School of Engineering and Computer Science

SWEN 304 **Database System Engineering**

Assignment 1

The objective of this assignment is to test your understanding of database foundations, basic terms, and the relational data model the entity relational model. It is worth 15% of your final grade. The assignment is marked out of 100.

The assignment is due on **Friday, 28 March, 23:59 pm**. Please submit your assignment in **pdf** via the submission system.

Question 1 [20 marks]

The TOP500 project lists the 500 most powerful non-distributed computer systems in the world (also called supercomputers). Suppose we use a relational database to manage the current and future data of this project. For this purpose, we use a relation schema with the attribute set {Performance, Name, Manufacturer, Country, Year}.

The following table shows a portion of the current instance of the SUPERCOMPUTERS relation schema that stores data for some supercomputers. Note that the performance in the table is measured in petaFLOPS.

SUPERCOMPUTERS

Performance	Name	Manufacturer	Country	Year
442010	Fugaku	Fujitsu	Japan	2020
148600	Summit	IBM	United States	2018
94640	Sierra	IBM	United States	2018
93015	Sunway	NRCPC	China	2016
64590	Perlmutter	HPE	United States	2021
63460	Selene	Nvidia	United States	2020
61445	Tianhe-2A	NUDT	China	2013
44120	JUWELS	Atos	Germany	2020
35450	HPC5	Dell EMC	Italy	2020
30050	Voyager-EUS2	Microsoft	United States	2021

a) [8 marks] For every set of attributes (that is, for every subset of the attribute set) decide whether you can deduce that it is ***not*** a candidate key, assuming the shown instance is legal. Justify your answer.

Answer: A candidate key is the minimal set of attributes that can uniquely identify a tuple in the database.

{Manufacturer} - The manufacturer cannot be the candidate key because the same manufacturer could make multiple supercomputers. For instance, IBM manufactured 2 supercomputers.

{Country} - The country cannot be the candidate key because the same country could produce multiple supercomputers. In this case, for example, the United States has made 5 supercomputers.

{Year} - The year cannot be the candidate key because there could be multiple supercomputers made in the same year. For example, there were four supercomputers made in the year 2020.

{Manufacturer, Country} – This subset cannot be the candidate key because there are supercomputers that are made in the same country by the same manufacturer. In this case, there are two supercomputers made in the United States by the same manufacturer IBM.

{Manufacturer, Year} – This subset cannot be the candidate key because there are supercomputers that are made by the same manufacturer in the same year. In this case, there are two supercomputers made by the manufacturer IBM in the year 2018

{Year, Country} – This subset cannot be the candidate key because there are supercomputers that are made in the same country in the same year. In this relation, in the United States, there were two supercomputers made in the year 2018.

{Manufacturer, Country, Year} – This subset cannot be the candidate key because there are supercomputers that are made by the same manufacturer in the same country in the same year. In this case, there are two supercomputers made by the manufacturer IBM in the United States in the year 2018.

{Name, Manufacturer}, {Name, Country}, {Name, Year}, {Performance, Name, Year} , {Performance, Name, Manufacturer} , {Performance, Name, Country}, {Name, Manufacturer, Country}, {Name, Manufacturer, Year}, {Performance, Name, Manufacturer, Country}, {Performance, Name, Manufacturer, Year}, {Performance, Name, Country, Year}, {Country, Name, Manufacturer, Year} – None of these subsets can be the candidate key because it goes against the minimality property of a candidate key which states that no proper subset of the candidate key can uniquely identify a tuple. In this case, the name attribute on its own can uniquely identify a tuple, and the name is a proper subset. Therefore none of these can be the candidate key.

{Performance, Manufacturer}, {Performance, Country}, {Performance, Year}, {Performance, Name} - This subset could be used but it is not the best option as it is not minimal. Since the performance attribute alone can identify unique tuples, these subsets are not ideal for a candidate key.

b) **[4 marks]** For every remaining set of attributes (that is, for every set not ruled out as a candidate key in part **a)**), discuss whether you consider it a suitable candidate key? Justify your answer.

