COMSW4111_003_2024_03:
Final Exam

Version 1.2, 2024-12-10 1350

Instructions

- We will be managing time **very strictly.**
    - Do NOT open this cover page until the exam moderator/proctor states that you can begin.
    - You will be given a 10-minute warning and a 5-minute warning before the exam time is complete.
    - You must stop writing immediately when the proctor announces that the exam is over. Failing to stop will result in substantial point deductions on the exam.
- You cannot use, look at, touch, … any electronic devices. All bags, books, coats, etc. should be in one of the collection areas at the front of the classroom.
- Your desk/writing area must be clear of everything other than this exam document and your two-page, double-sided "cheat sheet."
- Write your answers in the spaces provided after the questions. If you need additional space, you may use the backside of a page. **Please indicate that your answer continues on the backside of the page and on which page.**

"A man who uses a great many words to express his meaning is like a bad marksman who instead of aiming a single stone at an object takes up a handful and throws at it in hopes he may hit." — Samuel Johnson

We will deduct points if your answers are not concise and to the point.

NAME:

UNI:

A. Module I – Foundational Concepts

Written Questions, 2 points each

This section requires short, written answers. No question requires more than 5 sentences or bullet points.

1.1 ER Modeling

1. Describe the *Conceptual, Logical,* and *Physical Layers* of Data Modeling?
   What type of user is the audience for each layer?
   What does the modeler specify in each layer?

2. Explain *degree* of a relationship and *cardinality* of a relationship.

3. Describe/explain *alternate key* and *surrogate key.*

1.2 Relational Model

4. What is the difference between the schema *instructor (ID, name, dept_name, salary)* the schema *instructor (ID, name, dept_name, salary)?*

5. Explain Codd's Rule 4: *Active Online Catalog*
   Give an example of how SQL implements this rule.

1.3 SQL

6. SQL is more powerful/expressive than relational algebra. Give an example of a SQL statement that cannot be implemented with relational algebra operations.

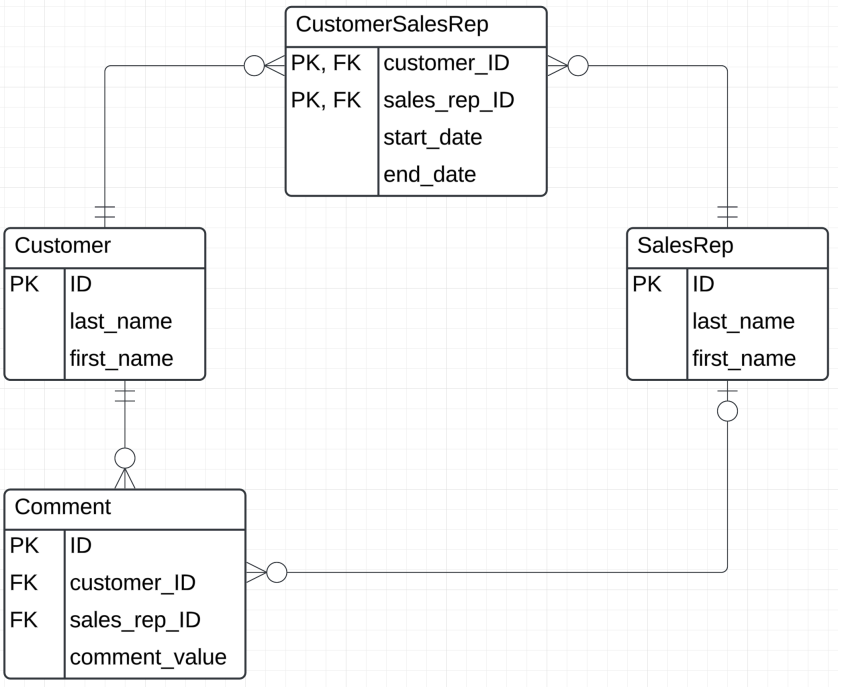
7. Give three examples of integrity constraints that apply to a single table.

8. Explain *cascading actions* in *referential integrity constraints.*

2 Practical Questions

2.1 ER Modeling, **8 points**

9. Write SQL DDL statements that implement the following *Crow's Foot* diagram. You can assume that all data types are *text.* We are focusing on your understanding of concepts. We are not focusing on memorization of SQL and perfectly following the syntax. Place your DDL on the next page.



