Knowledg

Data Science and Machine Learning

Fall 2022 Semester

Instructors: V. Kantere, D. Tsoumakos

Name: ____________

ID: ____________

Location: ____________

Date: ____________

Read each question carefully, then print the letter of the correct answer on the line next to the question.

Question Set 1 (70%)

1.______ Let a data warehouse cube consist of 4 dimensions, each of which has 5 hierarchy levels (including **all**). How many cuboids exist, including the base and apex cuboids? (4%)

a. 4^5

b. 5^4

c. 2^4

d. 20

2.______ How is the snowflake schema different from the star schema? (4%)

a. Snowflake does not contain a fact table

b. Dimension tables are split into further tables in the snowflake schema

c. Star schema has dimension tables, snowflake does not

d. The fact table in snowflake is smaller than in star

3.______ A data cube has n dimensions, and each dimension has exactly p distinct values in the base cuboid. Assume that there are no concept hierarchies. What is the maximum number of cells possible in the base cuboid? (4%)

a. p^n

b. p

c. pn

d. 2^n

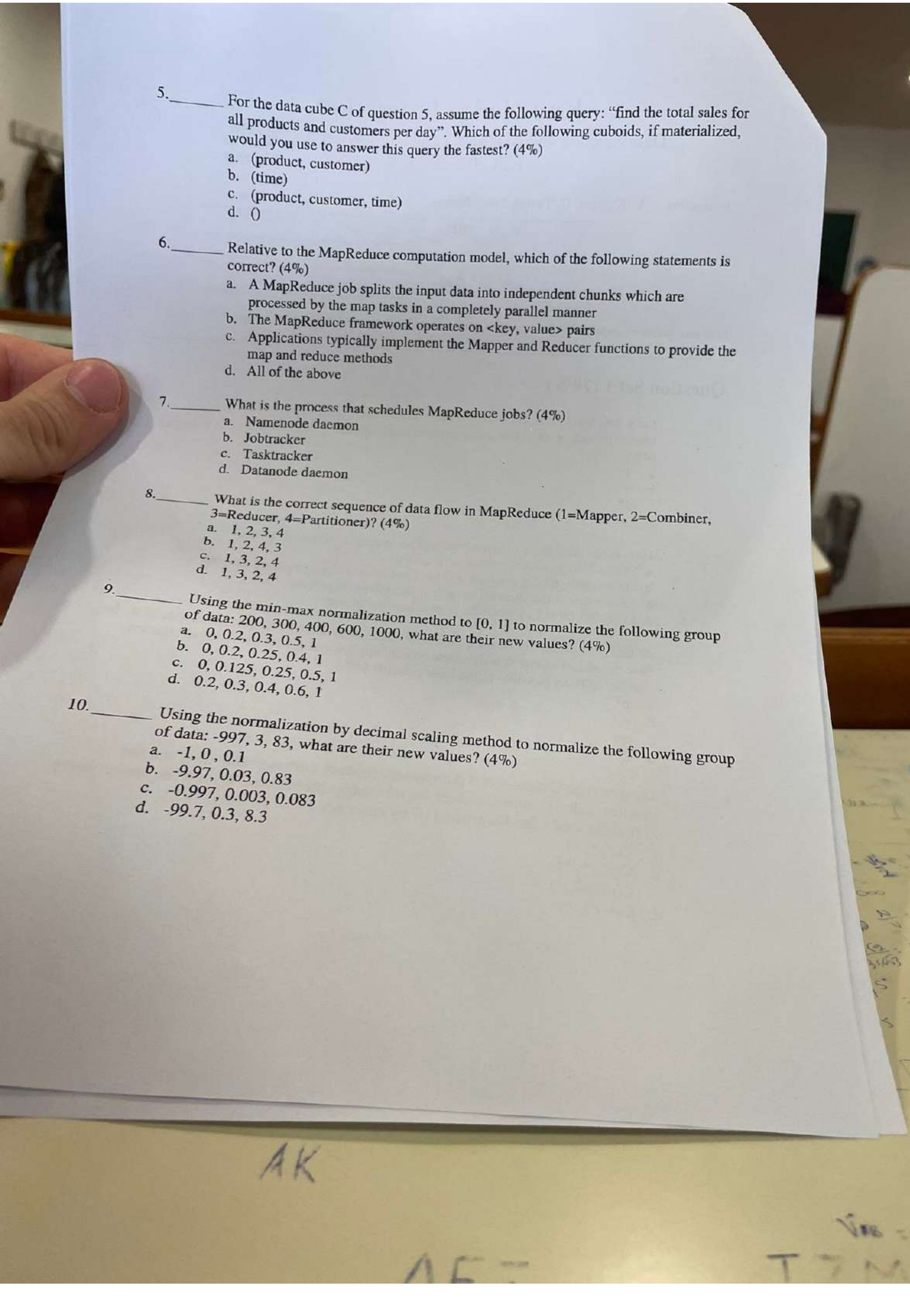
4.______ Let data cube C consist of 3 dimensions, product, customer and time, and one measure, sale. Assume that there are no concept hierarchies. What is the correct que to find the total sales for product p7 by customer c7 for all time? (4%)