Answer:

{Name} – Name is a good candidate key because each supercomputer must have a unique name, otherwise it will have copyright issues. Therefore, I think that the Name attribute is a suitable candidate key.

{Performance} – Performance can also be a candidate key because in the table above, all the supercomputers have unique performance, so this can be used to uniquely identify the tuples.

c) **[2 marks]** Which of the candidate keys identified in part **b)** would you choose as the primary key? Justify your answer.

Answer: {Name} – Because the name attribute is meaningful and descriptive as opposed to the performance. Furthermore, it is less likely to change, whereas the performance of a supercomputer may change over time.

d) **[2 mark]** Add a new tuple for a computer into the SUPERCOMPUTERS relation. How would you check that the primary key identified in part **c)** is still valid?

Answer: I would check that the new tuple includes a value for the {Name} Attribute and is not null. Also check that the Name does not already exist on the SUPERCOMPUTERS relation.

e) **[2 mark]** Create a relation that shows for each country in the table above the country and the capital, i.e., use a relation schema with attribute set {Country, Capital}. How many records are in your relation? Justify your answer.

Answer: The table below shows the relation for each country, with the two attributes being Country and Capital. I will have no duplicates in this table since there is no need for there to be duplicate countries to each have their own tuple since they will all have the same capital, so there are 5 tuples in the relation. The number of records in the relation equals the number of unique countries.

Country	**Capital**
United States	Washington
Japan	Tokyo
China	Beijing
Germany	Berlin
Italy	Rome

f) **[2 mark]**. Consider a relation schema with attribute set {Manufacturer, City} and assume that both attributes have a domain with ten values each. What would be the maximum number of records in an instance of this relation schema?

Answer: The maximum number of records would equal the number of unique manufacturers.

Question 2 [10 marks]

Your university is using a relational database to manage its data on students and their study performance. Suppose the underlying database schema includes the following two relation schemas:

- STUDENT (SID: STRING, Name: STRING, DoB: DATE, GPA: REAL) with primary key {SID}
- RESULTS (CourId: STRING, StudId: STRING, Grade: STRING) with primary key {StudId, CourID} and foreign key StudID $\subseteq$ STUDENT[SID]

Below you find instances of these two relation schemas:

RESULTS

CourID	StudID	Grade
SWEN102	53666	A
SWEN221	53688	B
SWEN301	53832	B
SWEN224	53650	C

STUDENT

SID	Name	DoB	GPA
50000	Dave	22/01/1985	3.3
53666	Jones	11/02/1986	3.4
53688	Smith	22/01/1985	3.2
53650	Smith	15/05/1986	3.8
53831	Mathew	16/06/1984	1.8
53832	Jack	25/11/1983	3.0

Your tasks are as follows. **Justify your answers**!

Primary Key entity integrity constraint= no primary key values can be null.

a) [5 marks] Decide which of the following tuples can be inserted into the given instances.

1. Insert tuple ('53688', 'Mike', '16/06/1985', 3.4) into STUDENT

   Answer: This will not be inserted because there is already a SID with the same value in the table. Because the SID is a primary key, all SID values must be unique.

2. Insert tuple (*null*, 'Mike', '16/06/1985', 3.4) into STUDENT

   Answer: This will not be inserted because the SID is null and the primary key cannot be null because it is used to uniquely identify records.

3. Insert tuple ('SWEN224', '53650', 'B') into RESULTS

   Answer: This will not be inserted because there already exists a StudID with the same number and because this is the primary key, it will violate the primary key constraint.

4. Insert tuple ('SWEN102', '50505', 'B') into RESULTS

   Answer: Yes this will work because all values are specified.

5. Insert tuple ('SWEN222', *null*, 'B') into RESULTS

   Answer: No this will not work because the StudID cannot be null as this will violate the primary key constraint.

b) [5 marks] Decide which of the following tuples can be deleted from the given instances.

1. Delete tuple ('53831', 'Matt', '16/06/1984', 1.8) from STUDENT

   Answer: This will not work because in there is no tuple with the name 'Matt'.

