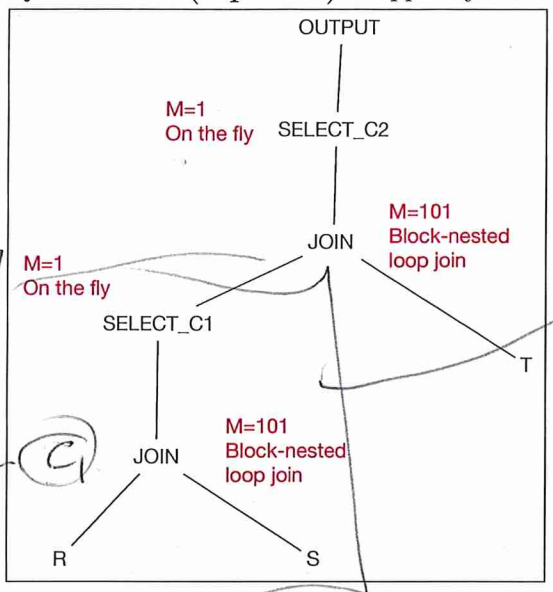
Write your answers in the box below only. Do not write on the back or outside the box.

**Question 6 (8 points).** Suppose you are given the following query plan and statistics:




```
PAGES(R) = 3000
PAGES(S) = 2000
PAGES(T) = 6000
PAGES(R JOIN S) = 4000
PAGES(SELECT_C1(R)) = 800
PAGES (SELECT_C1(R JOIN S)) = 800
PAGES((R JOIN S) JOIN T) = 8000
PAGES(SELECT_C2(SELECT_C1((R JOIN S) JOIN T)) = 80
```

(a) What is the cost of the first join (R with S)? Show your calculation and write a one sentence explanation.

(b) What is the cost of the second join (R and S) with T? Show your calculation and write a one sentence explanation.

(c) What is the total cost of the whole query plan assuming operations are pipelined? Show your calculation and write a one sentence explanation.

RWS =
C1 = $3000 + \left(\frac{3000}{100}\right) * 2,000 = 63,008$

Read T = $\frac{800}{100} * 6,000 = 48,000$

$63,000 + 48,000$

**Write your answers in the box below only. Do not write on the back or outside the box.**

**Question 4 (12 points).** You are given the following queries and statistics.

```
Q1. SELECT C,D FROM R WHERE A=10 AND B<20 ;
Q2. SELECT F FROM S WHERE S.I<= 200 AND S.H= 'ddd' ;
Q3. SELECT G FROM S WHERE S.F=10 AND S.H='ddd';
Index I1 on R(B,A,C) with 3 levels and 1000 leaf nodes
Index I2 on S(H,F,I) with 3 levels and 400 leaf nodes
TUPLES(R)=200,000 PAGES(R)=10,000
TUPLES(S)=100,000 PAGES(S)=2,000
TUPLES(R.A=10)=5,000 TUPLES(R.B<20)=1,000
    TUPLES(R.A=10 AND R.B<20)=80
```

| Attribute | Values | Minval | Maxval |
|---|---|---|---|
| S.F | 4,000 | 1 | 15,000 |
| S.G | 5 | 1 | 10 |
| S.H | 20 | 'aaa' | 'zzz' |
| S.I | 1,000 | 1 | 1,000 |

(a) What is the cost of answering query Q1 using index I1? Show your work.
(b) What is the cost of answering query Q2 using index I2? Show your work.
(c) If you were to create an index on a single attribute to help speed up Q3, what would that index be? Why? If you were allowed to create an index with multiple attributes, which index would you create and why?

(a) R(B,A,C)

- Scan for B condition → $\frac{1000}{200}$ = 5 leaf nodes

- Find tuples for A&B conditions → 80 tuples

- Read those tuples

cost = 2 + 5 + 80 = 87

I1 200,000/1000 = 200 tuples/page

(b) S(H,F,I)

Q2 → scan index for H condition only

2 + 400 * 1/20 = 20+2

Relation cost = 0 (all attributes in Q2 are in I2)

Total = 22 (or 23)

(c) S(H,F,G)

or S(F,H,G) ← multiple

single S(F)

$\frac{100000}{400} = 250$ tuples/page

$\frac{50,000}{20} = 1000$ tuples for H='ddd'

$\frac{5000}{250} = 20$ leaf nodes

A > 5 AND B < 10 AND C = 20

C, A, B

C, B, A

Write your answers in the box below only. Do not write on the back or outside the box.