Your SQL DDL here:

2.2 Relational Model/Algebra, **4 points**

For the following question, use the schema and relations below.

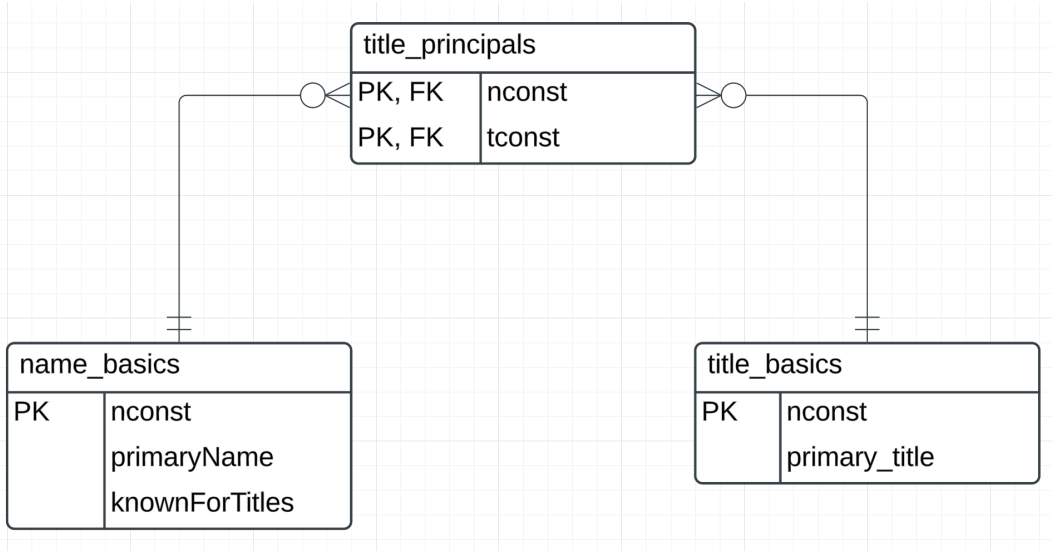
R.a	**R.b**	**R.c**		**S.b**	**S.d**		**T.b**	**T.d**
1	'a'	'd'		'a'	100		'a'	100
3	'c'	'c'		'b'	300		'd'	200
4	'd'	'f'		'c'	400		'f'	400
5	'd'	'b'		'd'	200		'g'	120
6	'e'	'f'		'e'	150			

10. What is the result of the following relational algebra statement?

```
π alpha←a, beta←b, charlie←c, delta←d
(σ d > 100
        (
                (S ⋈ R)
                ⋉
        T)
)
```

2.3 SQL, **16 points each**

11. Consider the following subset of the IMDB schema shown in the ER diagram and DDL.




```
create table if not exists w4111_f24_final.name_basics
(
    nconst         text         null,
    primaryName    text         null,
    knownForTitles text         null
);

create table if not exists w4111_f24_final.title_basics
(
    tconst        text null,
    primary_title text null
);


create table if not exists w4111_f24_final.title_principals
(
    nconst text null,
    tconst text null
);
```

You can assume that the following tables contain representative data. That is, the values indicate the size, type, and content of the columns. The fields are the following:

- nconst is a string that is a primary key identifying a row in name_basics.
- primaryName is a string of the form “first_name last_name.” You can assume that the strings always contain a first name, ‘ ‘, and last name.
- tconst is a string that is a primary key identifying a row in title_basics.
- primary_title is string representing the primary title of the film, episode, etc.
- The table title_principals “connects” name_basics and title_basics entires.
- knownForTitles is a comma delimited string containing 0, 1, 2, 3 or 4 tconst values for the titles for which the person is best known.

nconst	primaryName	knownForTitles
nm0000001	Fred Astaire	tt0072308,tt0050419,tt0053137,tt0027125
nm0000002	Lauren Bacall	tt0037382,tt0075213,tt0117057,tt0038355
nm0000003	Brigitte Bardot	tt0057345,tt0049189,tt0056404,tt0054452
nm0000004	John Belushi	tt0072562,tt0077975,tt0080455,tt0078723
nm0000005	Ingmar Bergman	tt0050986,tt0083922,tt0050976,tt0069467

tconst	primary_title
tt0000001	Carmencita
tt0000002	Le clown et ses chiens
tt0000003	Poor Pierrot
tt0000004	Un bon bock
tt0000005	Blacksmith Scene