2. Delete tuple ('53688', 'Smith', '22/01/1985', 3.2) from STUDENT

   Answer: Yes this will work and this tuple will be deleted because there is a tuple in the STUDENT database that matches this tuple.

3. Delete tuple ('50000', 'Dave', '22/01/1985', 3.3) from STUDENT

   Answer: Yes this will work and this tuple will be deleted because there is a tuple in the STUDENT database that matches this tuple.

4. Delete tuple ('SWEN301', '53832', 'B') from RESULTS

   Answer: Yes this will work and this tuple will be deleted because there is a tuple in the RESULTS database that matches this tuple.

5. Delete tuple ('SWEN301', '53832', *null*) from RESULTS

   Answer: This will not work because there is no row in the RESULTS database where the grade is null.

Question 3 **[20 marks]**

The Wellington Foreign Trade Office needs to translate hundreds of documents every day. To ensure professional translation in a timely manner the office cooperates with several translation agencies and expert translators in New Zealand. The processing of the data about translations as well as the checking of deadlines and quality requirements is time consuming and error prone if this is done manually on paper.

Therefore, the office wants to build a new database to record all relevant data that is needed for processing and checking translations. Suppose the following relation schemas have been proposed to be belong to the database schema for the new database.

- TranslationAgency (AgencyNumber, Name) with primary key {AgencyNumber}
- Translator (Name, Phone, Field, IRDNumber) with unknown primary key
- IsExpert (Name, Language) with primary key {Name, Language}
- TranslationOrder (OrderNumber, OrderDate, PageNumber, Budget, FromLanguage, ToLanguage, Deadline) with primary key {OrderNumber}
- Assignment (Agent, OrderNumber, Part, Language, Name) with primary key {OrderNumber, Part}

The following additional constraints are known:

1. Each translator has a unique IRD number, a unique phone and a unique name.
2. For each translator, the IRDNumber must be specified, while Field may be left blank (if not known).
3. Translators can be experts in up to four languages.
4. An agent can assign a translation order to multiple translators who can be distinguished by the assigned part of the order.

Your tasks are as follows:

a) [3 marks] For the relation schema **Translator**, identify all suitable candidate keys. Explain your answer. Which candidate key would you choose as the primary key? Justify your answer.

Answer: The possible candidate keys for the Translator relation schema would be {Name}, {Phone}, {IRDNumber}. Since the first constraint states that each translator has a unique IRD number, a unique phone number and a unique name, all three of these attributes can be used as candidate keys. Based on the second constraint, since the field value may be left blank, it cannot be the candidate key.

IRDNumber is the best option to be a primary key because a person's IRD number does not change frequently and it is unique for every record since 2 people cannot have the same IRDNumber. It is unique, minimal and stable. Also, based on the second constraint, the IRDNumber must be specified.

b) **[5 marks]** For each of the relation schemas, identify all suitable foreign keys (if there are any). Explain your answer.

Answer:

TranslationAgency[AgencyNumber, Name]

= no foreign keys

Translator[Name, Phone, Field, IRDNumber]

= no foreign keys

IsExpert[Name, Language]

= Constraint 1 states that each translator has a unique name. Since the name is unique for each translator, the name in the IsExpert schema can act as a foreign key referencing the Translator schema name.

TranslationOrder[OrderNumber, OrderDate, pageNumber, budget, Fromlanguage, ToLanguage, Deadline]

= no foreign key

Assignment[Agent, OrderNumber, Part, Language, Name] l

= The OrderNumber can act as a foreign key referencing the TranslationOrder Schema to determine what order number the given assignment is referring to.

Furthermore, the Name could also act as a foreign key referencing the Translator Schema to determine which translator is assigned to this assignment.

c) **[2 marks]** For each of the relation schemas, decide which attributes must be declared as not null. Explain your answer.

Answer:

TranslationAgency = Agency Number must be not null as it is the primary key and is used to identify a record in the schema,

Translator. = IRDNumber must be not null since according to the constraints, the IRD number must be specified.