a. C(*, *, *)

b. C(p7, *, *)

c. C(p7, c7, *)

d. C(*, c7, *)



5.______ For the data cube C of question 5, assume the following query: "find the total sales for all products and customers per day". Which of the following cuboids, if materialized, would you use to answer this query the fastest? (4%)

a. (product, customer)
b. (time)
c. (product, customer, time)
d. ()

6.______ Relative to the MapReduce computation model, which of the following statements is correct? (4%)

a. A MapReduce job splits the input data into independent chunks which are processed by the map tasks in a completely parallel manner
b. The MapReduce framework operates on <key, value> pairs
c. Applications typically implement the Mapper and Reducer functions to provide the map and reduce methods
d. All of the above

7.______ What is the process that schedules MapReduce jobs? (4%)

a. Namenode daemon
b. Jobtracker
c. Tasktracker
d. Datanode daemon

8.______ What is the correct sequence of data flow in MapReduce (1=Mapper, 2=Combiner, 3=Reducer, 4=Partitioner)? (4%)

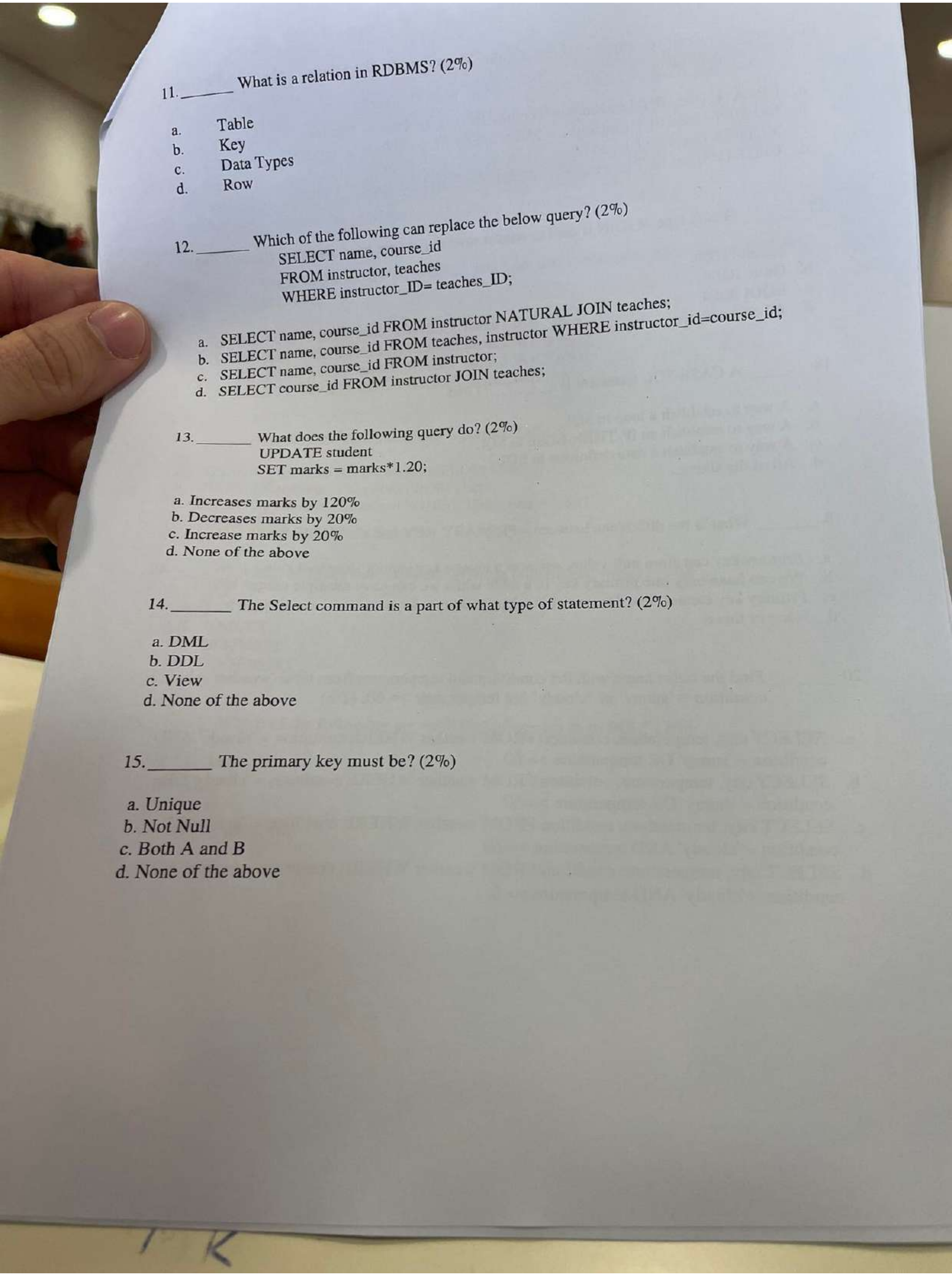
a. 1, 2, 3, 4
b. 1, 2, 4, 3
c. 1, 3, 2, 4
d. 1, 3, 2, 4