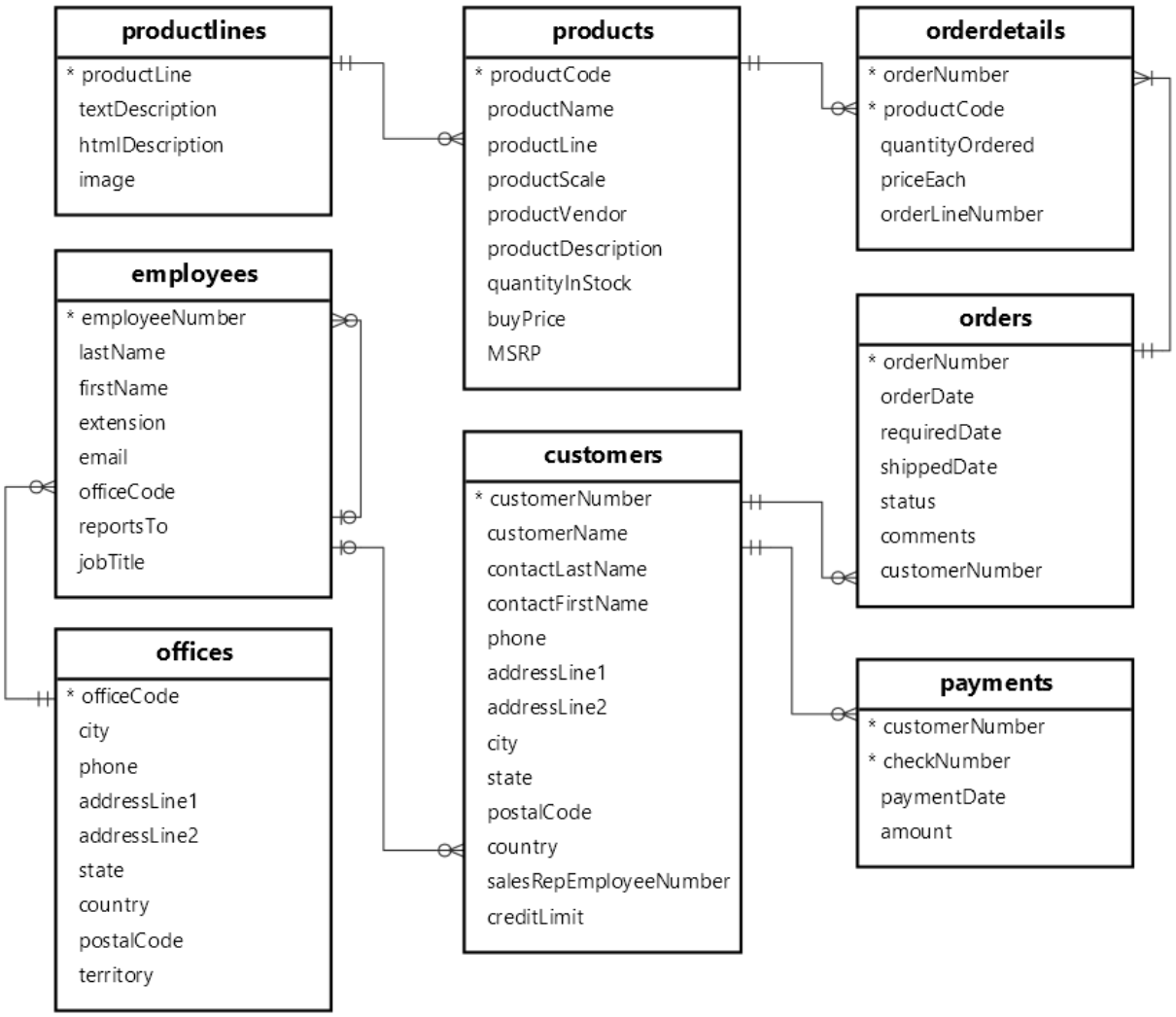
nconst	tconst
nm1588970	tt0000001
nm0005690	tt0000001
nm0005690	tt0000001
nm0374658	tt0000001
nm0721526	tt0000002

a) Please list below what changes you would make to the schema to make it better and why? **(4 points)**

b) Write the new DDL statements for the schema based on your changes. **(12 points)**

