CS41DS Database Systems

Kenny Kei Yun Sum

Student ID : 230215740

Database Coursework

Science Archives

Requirement 1

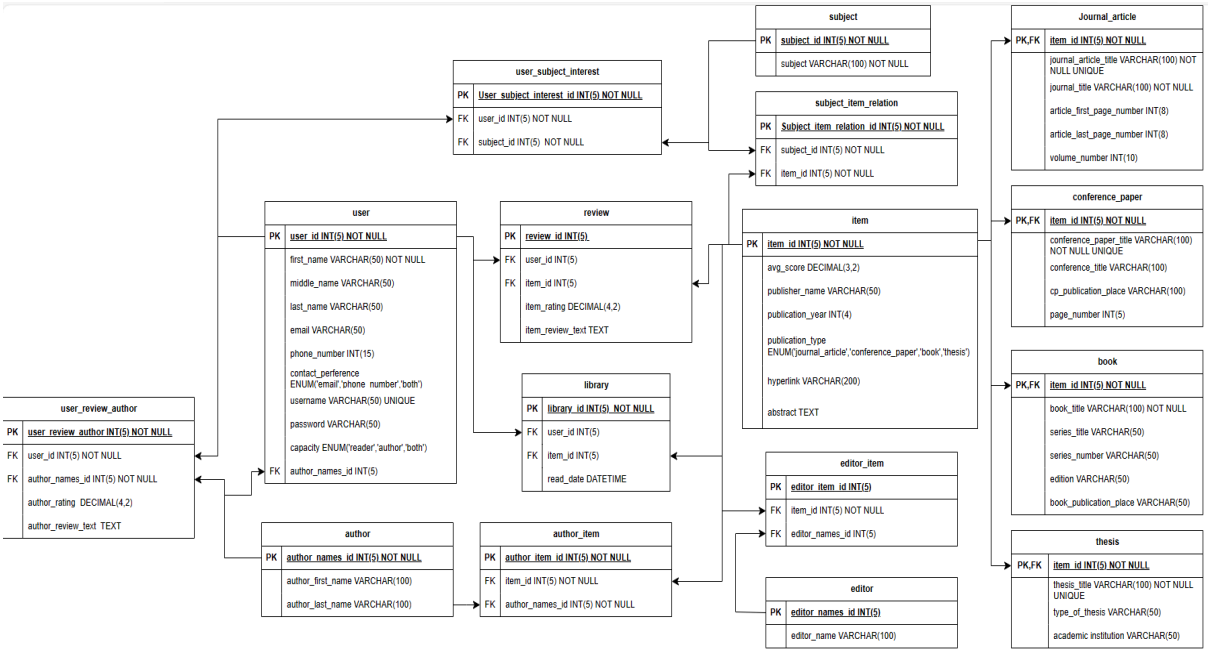


Figure 1 - Science Archives Relational Schema

The design decisions leading to the creation of the Relation Schema.

The relational schema of Science Archives shown in figure 1, represents the relationship between its main entities, which includes the user, items, journal articles, books, conference paper, thesis, author, subject and editor. Each entity contains attributes which further reveal information and its complexity. These entities are connected by various types of relationships such as one-to-one, one-to-many,many-to-one and many-to-many. This section will discuss each entity in more detail, the relationships between them and how it relates to the client specification.

Entities

User : The user entity contains user's id, full name (first name, middle name and last name), email, phone, contact preference, username, password, capacity and author name. The user id acts as a primary key and the user's full name establishes the user's identity. The email and phone attributes contain information on how the user can be contacted, as well as their contact preference. A username and password is also contained within the entity and their capacity, which identifies if the user is a reader, author or both. The user entity is also connected to the author entity, establishing a one-to-one relationship if the user is one of the authors that has written an item (conference paper, thesis, book, article). These attributes satisfy the client's specification.

Items : The item entity contains attribute item id, avg score, publisher name, publication year, publication type, hyperlink and abstract. The item id acts as a primary key. The column avg score shows the average score which rates the paper based percentage. The publisher name shows the name of the individual who published the paper and the column publication

year shows the year in which the item has been published. The entity also contains the column called publication_type which reveals the type of publication of the item. This can range between journal articles, books, thesis or conference papers.