9.______ Using the min-max normalization method to [0, 1] to normalize the following group of data: 200, 300, 400, 600, 1000, what are their new values? (4%)

a. 0, 0.2, 0.3, 0.5, 1
b. 0, 0.2, 0.25, 0.4, 1
c. 0, 0.125, 0.25, 0.5, 1
d. 0.2, 0.3, 0.4, 0.6, 1

10.______ Using the normalization by decimal scaling method to normalize the following group of data: -997, 3, 83, what are their new values? (4%)

a. -1, 0 , 0.1
b. -9.97, 0.03, 0.83
c. -0.997, 0.003, 0.083
d. -99.7, 0.3, 8.3



11.______ What is a relation in RDBMS? (2%)

a. Table
b. Key
c. Data Types
d. Row

12.______ Which of the following can replace the below query? (2%)

```
SELECT name, course_id
FROM instructor, teaches
WHERE instructor_ID= teaches_ID;
```

a. SELECT name, course_id FROM instructor NATURAL JOIN teaches;
b. SELECT name, course_id FROM teaches, instructor WHERE instructor_id=course_id;
c. SELECT name, course_id FROM instructor;
d. SELECT course_id FROM instructor JOIN teaches;

13.______ What does the following query do? (2%)

```
UPDATE student
SET marks = marks*1.20;
```

a. Increases marks by 120%
b. Decreases marks by 20%
c. Increase marks by 20%
d. None of the above

14.______ The Select command is a part of what type of statement? (2%)

a. DML
b. DDL
c. View
d. None of the above

15.______ The primary key must be? (2%)

a. Unique
b. Not Null
c. Both A and B
d. None of the above

16.______ How can you change "Thomas" into "Michel" in the "LastName" column in the Users table? (2%)

a. UPDATE User SET LastName = 'Thomas' INTO LastName = 'Michel'
b. MODIFY Users SET LastName = 'Michel' WHERE LastName = 'Thomas'
c. MODIFY Users SET LastName = 'Thomas' INTO LastName = 'Michel'
d. UPDATE Users SET LastName = 'Michel' WHERE LastName = 'Thomas'

17.______ Which type of JOIN is used to returns rows that do not have matching values? (2%)

a. Natural JOIN
b. Outer JOIN
c. EQUI JOIN
d. All of the above

18.______ A CASE SQL statement is ________? (2%)

a. A way to establish a loop in SQL.
b. A way to establish an IF-THEN-ELSE in SQL
c. A way to establish a data definition in SQL
d. All of the above.

19.______ What is the difference between a PRIMARY KEY and a UNIQUE KEY? (2%)

a. Primary key can store null value, whereas a unique key cannot store null value.
b. We can have only one primary key in a table while we can have multiple unique keys.
c. Primary key cannot be a date variable whereas unique key can be.
d. None of these.

