Examples of Open Question for the midterm Exam Preparation

Kindly note that these questions will not be in the midterm exam, but you will see similar questions.

Q1

- 1.1. (2 pnts) If we have the schema:

  game(*id*, mdate, stadium, team1, team2, tournament)

  eteam(*id*, teamname, coach)

  - Write an SQL query that returns the names of the coaches of the teams who played a game on 20-06-2008

- 1.2.(2 pnts) Write the Armstrong Axioms of the functional dependencies and use them to prove that if $(A \rightarrow B) \wedge (BD \rightarrow E)$ $then$ $(AD \rightarrow E)$

Q1

- 1.3. (2 pnts) Using referential integrity, we can avoid having dangling references. What do we mean by dangling references and how they can appear in the database?
- 1.4. (2 pnts): Formulate the definition of functional dependency and explain why a primary key will determine all the attributes in the relation.

Q2

- 2.1. (2 pnts) Explain how z-score normalization works and use it to normalize the data D = {1, 5, 9, 13, 17, 21, 25}.
- 2.2. (2 pnts) If you have the following data: D = {2, 5, 77, 6, 15, 30, 8, 17, 45, 25, 60, 28}, what will be the contents of the 3 bins that you will use to smooth the data by bin boundaries?
- 2.3. (2 pnts) Describe the parametric statistical approach for outlier detection and use the assumption that the data is normally distributed to detect the outliers in the following dataset:

  D = {2, 5, 3, 6, 4, 16, 3, 4, 2, 3, 7, 5}.

Additional Questions

- Given the following documents:
  - D1 = {com, omp, mpu, put, ute}
  - D2 = {imp, mpu, put, ute}
  - D3 = {sim, imp, mpl, ple}

a. Construct the document-shingle matrix

b. Given the permutations in the table, compute the signature matrix using minhash.

Index	shingles	P1	P2	P3	P4
1	com	2	9	9	7
2	omp	1	6	5	4
3	mpu	8	3	1	5
4	put	3	8	6	8
5	ute	4	5	2	9
6	imp	6	2	3	6
7	sim	9	7	8	1
8	mpl	7	4	4	2
9	ple	5	1	7	3

Additional Questions

How we compare the output of the following two queries.

Select * from students where sid == null

Select * from students where sid <> null

Justify your answer.

Define the following terms: relation schema, domain, relation cardinality, relation degree.

Additional Questions

- Consider the following schema (the underlined attributes represent primary keys):

  Emp(eid: integer, ename: string, age: integer, salary: real)

  Works(eid: integer, did: integer, pct_time: integer)

  Dept(did: integer, dname: string, budget: real, managerid: integer)

  I. Give an example of a foreign key constraint that involves the Dept relation.

  II. What are the options for enforcing this constraint when a user attempts to delete a Dept tuple?

  III. When defining the Dept relation in SQL, what should be done to make sure that every department is guaranteed to have a manager.

The Answers

Q1

- 1.1. (2 pnts) If we have the schema:

  game(*id*, mdate, stadium, team1, team2, tournament)

  eteam(*id*, teamname, coach)

  - Write an SQL query that returns the names of the coaches of the teams who played a game on 20-06-2008

Answer: the query is:

```
SELECT coach FROM eteam as et
JOIN game as g on et.id = g.team1
  WHERE mdate = '20-06-2008'
UNION
SELECT coach FROM eteam as et
JOIN game as g on et.id = g.team2
  WHERE mdate = '20-06-2008'
```

Q1

- 1.2.(2 pnts) Write the Armstrong Axioms of the functional dependencies and use them to prove that if $(A \rightarrow B) \wedge (BD \rightarrow E)\ then\ (AD \rightarrow E)$
- **Answer:**
  - Armstrong Axioms:
    - Reflexivity: if $X \subseteq Y$, then $Y \longmapsto X$
    - Augmentation: if X $\longmapsto Y$, then AX $\longmapsto AY$
    - Transitivity: if $X \longmapsto Y$ and $Y \longmapsto W$, then $X \longmapsto W$
  - $(A \rightarrow B) \wedge (BD \rightarrow E)\ then\ (AD \rightarrow E)$
    - Since $A \rightarrow B$ then $AD \rightarrow BD$ (augmentation)
    - $AD \rightarrow BD$ and $BD \rightarrow E$ then $AD \rightarrow E$ (transitivity)

