Exam

Module: Business Intelligence and Information Technology
Course: Databases and Business Intelligence
Date: 24 March 2017
Normal Exam time: 13:45 – 15:30

Lecturers: dr. C. Amrit
University of Twente

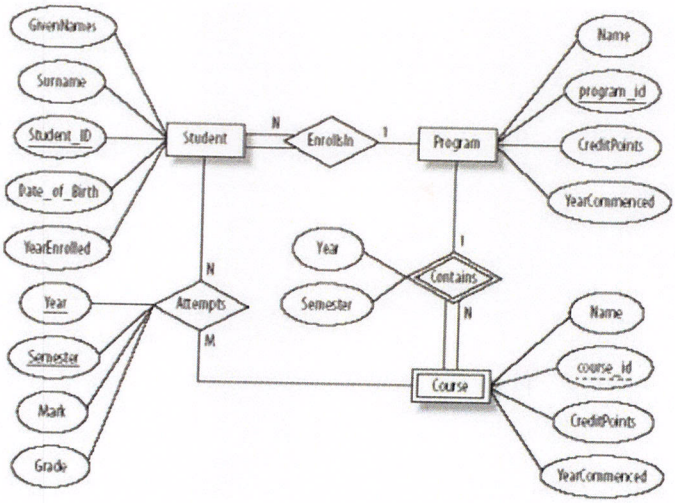
Name:_____

Student nr:_____

Please **return** the question paper after the exam!

Closed Book Exam
This is a closed book exam: No course materials (slides handouts, books, and papers) can be used during the exam.

Grading
This exam contains 30 multiple-choice questions (from the lectures, book chapters and from the papers that have been included in the course) and 1 optional question (for 2 marks). All the questions have only **1 right answer.**
Each of the following questions is for 1 mark.



1. If the above ER Diagram (where the cardinalities are mentioned in the relationships, and triangle also represents a relationship) were to be converted to a relational schema, the schema would contain

A) 4 tables
B) 5 tables
C) 6 tables
D) 7 tables

E) 8 tables

Answer: A.

2. In the same ER Diagram above the number of tables in the schema formed from entities and relations other than *strong entities* are:

A) 1 table
B) 2 tables
C) 3 tables
D) 4 tables
E) 5 tables

Answer: B.

3. In the query

SELECT A FROM B WHERE C ORDER BY D

A) A selects the rows and C the columns from Table D
B) A selects the columns and C the rows from Table B
C) A selects the columns and B the rows from Table C
D) A selects the rows and C the columns from Table B
E) A selects the columns and D the rows from Table B

Answer B.

Consider the following schema:

Salesperson

ID	Name	Age	Salary
1	Abe	61	140000
2	Bob	34	44000
5	Chris	34	40000
7	Dan	41	52000
8	Ken	57	115000
11	Joe	38	38000

Customer

ID	Name	City	Industry Type
4	Samsonic	pleasant	J
6	Panasung	oaktown	J
7	Samony	jackson	B
9	Orange	Jackson	B

Orders

Number	order_date	cust_id	salesperson_id	Amount
10	8/2/96	4	2	540
20	1/30/99	4	8	1800
30	7/14/95	9	1	460
40	1/29/98	7	2	2400
50	2/3/98	6	7	600
60	3/2/98	6	7	720
70	5/6/98	9	7	150

It contains information on sales and orders of a company. Where ID is the primary key of both Salesperson and Customer tables, while Number is the primary key for the table Orders.

4. Based on the above, which of the following potential functional dependencies, is NOT a functional dependency.

A. Customer.City -> Customer.Industry Type
B. Customer.Industry Type -> Customer.Name
C. Salesperson.Name->Salesperson.Age
D. Orders.Number -> Orders.cust_id
E. Orders.Number -> Orders.salesperson_id

Answer B

5. The query:

```
SELECT COUNT(Ord.Number)
FROM Orders Ord
WHERE Ord.Amount <1000
GROUP BY Ord.salesperson_id
```

Order By Ord.salesperson_id Asc;

will return:

A. 3
B. 3, 2, 1
C. 1, 2, 3
D. 3, 1, 1
E. 1, 1, 3

Answer E.