# Database Systems — CSci 4380
# Final Exam
# December 14, 2023

**RCS ID:** _____ @rpi.edu **Name:** _____

**RIN # :** _____

**Rules.** The exam is 180 minutes long for a total of 100 points. Open book and notes. During the exam, you can refer back to the text books you already have with you or on your computer, any notes I have made available, and any course notes you took. You may not use a search engine to search outside of your personal notes and books or use any other online tool. You may not open the camera of any device. If you are observed doing so, you will automatically fail the exam.

Until the exam is over for everyone, you may not discuss the exam with anyone else in any form, shape, and on any platform. You may not receive help from anyone during the exam, and you may not provide answers or questions to anyone until the exam is over.

Read questions carefully and make any reasonable assumptions.

**Question 1 (12 points).** You are given the following schedule S:

```
S: r1(x) r1(m) w1(m) r2(y) w1(z) r2(m) r3(y) w3(y) w2(y) w3(x) w2(m)
```

(a) List all conflicts in S.

(b) Draw the conflict graph for S. Is S a serializable schedule? Explain why or why not. If the schedule is serializable, find an equivalent serial schedule.

$r_1(x)$ $w_3(x)$

$r_1(m)$ $w_2(m)$

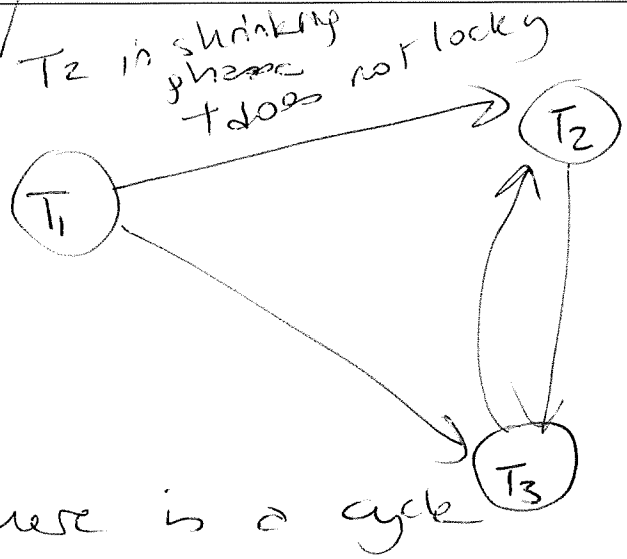
$w_1(m)$ $r_2(m)$

$w_1(m)$ $w_2(m)$

$r_2(y)$ $w_3(y)$

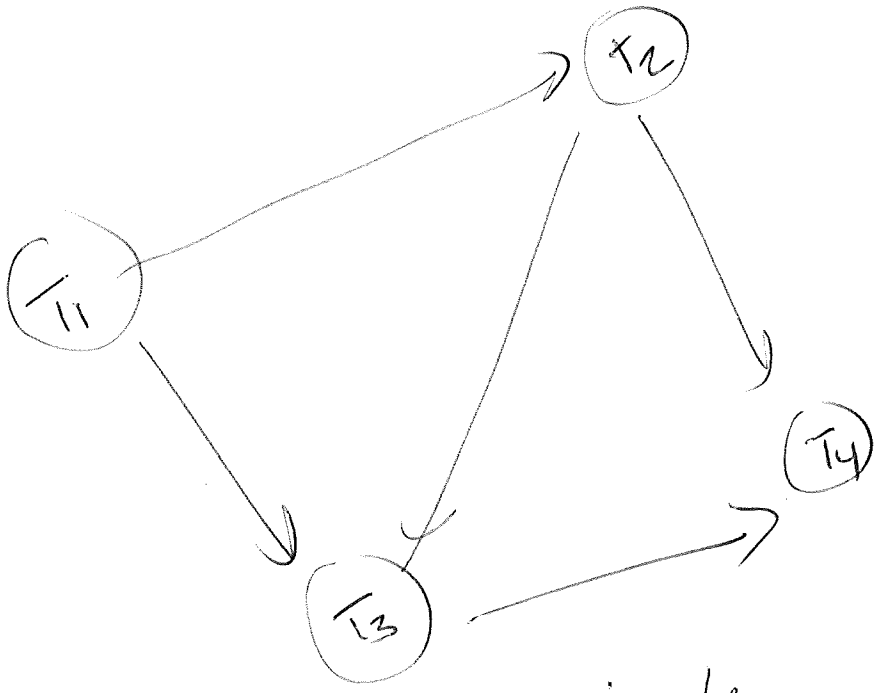
$r_3(y)$ $w_2(y)$

$w_3(y)$ $w_2(y)$



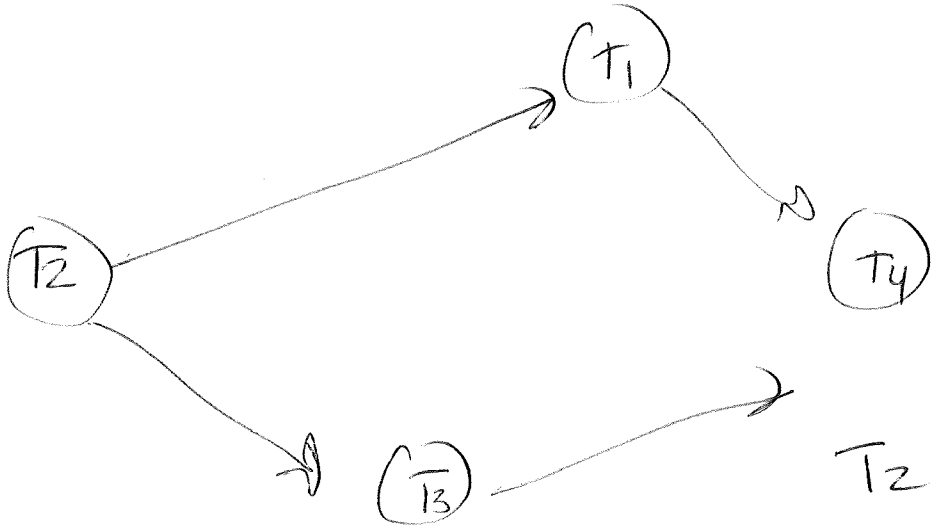

There is a cycle

Not serializable




No cycle, serializable

$T_1, T_2, T_3, T_4$




$T_2\ T_1\ T_3\ T_4$

$T_2\ T_3\ T_1\ T_4$

$T_2$ $T_1$ $T_4$
$T_3$

## Question 1 continued.

(c) Which of the following schedules and sequence of locks are possible under two phase locking? Note: `sl`, `xl`, `ul` stand for shared lock, xlock and unlock of a data item. Assume that `xl` works like an upgrade lock as needed. Write yes/no and provide a short explanation for each.

S1: sl1(x) r1(x) xl1(m) r1(m) w1(m) ul1(m) sl2(y) r2(y) xl1(z) w1(z) sl2(m) ul2(y) r2(m)
sl3(y) r3(y) xl3(y) w3(y) xl2(y) w2(y) xl3(y) w3(x) xl2(m) w2(m)

S2: sl1(x) r1(x) xl1(m) r1(m) w1(m) sl2(y) r2(y) xl1(z) w1(z) sl2(m) r2(m) sl3(y) r3(y)
xl3(y) w3(y) xl2(y) w2(y) xl3(y) w3(x) xl2(m) ul1(m) w2(m)

S3: sl1(x) r1(x) xl1(m) r1(m) w1(m) sl2(y) r2(y) xl1(z) w1(z) sl2(m) ul2(y) r2(m) sl3(y)
r3(y) xl3(y) w3(y) xl2(y) w2(y) xl3(y) w3(x) xl2(m) ul1(m) w2(m)

