*, then delete 3*. What is the resulting tree? Choose **only one** option:

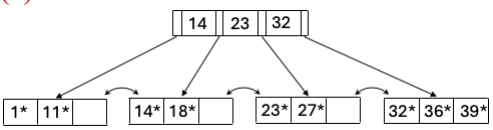
(a)



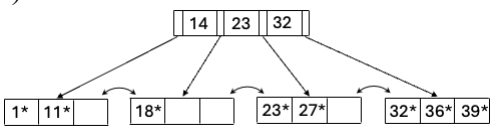
(b)



(c)



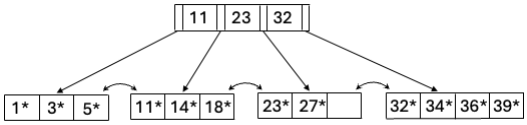
(d)



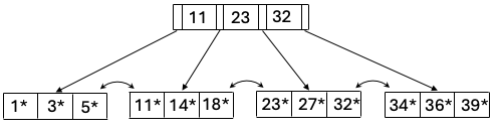
(e) None of the above

5.2.2 (4 points) Consider the **original tree** in Figure 5 (ignore the operations in Question 5.2.1). Insert 34* into the tree. What is the resulting tree? Choose **only one** option:

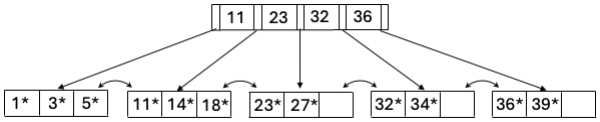
(a)



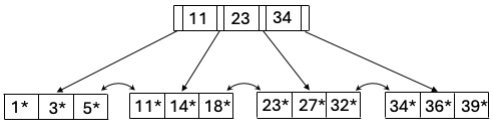
(b)



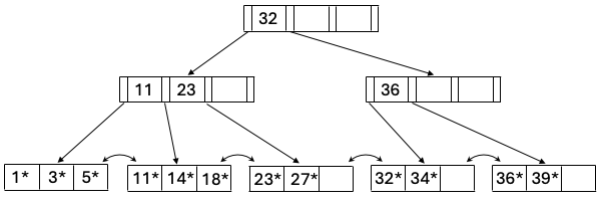
(c)



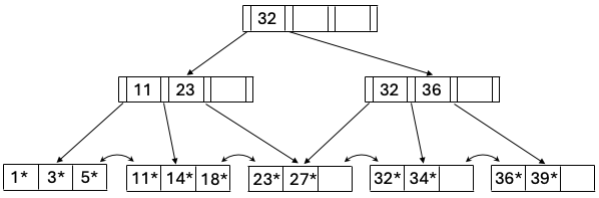
(d)



(e)



(f)



(g) None of the above

AALBORG UNIVERSITY
DENMARK

6. Query Processing and Optimization (14 points)

Consider a database containing relations r_1 and r_2 with the following properties: r_1 has 100,000 tuples, r_2 has 500,000 tuples, 50 tuples of r_1 fit on one page, and 10 tuples of r_2 fit on one page. Assume the computer has 300 buffer pages.

Answer the following questions:

6.1 (2 points) Suppose you need to sort r_1 using the 2-Way External Merge Sort algorithm. How many passes are needed in Phase 2? Choose **only one** option:

(a) $\lceil \log_2 100{,}000 \rceil$

(b) $\lceil \log_2 2{,}000 \rceil$

(c) $\lceil \log_{299} \lceil 2{,}000/300 \rceil \rceil$

(d) $\lceil \log_{299} \lceil 100{,}000/300 \rceil \rceil$

(e) None of the above

6.2 (4 points) Suppose you need to sort r_2 using the Multi-Way External Merge Sort algorithm. Which of the following statements is **correct**? Choose **one or more** options:

(a) The I/O cost in Phase 1 is 100,000.

(b) The I/O cost in Phase 1 is 1,000,000.

(c) The I/O cost in Phase 2 is $\lceil \log_{299} \lceil 50{,}000/300 \rceil \rceil$.