6. Which of the following return less than 3 tuples

A. SELECT S.Name FROM Salesperson S WHERE S.Salary > 40000;
B. SELECT COUNT(distinct(C.Name)) FROM customer C GROUP BY C.city;
C. SELECT Ord.Number FROM Orders Ord WHERE Ord.Amount>550;
D. SELECT C.Name FROM Customer C, Orders Ord WHERE (C.ID=Ord.cust_id) AND (Ord.Amount>600);
E. SELECT Ord.Number FROM Salesperson S, Orders Ord WHERE (S.ID=Ord.salesperson_id) AND (S.Salary>100000);

Answer E.

7. Which of the following queries gives the names of all salespeople that have an order with 'Samsonic'

A) SELECT DISTINCT S.Name FROM Salesperson S, Customer C, Orders Ord WHERE S.ID=Ord.salesperson_id AND C.ID=Ord.cust_id;
B) SELECT DISTINCT S.Name FROM Salesperson S, Customer C, Orders Ord WHERE S.ID=Ord.salesperson_id AND C.Name = 'Samsonic';
C) SELECT S.Name FROM Salesperson S, Orders Ord WHERE S.ID = Ord.salesperson_id and Ord.cust_id = '4';
D) SELECT S.Name FROM Salesperson S where S.ID= ANY (select DISTINCT (Ord.salesperson_id) FROM Orders Ord, Customer c WHERE C.Name = 'Samsonic');
E) SELECT DISTINCT S.Name FROM Salesperson S, Customer C, Orders Ord WHERE C.ID=Ord.cust_id AND C.Name = 'Samsonic';

Answer C

8. The query:

```
SELECT DISTINCT S.Name
FROM Orders Ord, Salesperson S
WHERE Ord.salesperson_id = S.ID
GROUP BY S.Name, Ord.salesperson_id
HAVING COUNT( Ord.cust_id ) >1
```

Returns:

A) The names of all salespersons
B) The names of salespersons with more than one order.
C) The names of only those salespersons with more than one customer
D) The names of all distinct salespersons with two or more recent orders
E) The names of all distinct salespersons who serve customers with more than one order

Answer C

9. The above schema is in/only in the

A. First Normal Form
B. Second Normal Form
C. Third Normal Form
D. None of the above

Answer C.

Consider the following Employee Department Project DB schema (the primary keys are underlined):

EmpID	Department	ProjID	ProjName	ProjDescription	ProjYear
123	IEBIS	P1	PName1	Social Media	2012
123	IEBIS	P2	PName2	Social Network	2013
324	BA	P1	PName1	Database	2014