$sl_1(x) \leftarrow \sigma_1(x)$ $\quad xl_1(x)$ $\quad xl_1(x) \leftarrow \varphi_1(x)$

$T_2$ in shrinking phase

Basis form

~~AB → A~~

AB → C

CD → ~~E~~

CD → F

~~AB~~ → F

AB → F

~~AB~~ → G

~~ABD → F~~

AB → AC

CD → EF

ABC → FG

~~G → H~~

ABD → F

$ABD^+ = \{A, B, D, C, F, E, G\}$

$ABC^+ = \{A, B, C, G\}$

$CD^+ = \{C, D, E\}$

$AB^+ = \{A, B, C, F, G\}$

AB → CFG

CD → EF

AB → C

CD → E

CD → F

AB → F

AB → G

ABC

CDE

CDF

ABF

ABG

ABD

ABCFG

CDEF

ABD

Write your answers in the box below only. Do not write on the back or outside the box.

**Question 3 (12 points).** You are given the following relation and functional dependencies to store information about the staff working in the Alumni office:

`Staff(sid, name, phone, email, title, startdate, region, url, eventtype, bio)`

`sid` → `name title bio startdate`
`region` → `url email`

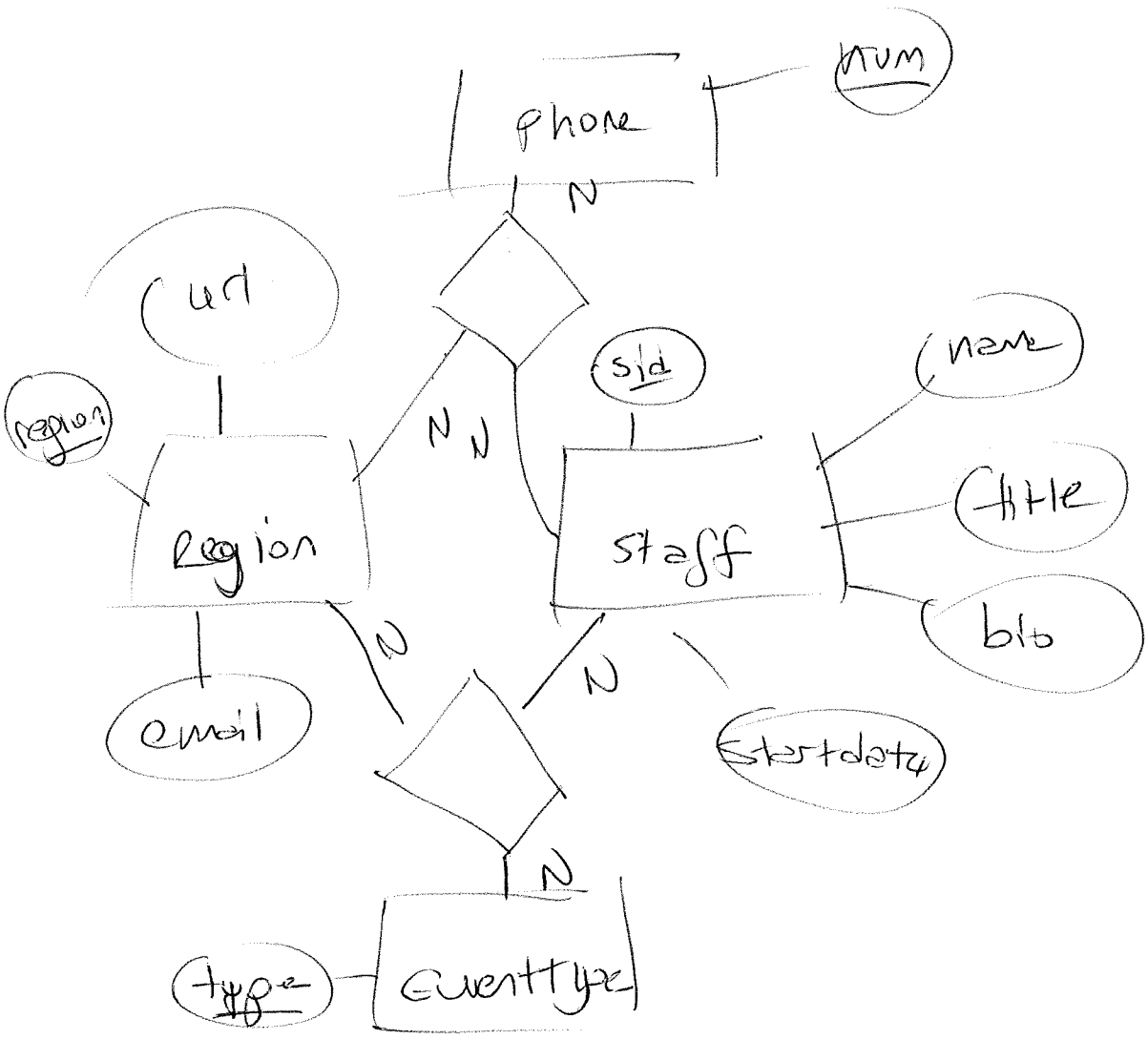
(a) According to this model, is it possible to store for a staff member with a given `sid`:
(i) two different `names`? YES/NO
(ii) two different `titles`? YES/NO
(iii) two different `regions`? YES/NO
(iv) two different `eventtypes`? YES/NO
Circle yes/no above and explain why with a short sentence.