(d) The I/O cost in Phase 2 is 100,000 * $\lceil \log_{299} \lceil 50{,}000/300 \rceil \rceil$.

(e) The I/O cost in Phase 2 is 1,000,000 * $\lceil \log_{299} \lceil 500{,}000/300 \rceil \rceil$.

(f) None of the above

6.3 (4 points) Suppose you need to join r_1 and r_2 using the page-oriented nested-loop join algorithm with r_1 as outer. Which of the following statements is **correct**? Choose **only one** option:

(a) The total I/O cost required is 2,000 + 100,000 * 50,000.

(b) The total I/O cost required is 2,000 + 2,000 * 50,000.

(c) The total I/O cost required is 100,000 + 100,000 * 500,000.

(d) The total I/O cost required is 50,000 + 50,000 * 2,000.

(e) None of the above

6.4 (4 points) Given a relation $r(A, B, C)$ with $n_r = 4000$, $V(A, r) = 100$, $V(B, r) = 200$, and $V(C, r) = 40$, where n_r is the number of tuples in relation r and $V(A, r)$ is the number of distinct values that appear in r for attribute A.
Assume the range of values for attribute A is [0, 100], for attribute B is [20, 60], and for attribute C is [100, 200]. The values are uniformly distributed.

Which of the following statements is **correct**? Choose **one or more** options:

(a) The size estimation of the selection operation $\sigma_{C=150}(r)$ is 100.

(b) The size estimation of the selection operation $\sigma_{A=50}(r)$ is 100.

(c) The size estimation of the selection operation $\sigma_{A<50}(r)$ is 2000.

(d) The size estimation of the selection operation $\sigma_{A<50}(r)$ is 50.

(e) The size estimation of the selection operation $\sigma_{B<30}(r)$ is 1000.

(f) None of the above

AALBORG UNIVERSITY
DENMARK

7. Transactions, Concurrency Control, and Recovery (14 points)

7.1 Evaluate whether the following statements are true or false.
7.1.1 (2 points) During log-based recovery we first do *undo* and then *redo*.
(a) True (b) False
7.1.2 (2 points) In two-phase locking, a transaction cannot obtain a lock if it has released a lock.
(a) True (b) False

7.2 We consider three transactions T1, T2, and T3 consisting of the following actions:

T1: R(X) W(X) R(Z) C
T2: R(Y) R(X) R(Z) W(Z) C
T3: R(X) W(Y) C

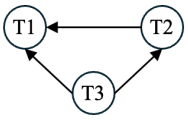
R(X) represents reading object X, W (X) represents writing object X, and C represents Commit. Consider the following two schedules:

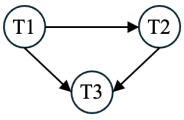
Schedule (1):
T1: R(X) W(X) R(Z) C
T2: R(Y) R(X) R(Z) W(Z) C
T3: R(X) W(Y) C

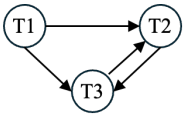
Schedule (2):
T1: R(X) W(X) R(Z) C
T2: R(Y) R(X) R(Z) W(Z) C
T3: R(X) W(Y) C

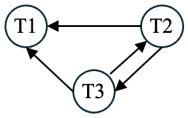
Answer the following questions:

7.2.1 (3 points) Which precedence graph of Schedule (1) is **correct**? Choose **one** option:

(a) 

(b) 

(c) 

(d) 

(e) None of the above

7.2.2 (4 points) Which of the following statements about Schedule (1) is **correct**? Choose **one or more** options:
(a) Schedule (1) is a serial schedule.
(b) Schedule (1) is a conflict serializable schedule.
(c) Schedule (1) is a serializable schedule.
(d) The equivalent serial schedule of Schedule (1) can be T3, T2, T1.
(e) The equivalent serial schedule of Schedule (1) can be T1, T2, T3.
(f) None of the above

7.2.3 (3 points) Which type(s) of schedule does Schedule (2) belong to? Choose **one or more** options:
(a) Serial schedule
(b) Serializable schedule
(c) Conflict serializable schedule
(d) None of the above