EmpID	Name	Address
123	ABC	Abcdstraat, Enschede
324	BCD	Asasdastraat, Enschede

10. The above schema is in/only in the

A. First Normal Form
B. Second Normal Form
C. Third Normal Form
D. None of the above

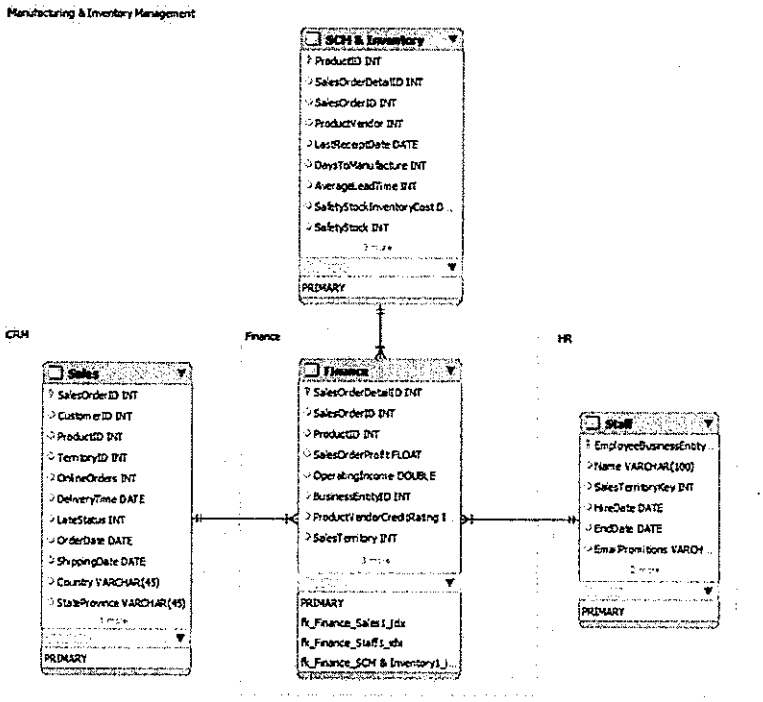
Answer B

11. The situation where 'deleting the IT department removes an employee from the database' is known as the Deletion Anomaly and can be addressed by

A. First Normal Form
B. Second Normal Form
C. Third Normal Form
D. None of the above

Answer C

12. Consider the following data warehouse schema:

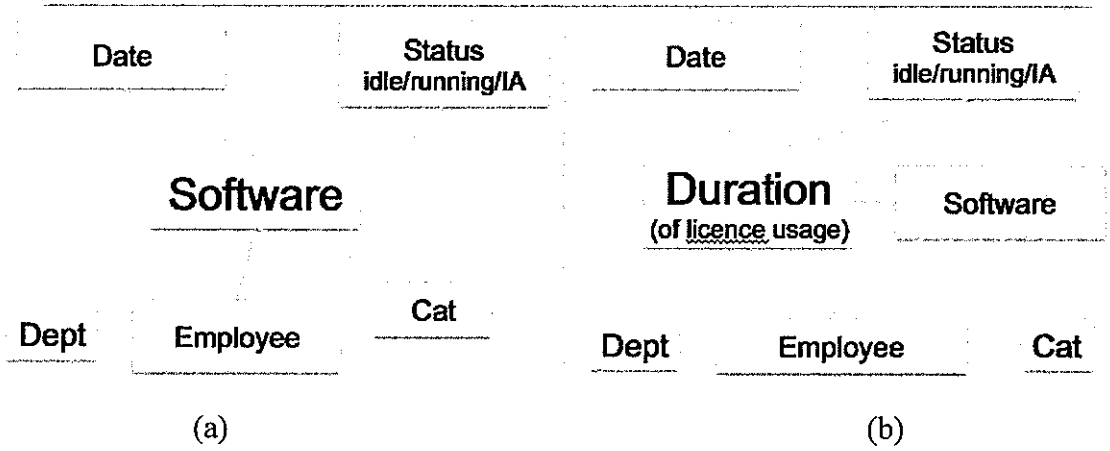


A. It is a good Data Warehouse schema
B. The problem with the schema is that the fact table Sales is not in the center
C. The problem with the schema is that the snowflake schema has too few dimensions
D. The problem with the schema is that the fact table Finance does not have all the facts
E. The problem with the schema is that the fact table Sales does not contain SalesOrderProfit

Answer D

13. Company spends much money on licences for software. You start paying when you open software and stop when you terminate it. Software use is interactive or running (e.g., simulation), but software can also be idle. Management

wants to know if they pay a lot of money of started software per category that is idle for a long time.



(a) (b)

In the figure above comparing (a) Data Warehouse schema with (b) we can conclude

A. Both (a) and (b) are good Data Warehouse schemas
B. Schema (b) is better than (a) as type of Software is not a fact
C. The business is concerned about the software and that is central to the main problem so Schema (a) is better than (b)
D. Schema (b) is not a valid star schema so (a) is better than (b)
E. For the business problem one requires only the duration of the licence information and therefore schema (b) is better

Answer E.