Q1

- 1.3. (2 pnts) Using referential integrity, we can avoid having dangling references. What do we mean by dangling references and how they can appear in the database?
  - **Answer:** Dangling references appear when we modify/delete a value(s) of a primary key attribute(s) that is included among the value(s) of foreign key attribute(s) in another table

Q1

- 1.4. (2 pnts): Formulate the definition of functional dependency and explain why a primary key will determine all the attributes in the relation.
- Answer: the functional dependency definition states that: Let T be a relation (table), if we have a functional dependency from set of attributes X to another set of attributes Y then: if there are two records in table with the same values for the set of attributes X, then they should have the same values for the set of attributes Y.
- That is:
  - If $X \longmapsto Y \ \ in\ T\ then\ \forall r_1, r_2 \in T \qquad r_1[X] = r_2[X] \Rightarrow r_1[Y] = r_2[Y]$

Q1

FD: If $X \longmapsto Y \ \ in\ T\ then\ \forall r_1, r_2 \in T \quad r_1[X] = r_2[X] \Rightarrow r_1[Y] = r_2[Y]$

- Why a primary key will determine all the attributes in the relation
- **Answer:** since the attributes that represent primary key have different values in each record, the condition $r_1(X) = r_2(X)$ will never be true so the Boolean expression $r_1(X) = r_2(X) \Rightarrow r_1(Y) = r_2(Y)$ will be always true.

A	B	$A \Rightarrow B$
T	T	T
T	F	F
F	T	T
F	F	T

Truth table for the ⇒ operator

Q2

- 2.1. (2 pnts) Explain how z-score normalization works and use it to normalize the data D = {1, 5, 9, 13, 17, 21, 25}.
- Mean(D) = 13
- Std (D) = 8
- Normalized(D) = {-1.5, -1, -0.5, 0, 0.5, 1, 1.5}

Q2

- 2.2. (2 pnts) If you have the following data: D = {2, 5, 77, 6, 15, 30, 8, 17, 45, 25, 60, 28}, what will be the contents of the 3 bins that you will use to smooth the data by bin boundaries?
- Answer:
  - Since it didn't mention what technique should we use, we can select any technique
- 2.3. (2 pnts) Describe the parametric statistical approach for outlier detection and use the assumption that the data is normally distributed to detect the outliers in the following dataset:

  D = {2, 5, 3, 6, 4, 16, 3, 4, 2, 3, 7, 5}.

Q2

- 2.2. (2 pnts) If you have the following data: D = {2, 5, 77, 6, 15, 30, 8, 17, 45, 25, 60, 28}, what will be the contents of the 3 bins that you will use to smooth the data by bin boundaries?
- **Answer:**
  - Since it wasn't mentioned, which technique we should use, we can select any technique
  - I will use both:
  - Equi-depth binning: We sort the data to get DS = {2, 5, 6, 8, 15, 17, 25, 28, 30, 45, 60, 77} then
    - B1 = {2, 5, 6, 8}, B2 = {15, 17, 25, 28}, B3 = {30, 45, 60, 77}
    - B1s = {2, 2, 8, 8}, B2s = {15, 15, 28, 28}, B3s = {30, 30, 77, 77}
  - Equi-Width binning: We will use the DS also: bin width = (max(D) – min(D) + 1) / #bins = 76 / 3 = 25.33
    - B1 = {2, 5, 6, 8, 15, 17, 25} elements in [2, 27.33) then B1s = {2, 2, 2, 2, 25, 25, 25}
    - B2 = {28, 30, 45} elements between [27.33, 52.66) then B2s = {28, 28, 45}
    - B3 = {60, 77} elements between [52.66, 78) then B2s = {60, 77}

Q2

- 2.3. (2 pnts) Describe the parametric statistical approach for outlier detection and use the assumption that the data is normally distributed to detect the outliers in the following dataset:

  D = {2, 5, 3, 6, 4, 16, 3, 4, 2, 3, 7, 5}.