12. Below is an ER diagram for the schema of classic models. Write a SQL query that uses customers, orders and orderDetails to produce the result below. The requirements on the result are:

- The result is sorted by country.
- total_revenue is the sum of quantityOrdered*priceEach for orderDetails in orders placed by customers in the country.
- The result only contains an entry for countries with total_revenue >= 250000.
- If total_revenue >= 1,000,000, then revenue_tier is “Tier 1.” If total_revenue < 1,000,000 and >= 400,000, the revenue_tier is “Tier 2.” Otherwise, the revenue_tier is “Tier 3.”
- The result only contains countries with total revenue >= 250000.



	country	total_revenue	revenue_tier
1	Australia	562582.59	Tier 2
2	Finland	295149.35	Tier 3
3	France	1007374.02	Tier 1
4	Italy	360616.81	Tier 3
5	New Zealand	476847.01	Tier 2
6	Singapore	263997.78	Tier 3
7	Spain	1099389.09	Tier 1
8	UK	436947.44	Tier 2
9	USA	3273280.05	Tier 1

a) Write your SQL in the space below. We are not focused on ensuring that you get the syntax exactly correct. **(12 points)**

b) Explain the choices, approaches, etc. that you used in your query and why. More specifically, explain each clause, use of language keywords, subqueries, … … For clarity, keywords are things like SELECT, FROM, WHERE, ORDER BY, … … A clause is a larger unit, for example SELECT x, y, average(x, y) … …, or A JOIN B JOIN C. **(4 points)**

Written Questions, 2 points each

This section requires short, written answers. No question requires more than 5 sentences or bullet points.

B. Module II – DBMS Architecture and Implementation

13. Briefly explain logical block addressing and cylinder-head-sector block addressing.

14. What are four primary approaches to organizing records in files?

15. For which types of SQL queries is Columnar Representation/Storage of records useful.

16. What is a query processing scenario for which least recently used is an extremely poor performing buffer replacement strategy?

17. What is a clustering index. How many clustering indexes can a table have?

18. What is a major advantage of hash indexes and a major disadvantage of hash indexes? How does a B+ Tree resolve the disadvantages?

19. Give a simple example of an SQL statement for which the query optimizer might choose to implement a JOIN using a hash join.

20. For the sample database associated with the recommended textbook, what is an equivalent SQL statement to the following that the query optimizer might choose for larger datasets?

```
select
    *
from
   instructor join teaches
where
   instructor.dept_name="Comp. Sci." and teaches.year = 2018;
```

21. What are two primary “Evils of Redundancy” relative to a schema design?

22. Briefly explain the concepts of functional dependency and decomposition to eliminate functional dependencies.

23. Relative to transactions/recovery, briefly explain the concepts of steal/no steal and force/no force. What technique is used to enable the use of steal/no force without compromising ACID properties?