A supermarket stores all the transactions in a large database. These transactions database can be used for "basket analysis". For the sake of simplicity and time we focus only on the following small part of the database and the items:

A supermarket stores all the transactions in a large database. These transactions data base can be used for "basket analysis". For the sake of simplicity and time we focus o nly on the following small part of the database and the items:

TID	Bread	Milk	Beer	Eggs	Cola
1	1	1	1	1	1
2	0	1	0	1	1
3	1	0	1	1	1
4	0	1	1	0	0
5	1	1	1	1	0
6	1	1	1	0	0
7	1	0	1	0	1
8	0	1	1	1	1
9	1	0	0	0	1
10	1	1	1	1	0

11	0	1	1	1	1
12	1	1	1	0	0
13	1	1	1	0	1
14	1	0	0	0	1
15	0	1	1	1	0

Part of a transaction data base

14. In the above table, the support and confidence of the following association rules: {eggs} => {milk}

A. 7/15, 7/11
B. 7/8, 7/15
C. 7/15, 7/8
D. 8/12, 8/15
E. 7/11, 7/15

Answer C

15. The support and confidence for {bread} => {cola} is:

A. 7/15, 7/10
B. 7/10, 7/15
C. 6/15, 6/10
D. 6/15, 6/9
E. 6/10, 6/15

Answer C

16. The support and confidence of {beer, eggs} => {milk} is:

A. 6/15, 6/12
B. 6/11, 6/15
C. 6/12, 6/15
D. 6/15, 6/7
E. 6/15, 6/9

Answer D

17.The frequent 1 item sets with support ≥ 50% from Table 1 are:

A. {Bread}, {Beer}, {Eggs},{Milk}, {Cola}
B. {Bread}, {Beer}, {Eggs},{Milk},
C. {Beer}, {Eggs},{Milk}, {Cola}
D. {Bread}, {Eggs},{Milk}, {Cola}
E. {Bread}, {Eggs},{Milk},

Answer A.

18. All the frequent 2 item sets with support ≥ 50% from Table 1 are:

A. {bread, milk}, {bread, beer}, {bread, cola},{milk, beer}
B. {bread, milk}, {bread, beer}, {bread, cola}
C. {bread, beer}, {bread, cola},{milk, beer

D. {bread, beer},{milk, beer}

E. {bread, milk}, {bread, beer}

Answer D

19. All the frequent 3 item sets with support ≥ 50% from Table 1 are:

A. {bread, milk, cola}, {bread, cola, beer}
B. {bread, eggs, beer}, {bread, cola, milk}
C. {bread, cola, milk}
D. {bread, milk, beer}
E. None of the above

Answer E.

20. Assume that the confidence of the decision rule, *a => b,* is *80%,* then the confidence of the decision rule, *b=>a* is always:

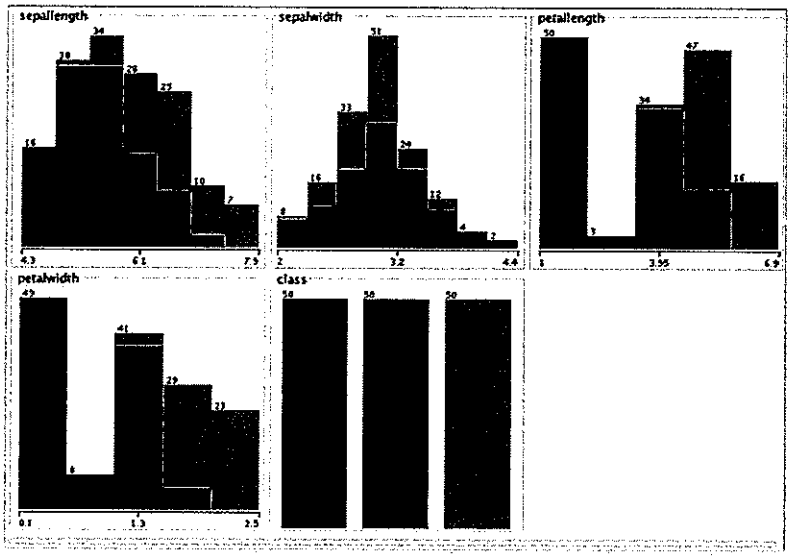
A) 80%.
B) <80%.
C) >=80%
D) <=80%
E) None of the above

Answer: E

21. Assume that the confidence of the decision rule, *a => {b,c}* is *<50%* then the support of the decision rule, *b=>{a,c}* is always:
A) 50%.
B) <50%.
C) >=50%
D) >50%
E) None of the above

Answer B

The figure below shows the relation of 4 attributes to the class of a flower. The attributes being sepal length, sepal width, petal length and petal width. The class of the flower is 3 types: iris setosa (blue), iris versicolor (red) and iris virginica (light blue).