IsExpert = Both the Name and the Language must be not null because they are the primary key. If there is one of these attributes missing in the record, it wont make sense who is an expert in what language.

TranslationOrder. = The OrderNumber must be not null because it is the primary key.

Assignment = the OrderNumber and Part must be not null because they are the primary key.

d) **[5 marks]** Assume, the translator with name 'Peter Pan' in the Translator relation retires. When deleting the record of this translator from the Translator relation, all the assignments made to him **should not** be lost. How would you ensure this requirement? Explain your answer.

Answer:

There are three different options. The first is to reject the deletion, however in this case it is not applicable because if a translator retires then he must be deleted from the database because he will no longer be able to work.

The second option is to cascade, which will delete tuples in other relations that refer to Peter Pan. This is not an ideal solution because all the assignments made to him will be deleted.

The third option is to insert a null or a default value in the foreign key attributes of tuples in other relations that refer to Peter Pan. I think this is the best option in this case because if this translator Peter Pan is deleted, then set the foreign key attributes to null. This will ensure that the assignments are kept intact so if he retired in the middle of an assignment, the agency can see that the assignment he was working on now has a null value for the IRDNumber attribute in the assignment schema and so can be reassigned to another translator.

e) **[5 marks]** Assume, a translation order with order number '42' in the TranslationOrder relation is cancelled. Suppose, however, that already some assignments have been made to translate parts of this translation order. When deleting the record of the translation order from the TranslationOrder relation, then all the assignments should be deleted, too. How can this requirement be ensured? Explain your answer.

Answer:

In this case, the best way to ensure this requirement is met, is to use the Cascade option. This will delete all the tuples in other relations that refer to this order number. So if the order number 42 tuple is deleted from the TranslationOrder schema, all the tuples in the assignment schema with the order number 42 will be deleted.

Question 4 [30 marks]

You are asked to design a new database for your grandma's collection of books. A book has a title, a release year, a unique international standard book number, a number of pages and was published by a certain publishing house. For each book, your grandma only bought one copy.

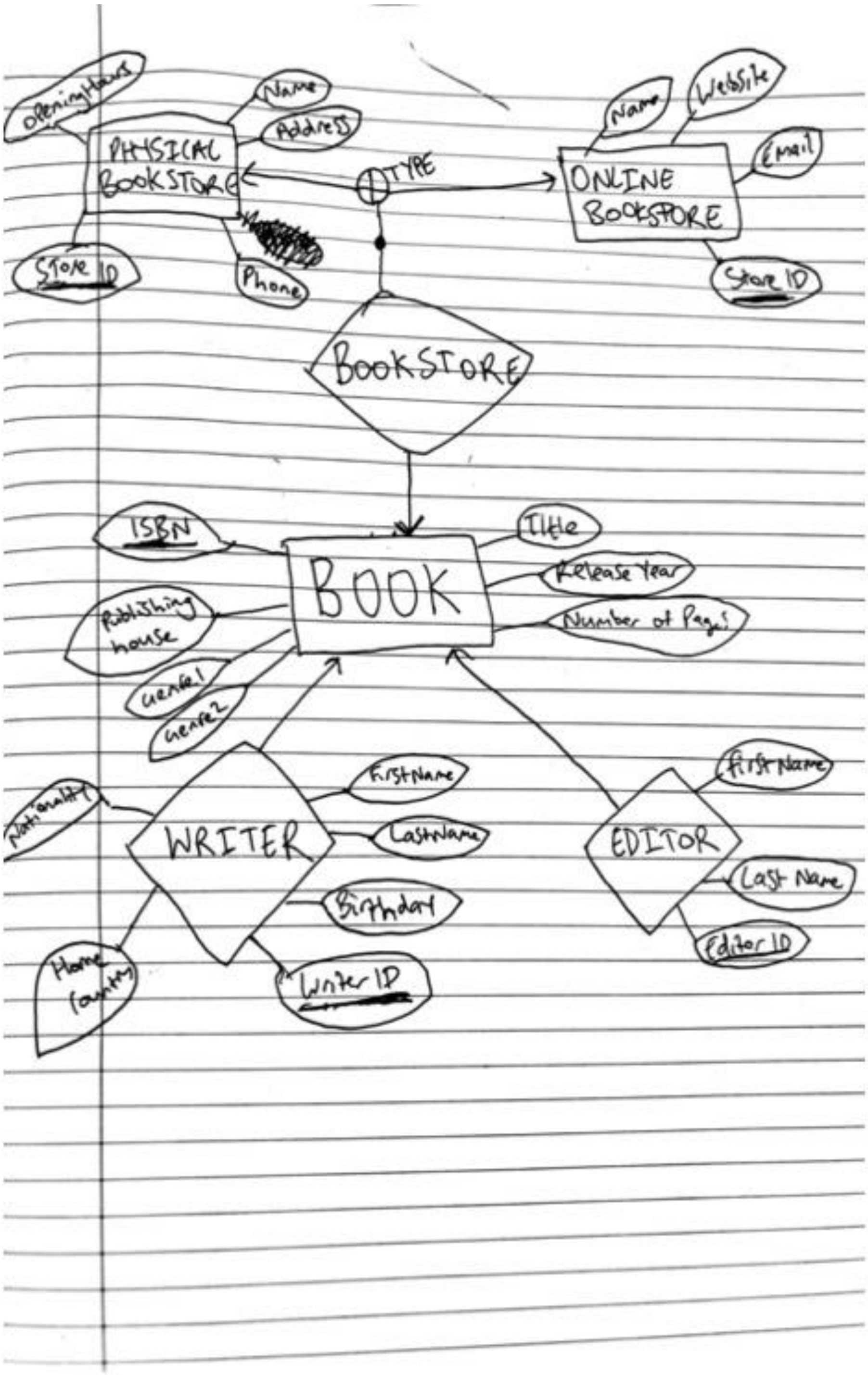
A book can have one or more authors. The authors of a book are writers. A writer has a first name, a last name, a birthday, a nationality, and a home country. A book can have one or more editors. An editor has a first name, a last name. Editors oversee the emergence of a book from the first manuscript to the print-ready form. A book without authors has at least one editor.