(b) List all keys for this relation. Is the relation in 3NF? Explain why or why not.
(c) If the relation is not in 3NF, decompose it using 3NF decomposition. For each resulting relation, list the functional dependencies and indicate briefly whether it is also in 4NF or not.

key:
eventtype, phone, sid, region

(b) Not in 3NF
region → url email
not superkey
not prime

(c) (sid, name, title, bio, startdata)
(region, url, email)
(eventtype, phone, sid, region)


phone
num
N
url
sid
name
region
N
N
Region
staff
title
blb
email
N
N
startdate
N
type
eventtype

3NF decomp:

A, B, C

C, D, E, F

A, F, D

A, G, D

{A → BC}

{CD → EF}

{FD → A}

{}

keys

$R(A, B, C, D, E, F, G)$

CDG

AGD

GDF

$A \rightarrow BC$ ← not in 3NF

$CD \rightarrow EF$

3NF $\vee$ $FD \rightarrow A$ ← prime

ABCGDE

Prime attrs = A, G, D, F, C

BCNF ⇒ Trivial or All left is a superkey

Not in BCNF, $A \rightarrow BC$

A is not a superkey

3NF → Trivial or Left is a superkey

or All right side are prime

Not → $CD \rightarrow EF$

CD is not a superkey

E is not prime

$R(A, B, C, D, E)$

$A \rightarrow BC$

~~$C \rightarrow D$~~

$D \rightarrow E$

Key: A

NOT IN BCNF

$A \rightarrow BC$ ok for BCNF A is a superkey

$C \rightarrow D$ not ok → C is not a superkey

---

BCNF decomp.