22. The attribute that most clearly determines the class of the flower is:
    A) Sepal length followed by sepal width
    B) Sepal width followed by petal width
    C) Petal length followed by sepal length
    D) Petal width followed by petal length
    E) Petal length followed by sepal width

Answer: D

23. In _____, the problem is to group an unlabeled collection of objects, such as documents, customer comments, and Web pages into meaningful groups without any prior knowledge.
    A) search recall
    B) classification
    C) clustering
    D) grouping
    E) neural networking

Answer C

24. Data that has an arbitrary zero point is _____ data.
    A) categorical
    B) nominal
    C) interval
    D) ratio
    E) numerical

Answer: C

_____ data mining begins with a proposition by the user, who then seeks to validate

the truthfulness of the proposition. For example, a marketing manager may begin with the following proposition: "Are BluRay player sales related to sales of HDTV sets?"

A) Hypothesis-driven
B) Theory-driven
C) Discovery-driven
D) Data-driven
E) Predictive

Answer A

26. Which of the following issues related to BI implementation is false?
A) BI and predictive analytics can lead to serious ethical issues such as privacy and accountability.
B) Developing an effective BI application is no longer complex.
C) Smaller organizations can make the solutions cost effective if they leverage existing databases rather than create new ones.
D) The quality and timeliness of business information for an organization is not the choice between profit and loss--it may be a question of survival.
E) All the above

Answer: B

27. Which of the following correctly defines a text mining term?
A. Tagging is the number of times a word is found in a specific document.
B. A token is an uncategorized block of text in a sentence.
C. Rooting is the process of reducing inflected words to their base form.
D. A term is a single word or multiword phrase extracted directly from the corpus by means of NLP methods.
E. All the above

Answer: D

28. In 10 fold cross validation:
A) the classifier is iteratively trained on one fold and tested on one fold and the result is averaged
B) the classifier is trained on ten folds and tested on ten folds
C) the classifier is iteratively trained on one fold and tested on ten folds and the result is averaged
D) the classifier is trained on one fold and tested on the remaining nine folds iteratively and the result is averaged
E) the classifier is trained on nine folds and tested on the remaining one fold iteratively and the result is averaged

Answer: E

29. Which of the following is <u>not</u> a balanced scorecard perspective
A) Internal business process
B) Financial
C) Marketing and advertising
D) Innovation and learning

E) Customer

Answer C

30. Which of the following would be a lead measure for a balanced scorecard?
A) customer profitability
B) cost per employee
C) return on investment
D) employee training hours
E) sales per employee

Answer D

BONUS Question for 2 points,

Please note: There will be no points given for answers without clear Gini index calculation – if you are doing this submit the extra sheets along with the Multiple Choice answer sheets

31. ***Data mining exercise.***
A financial company wants to automatize the applications for a loan. In order to build a classification model the company uses the part of the database concerning loan applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. The relevant attributes, determined by domain experts, are for convenience denoted by A, B, C and D. The values for A are a and b, the values for B are u and y, the values for C are g and p and the values for D are w, q and m. A '+' means that the customer did pay his or her loan in time, a '–' means that he or she did not pay back the loan in time.

A	B	C	D	Approval
b	u	g	w	+
a	u	g	q	+
a	u	g	q	+
b	u	g	w	+
b	u	g	w	+
b	u	g	m	+
b	u	g	m	+
b	u	g	m	-
a	u	g	m	-
b	u	g	m	-
b	y	p	m	-
b	u	g	q	-
b	u	g	w	-
a	y	p	w	-
b	y	p	m	-
b	u	g	q	-

The GINI Index for a given node t in the decision tree is:

$$GINI(t) = 1 - \sum_{j} [p(\frac{j}{t})]^2$$

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Where $p(\frac{j}{t})$ is the relative frequency of class j at node t.

n_i = number of instances at child i,

n = number of instances at node t.

Based on the Gini Index, the best attribute for the root node of the decision tree is:

A. A
B. B
C. C
D. D
E. Approval

Answer: B or C