20.______ Find the cities name with the condition and temperature from table 'weather' where condition = 'sunny' or 'cloudy' but temperature >= 60. (2%)

a. SELECT city, temperature, condition FROM weather WHERE condition = 'cloudy' AND condition = 'sunny' OR temperature >= 60
b. SELECT city, temperature, condition FROM weather WHERE condition = 'cloudy' OR condition = 'sunny' OR temperature >= 60
[✓ c.] SELECT city, temperature, condition FROM weather WHERE condition = 'sunny' OR condition = 'cloudy' AND temperature >= 60
d. SELECT city, temperature, condition FROM weather WHERE condition = 'sunny' AND condition = 'cloudy' AND temperature >= 6

21.______ Which of the following statement is correct to display all the cities with the condition, temperature, and humidity whose humidity is in the range of 60 to 75 from the 'weather' table? (2%)

a. SELECT * FROM weather WHERE humidity IN (60 to 75)
b. SELECT * FROM weather WHERE humidity BETWEEN 60 AND 75
c. SELECT * FROM weather WHERE humidity NOT IN (60 AND 75)
d. SELECT * FROM weather WHERE humidity NOT BETWEEN 60 AND 75

22.______ Which statement is used to get all data from the student table whose name starts with p? (2%)

a. SELECT * FROM student WHERE name LIKE '%p%';
b. SELECT * FROM student WHERE name LIKE 'p%';
c. SELECT * FROM student WHERE name LIKE '_p%';
d. SELECT * FROM student WHERE name LIKE '%p';

23.______ Which of the SQL statements is correct? (2%)

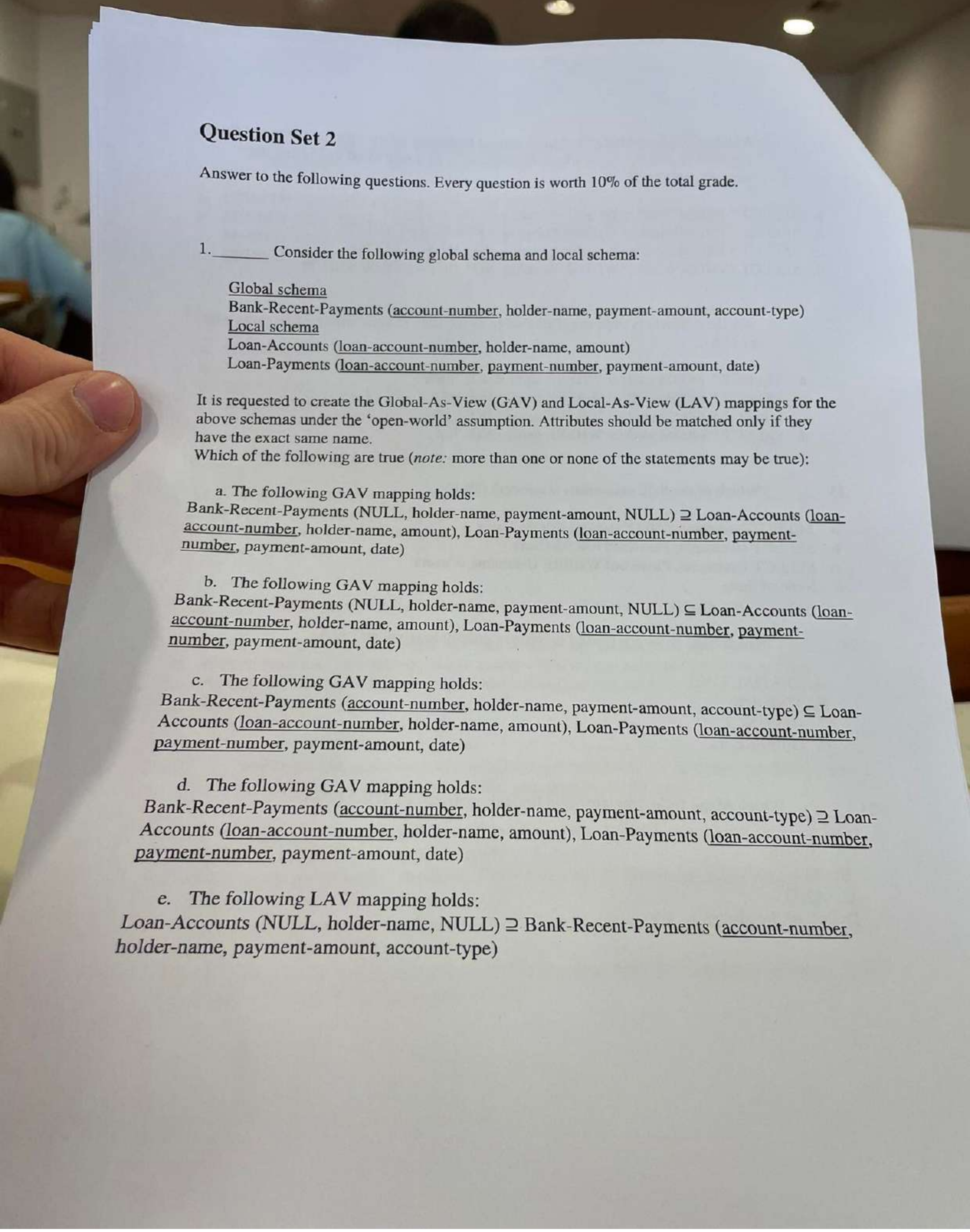
a. SELECT Username AND Password FROM Users
b. SELECT Username, Password FROM Users
c. SELECT Username, Password WHERE Username = 'user1'
d. None of these