Answer:

With the normality assumption (data is extracted from normal distribution), we consider a value as outlier if the distance from the value to the mean of the data is greater than 3 times the standard deviation of the data. That is

Let m = mean (D) and s = std(D), $\forall x \in D, if\ |x - m| > 3 * s$ then outlier.

For the given data: m = 5, s = 3.63

Outliers = {16}

Additional Questions

- Given the following documents:
  - D1 = {com, omp, mpu, put, ute}
  - D2 = {imp, mpu, put, ute}
  - D3 = {sim, imp, mpl, ple}

a. Construct the document-shingle matrix

b. Given the permutations in the table, compute the signature matrix using MinHash.

Index	shingles	P1	P2	P3	P4
1	com	2	9	9	7
2	omp	1	6	5	4
3	mpu	8	3	1	5
4	put	3	8	6	8
5	ute	4	5	2	9
6	imp	6	2	3	6
7	sim	9	7	8	1
8	mpl	7	4	4	2
9	ple	5	1	7	3

Additional Questions

a. Construct the document-shingle matrix

b. Given the permutations in the table, compute the signature matrix using MinHash.

P/D	D1	D2	D3
P1	1	3	5
P2	3	2	1
P3	1	1	3
P4	4	5	1

P1	P2	P3	P4	Index	shingles	D1	D2	D3
2	9	9	7	1	com	1	0	0
1	6	5	4	2	omp	1	0	0
8	3	1	5	3	mpu	1	1	0
3	8	6	8	4	put	1	1	0
4	5	2	9	5	ute	1	1	0
6	2	3	6	6	imp	0	1	1
9	7	8	1	7	sim	0	0	1
7	4	4	2	8	mpl	0	0	1
5	1	7	3	9	ple	0	0	1

Additional Questions

How we compare the output of the following two queries.

Select * from students where sid == null

Select * from students where sid <> null

Justify your answer.

These two queries are the same

Any operation that involves null will result in unknown, so the result of each condition is unknown (See the next slide)

Null Values and Three Valued Logic

∧	*T*	*F*	*U*
T	T	F	U
F	F	F	F
U	U	F	U

∨	*T*	*F*	*U*
T	T	T	T
F	T	F	U
U	T	U	U

- Three values – *true*, *false*, *unknown*
- Any comparison with *null* returns *unknown*
  - Example*: 5 < null or null <> null or null = null*
- Three-valued logic using the value *unknown*:
  - OR: (*unknown* **OR** *true*) = *true*,
    (*unknown* **OR** *false*) = *unknown*
    (*unknown* **OR** *unknown*) = *unknown*
  - AND: *(true* **AND** *unknown) = unknown,*
    (false **AND** *unknown) = false,*
    (unknown **AND** *unknown) = unknown*
  - NOT*: (***NOT** *unknown) = unknown*
  - "*P* **is unknown**" evaluates to true if predicate *P* evaluates to *unknown*
- Result of **WHERE** clause predicate is treated as *false* if it evaluates to *unknown*

Additional Questions

Define the following terms: relation schema, domain, relation cardinality, relation degree.

Find the definitions in the book

Additional Questions

- Consider the following schema (the underlined attributes represent primary keys):

  Emp(eid: integer, ename: string, age: integer, salary: real)

  Works(eid: integer, did: integer, pct_time: integer)

  Dept(did: integer, dname: string, budget: real, managerid: integer)

  I. Give an example of a foreign key constraint that involves the Dept relation.

  Answer:

  The did is a primary key in the `Dept' relation and a foreign key in the `Works' relation

Additional Questions

II. What are the options for enforcing this constraint when a user attempts to delete a Dept tuple?

Answer:

There are three options:

1. Set default/null: set the foreign key value to a default/null value when the corresponding primary key value is deleted/modified
2. Cascade: Apply the same changes that happened on the primary key values on the corresponding foreign key values.
3. no action: reject the modifications of the primary key values if there are corresponding foreign key values.

Additional Questions

- Compare Jaro and Jaro-Winkler similarity. Give an example of two strings and compute both similarities between the strings.
- **Answer:** Jaro-Winkler gives more weight for the leading common prefix. Examples can be found in the slides.