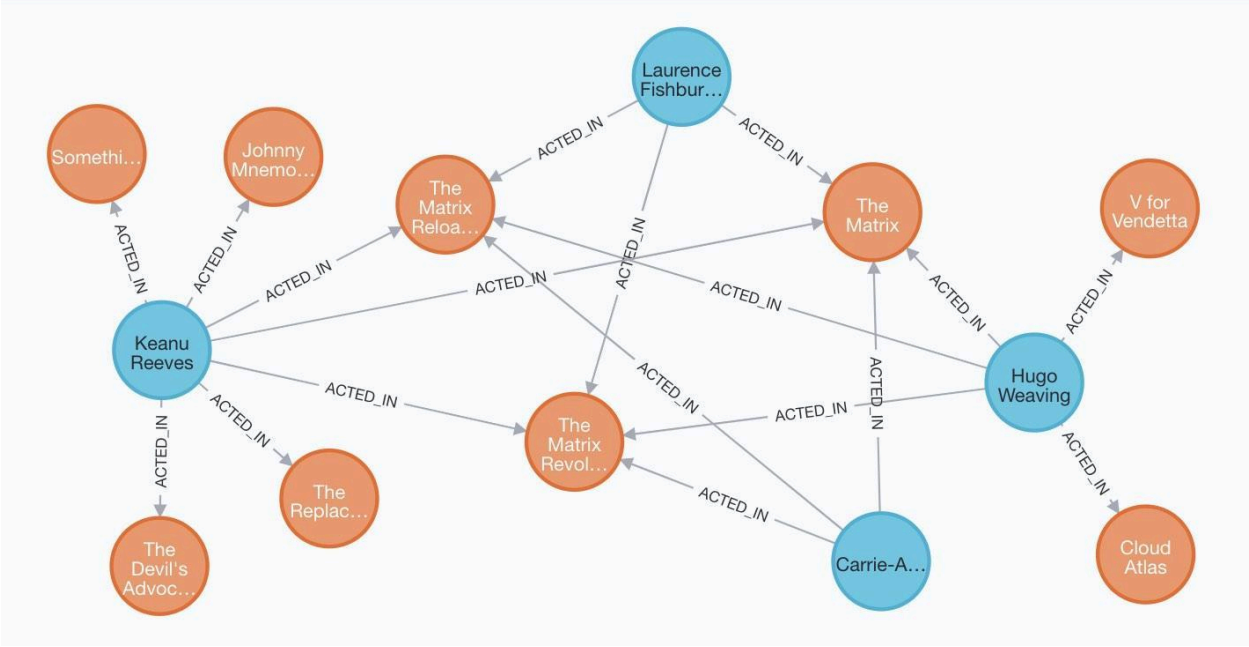
24. Explain the concepts of 2-Phase Locking and Strict 2-Phase Locking.
What is the primary benefit of Strict 2-Phase Locking compared to non-strict?

25. Relative to transactions, what is a deadlock? What are two techniques used to resolve a deadlock?

C. Module III – NoSQL

26. How would you represent the IMDB data from the SQL-P1 question in MongoDB and Neo4j. Give a couple of simple examples of what the data might look like. The examples should be sample documents for MongoDB and node/edge drawings for Neo4j. *Your documents for MongoDB and diagrams for Neo4j should show field names, labels, etc.*

27. For the sample movie data that you used in the Neo4j tutorial, write a Cypher/Neo4j query that finds all people that were associated with/related to a movie with which Tom Hanks was associated. For reference, this is an example of the data in the database.



28. The following are 3 sample documents from a MongoDB collection that contains information about orders from classic models. On the next page, write a MongoDB aggregation pipeline that would produce a result in the format in the example that follows the representative documents. We are not very concerned about syntax and want you to demonstrate your understanding of the concepts.

Sample Documents

Sample Answer

```
orderNumber : 10100
orderDate : "2003-01-06"
status : "Shipped"
productCode : "S18_4409"
quantityOrdered : 22
priceEach : 75.46
```

```
productCode : "S18_2795"
quantityOrdered : 26
priceEach : 167.06
orderNumber : 10101
orderDate : "2003-01-09"
status : "Shipped"
```

```
orderNumber : 10101
orderDate : "2003-01-09"
status : "Shipped"
productCode : "S24_2022"
quantityOrdered : 46
priceEach : 44.35
```

```
orderNumber : 10101
orderDate : "2003-01-09"
status : "Shipped"
productCode : "S24_1937"
quantityOrdered : 45
priceEach : 32.53
```

```
status : "Shipped"
productCode : "S18_2325"
quantityOrdered : 25
priceEach : 108.06
orderNumber : 10101
orderDate : "2003-01-09"
```

```
orderNumber : 10102
orderDate : "2003-01-10"
status : "Shipped"
productCode : "S18_1367"
quantityOrdered : 41
priceEach : 43.13
```

Write your MongoDB aggregation here:

D. Module IV – Data Engineering and Insight

29. Explain the concept of Algebraic Operation in Spark and how it compares to the operations in Map-Reduce.

30. Using the Classic Models ER-diagram from above. Explain and give examples of the OLAP/Star Schema concepts of:
    - Fact Table
    - Dimension Table
    - Drill Down and Roll Up

31. Using the Classic Models ER-diagram from above. Draw the ER-Diagram of a star schema for classic models that shows your fact table and dimension tables.

32. What are the Five Vs of big data?