24.______ Which SQL keyword is used to retrieve only unique values? (2%)

A. DISTINCTIVE
B. UNIQUE
C. DISTINCT
D. DIFFERENT

25.______ Which of the following are valid logical operators in SQL? (2%)

A. SOME
B. ALL
C. AND
D. All of the above



Question Set 2

Answer to the following questions. Every question is worth 10% of the total grade.

1.______ Consider the following global schema and local schema:

Global schema
Bank-Recent-Payments (account-number, holder-name, payment-amount, account-type)
Local schema
Loan-Accounts (loan-account-number, holder-name, amount)
Loan-Payments (loan-account-number, payment-number, payment-amount, date)

It is requested to create the Global-As-View (GAV) and Local-As-View (LAV) mappings for the above schemas under the 'open-world' assumption. Attributes should be matched only if they have the exact same name.
Which of the following are true (*note:* more than one or none of the statements may be true):

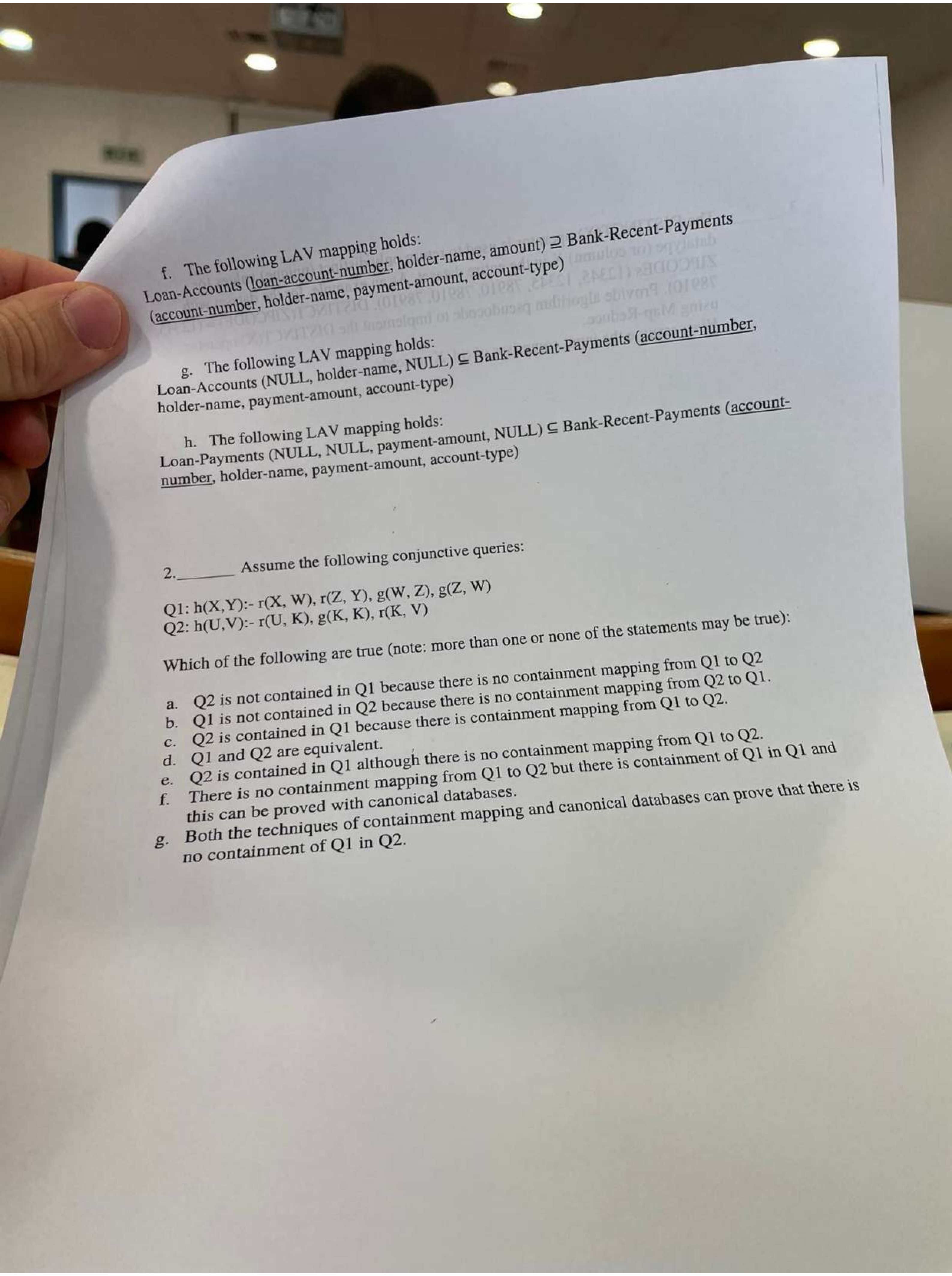
a. The following GAV mapping holds:
Bank-Recent-Payments (NULL, holder-name, payment-amount, NULL) $\supseteq$ Loan-Accounts (loan-account-number, holder-name, amount), Loan-Payments (loan-account-number, payment-number, payment-amount, date)