Furthermore, your grandma buys books at certain bookstores which are either physical ones or online ones. Physical bookstores have a name, an address, a contact phone number and opening hours, while online bookstores have a name, a website and a contact email address. There are different genres such as adventure, comedy, crime, mystery, fantasy or science fiction. For every book at most two genres should be recorded in the database.

a) [24 marks] Draw an extended ER diagram for the database above. Write down the corresponding extended ER schema, including declarations of all the entity types (showing attributes and keys) and relationship types (showing components, attributes and keys).

Answer:

Below is my ER diagram for the database above:



Scanned with CamScanner

Below is the ER schema

Entity Types:

- **Book**
  - **Attributes:** Title, Release Year, ISBN, number of pages, publishing house, Genre1, Genre2
  - **Primary Key:** ISBN
- **Physical BookStore**
  - **Attributes:** Name, Address, Phone, Opening Hours, Store ID
  - **Primary Key:** Store ID
- **Online BookStore**
  - **Attributes:** Name, Website, Email, Store ID
  - **Primary Key:** Store ID

Relationship Types

- **Writer**
  - **Components:** Book
  - **Attributes:** First Name, Last Name, Birthday, Nationality, Home Country, Writer ID
  - **Primary Key:** Writer ID
- **Editor**
  - **Components:** Book
  - **Attributes:** First Name, Last Name, Editor ID
  - **Primary Key:** Editor ID
- **BookStore**
  - **Components: Book**
  - **Attributes:**
  - **Primary Key:**