$C \rightarrow D$ violates BCNF

$C^+ = \{C, D, E\}$

All attrs in R but $C^+ - \{C\}$ remove

$X(C, D, E)$ logic

$C \rightarrow D$

$D \rightarrow E$

$(A, B, C)$ ✓

$A \rightarrow BC$

Key: A

$(D, E)$ $D \rightarrow E$

$(C, D)$ $C \rightarrow D$

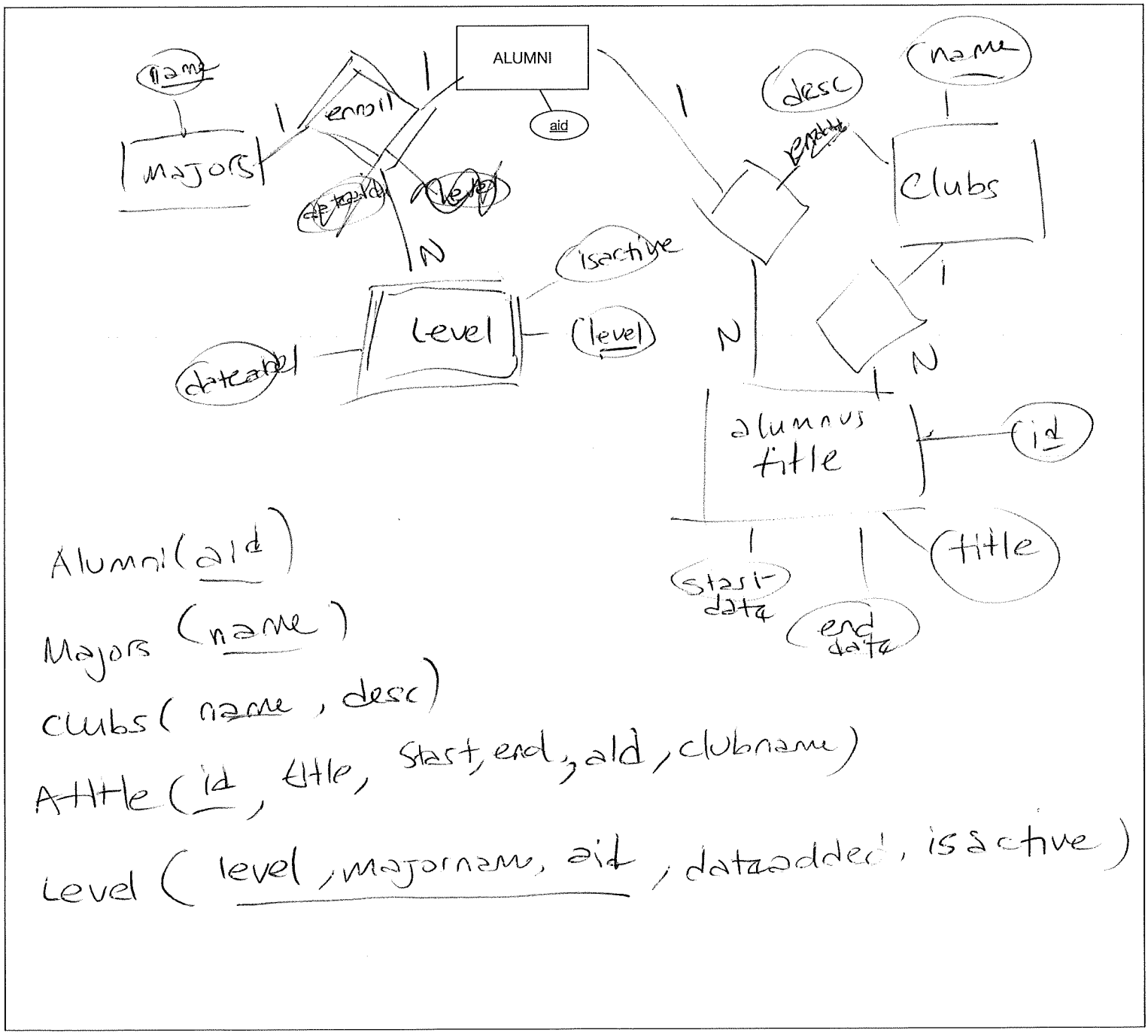
$(A, B, C)$ $A \rightarrow BC$

Write your answers in the box below only. Do not write on the back or outside the box.

**Question 4 (10 points).** Add the following information to the ER Diagram below and then convert your ER Diagram to the relational model. Underline the key of each relation. Make sure all your resulting relations are in BCNF.

Majors have unique names. Alumni enroll in a major (e.g. CSCI) at a specific level (e.g. MS) on a given date (dateadded). They can only enroll in a specific major and level once. For each major, level and alumnus, we store whether the alumnus has graduated and whether his/her enrollment is active.

The database has clubs, identified by a unique name. Each club has a description. Alumni hold titles in a club. For each title, there is a start and an end date. An alumnus can hold multiple titles at the same time period and can hold the same title in different time periods.




Alumni(aid)

Majors (name)

clubs( name , desc)

Attle ( id , title, start, end, aid, clubname)

Level ( level, majorname, aid , dataadded, isactive )