b. The following GAV mapping holds:
Bank-Recent-Payments (NULL, holder-name, payment-amount, NULL) $\subseteq$ Loan-Accounts (loan-account-number, holder-name, amount), Loan-Payments (loan-account-number, payment-number, payment-amount, date)

c. The following GAV mapping holds:
Bank-Recent-Payments (account-number, holder-name, payment-amount, account-type) $\subseteq$ Loan-Accounts (loan-account-number, holder-name, amount), Loan-Payments (loan-account-number, payment-number, payment-amount, date)

d. The following GAV mapping holds:
Bank-Recent-Payments (account-number, holder-name, payment-amount, account-type) $\supseteq$ Loan-Accounts (loan-account-number, holder-name, amount), Loan-Payments (loan-account-number, payment-number, payment-amount, date)

e. The following LAV mapping holds:
Loan-Accounts (NULL, holder-name, NULL) $\supseteq$ Bank-Recent-Payments (account-number, holder-name, payment-amount, account-type)



f. The following LAV mapping holds:
Loan-Accounts (loan-account-number, holder-name, amount) ⊇ Bank-Recent-Payments (account-number, holder-name, payment-amount, account-type)

g. The following LAV mapping holds:
Loan-Accounts (NULL, holder-name, NULL) ⊆ Bank-Recent-Payments (account-number, holder-name, payment-amount, account-type)

h. The following LAV mapping holds:
Loan-Payments (NULL, NULL, payment-amount, NULL) ⊆ Bank-Recent-Payments (account-number, holder-name, payment-amount, account-type)

2._______ Assume the following conjunctive queries:

Q1: h(X,Y):- r(X, W), r(Z, Y), g(W, Z), g(Z, W)
Q2: h(U,V):- r(U, K), g(K, K), r(K, V)

Which of the following are true (note: more than one or none of the statements may be true):

a. Q2 is not contained in Q1 because there is no containment mapping from Q1 to Q2
b. Q1 is not contained in Q2 because there is no containment mapping from Q2 to Q1.
c. Q2 is contained in Q1 because there is containment mapping from Q1 to Q2.
d. Q1 and Q2 are equivalent.
e. Q2 is contained in Q1 although there is no containment mapping from Q1 to Q2.
f. There is no containment mapping from Q1 to Q2 but there is containment of Q1 in Q1 and this can be proved with canonical databases.
g. Both the techniques of containment mapping and canonical databases can prove that there is no containment of Q1 in Q2.

3._______ The DISTINCT(X) operator is used to return only distinct (unique) values for datatype (or column) X in the entire dataset. As an example, for the dataset with ZIPCODEs (12345, 12345, 78910, 78910, 78910), DISTINCT(ZIPCODE) = (12345, 78910). Provide algorithm pseudocode to implement the DISTINCT(X) operator using Map-Reduce.

Hint: assume the input comprises of (id, X) records