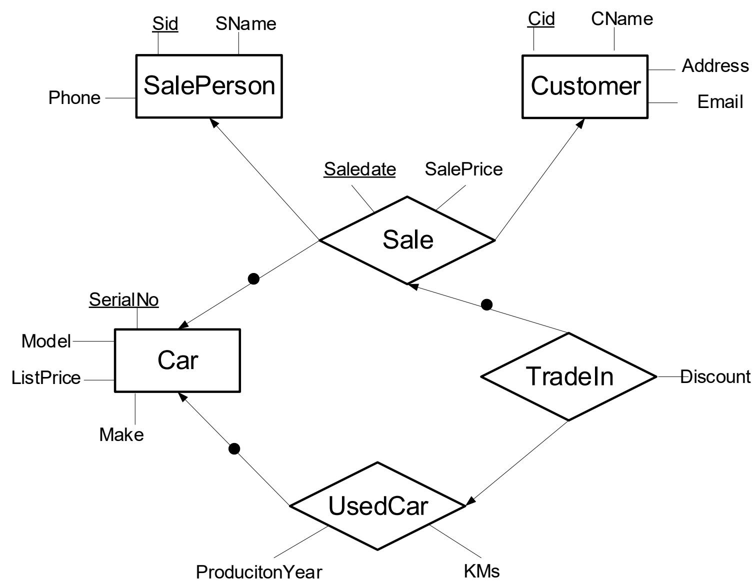
b) **[6 marks]** There may be information, requirements or integrity constraints that you are not able to represent in your diagram. Give three examples of integrity constraints that have not been represented in your diagram.

Remark**:** Whenever you feel that information is missing in the problem description above, add an assumption and make your assumption explicit. In practice you would consult the domain experts or potential users for clarification.

Answer:

Question 5 [20 marks]

Consider the following extended ER diagram:



a) [5 marks] Present the extended ER schema of the extended ER diagram above.

Answer:

Entity Types:
- **SalesPerson**
  - **Attributes:** Sid, SName, Phone
  - **Primary Key:** Sid
- **Customer**
  - **Attributes:** Cid, CName, Address, Email
  - **Primary Key:** Cid
- **Car**
  - **Attributes:** SerialNo, Model, ListPrice, Make
  - **Primary Key:** SerialNo

Relationship Types
- **Sales**
  - **Components:** SalesPerson, Customer, Car
  - **Attributes:** Saledate, SalePrice
  - **Primary Key:** Saledate, Car
- **TradeIn**
  - **Components:** UsedCar, Sale
  - **Attributes:** Discount
  - **Primary Key:** Sale
- **UsedCar**
  - **Components:** Car, TradeIn
  - **Attributes:** ProductionYear, KMs
  - **Primary Key:** Car

b) [10 marks] Transform your extended ER schema into a relational database schema. In particular, list all the relation schemas in your relational database schema. For each relation schema, list all attributes, the primary key, the NOT NULL constraints, and the foreign keys.

Answer:

- **SalesPerson**
  - **Attributes:** Sid (NOT NULL), SName, Phone
  - **Primary Key:** Sid
- **Customer**
  - **Attributes:** Cid (NOT NULL), CName, Address, Email
  - **Primary Key:** Cid
- **Car**
  - **Attributes:** SerialNo (NOT NULL), Model, ListPrice, Make
  - **Primary Key:** SerialNo
- **Sales**
  - **Attributes:** Saledate(NOT NULL), SalePrice, Cid(NOT NULL), Sid(NOT NULL), SerialNo (NOT NULL)
  - **Primary Key:** Saledate, Cid, Sid, SerialNo
  - **Foreign Key:** [Cid] ⊆ Customer[Cid], [Sid] ⊆ SalePerson[Sid], [SerialNo] ⊆ Car[SerialNo]
  -
- **TradeIn**
  - **Attributes:** Discount, Saledate (NOT NULL),
  - **Primary Key:** Saledate,
  - **Foreign Key:** [Saledate] ⊆ Sales[Saledate],
- **UsedCar**
  - **Attributes:** ProductionYear, KMs, SerialNo(NOT NULL)
  - **Primary Key:** SerialNo
  - **Foreign Key:** [SerialNo] ⊆ Car[SerialNo],

c) [5 marks] We also want to record information about expenses related to services for all the cars sold by the company. Each related expense has a date, a cost, and a description. Enhance the given extended ER diagram to reflect this additional information. Present the extended ER diagram with your proposed enhancements. Justify your answer.

Answer:

****