Due to the various types of publication, four entities have been created which connect to the item entity. These entities are thesis, book, conference paper and journal article. When an item is added onto the database, information is also inputted into one of these tables which establishes their type of publication. Each publication type has different attributes, so various entities were created to suit each one. For example, a conference paper is related to the name of the conference but a book is not involved in a conference but instead contains a book publication place.

An item can be stored within the entity item and the entity journal article if the item itself is an article from a journal, both containing the same item id and are therefore connected. The same can be done if the item were a different publication type, such as if the item were a book or a thesis. This fulfils the requirement of the database being capable of containing items of different types of publication based on the client specification.

Other entities : The author entity contains various authors who were responsible for creating the item, the subject entity holds numerous subjects that the user is interested in and the item is related to and the editor entity accommodates several editors responsible for editing the item.

Relationship between entities

Entities have relationships between themselves which shows how they interact. Relationships can range from one-to-one, one-to-many, many-to-one and many-to-many. This section will highlight the relationships between selected entities and why the type of relationship was chosen between the two shown on the diagram (figure 1).

Relationship between user and subject : According to the client specification, users will have a scientific interest within a certain subject or multiple subjects. An entity based on user and subject is created and a many-to-many relationship is established, as shown on the diagram (figure 1). This is due to the fact that the database can contain multiple users and subjects. In addition, each user can be interested in multiple subjects and each subject can be interested by multiple users. As a result, a table known as user_subject_interest is created in order to establish the many-to-many relationship between the two entities.

Relationship between user and item : The diagram (figure 1) shows that each user is able to review and read each item. As the database can contain many items, each can be read and reviewed by many users. A review and library entity is therefore created to connect the user and item entity together, establishing a many-to-many relationship. The attributes within the review entity allows users to rate and write a review on the items that they have read and the library attribute keeps track of these items and the date in which the item is read.

Relationship between user and subject : A relationship between user and subject is established to connect the user to a single or various subjects that they are interested in. In addition, a subject can intrigue many users, creating a many-to-many relationship.

Relationship between item and subject : Each item in the item entity is related to a or several subjects within the subject entity. An entity called item_subject_relation is created to show the relationship between the two entities, which is a many-to-many relationship. Since an item can be related to many or one subject and a subject can be connected to many items.

Relationship between item and author : Every item contains an author or multiple authors. In addition, an author can write many items and an item can be written by many authors. A many-to-many relationship is established because of this.

Relationship between item and editor : Every item can have an editor or many editors. Furthermore, an editor can edit many items, while an item can have many editors. This shows a many-to-many relationship between the two entities.

Each entity and their attributes were determined by the information acquired from the client's specification in order to achieve the system's functionality. The database and its relational schema was examined and improved by using the process of normalisation. This is done to eliminate redundant data, reduce data modification errors, and to simplify the query process. In addition, various data types and constraints were used for each attribute to promote data integrity and assist in queries. For example, the data type for INT was used for entity ID to ensure the cell contains only numbers. While DECIMAL data type is used to show the percentage of the average score for an item. The primary key was used to identify each record within the table, while the foreign key was used to establish the relationship between each table. In conclusion, the relational schema ensures that the database fulfils the requirement for the clients specification.

Database SQL commands and dummy data

The SQL commands used to create the database and its dummy data are shown within the SQL file.

Constraints type	Purpose of constraint	How to test constraint	SQL statement	Results	Conclusion
Primary key	The purpose of a primary key is to serve as a unique identifier for each record and to establish relationships between tables. It is used to prevent duplicate data and used to effectively retrieve specific data.	To ensure that the primary key is functional, a test scenario has been done to ensure its integrity. An SQL statement with a primary key value of 1 has been inserted into the table nuser, which already has primary key value of 1.	INSERT INTO `nuser`(`user_id`, `first_name`, `middle_name`, `last_name`, `email`, `phone_no`, `contact_preference`, `username`, `password`) VALUES ('1','Ben','Long','Green','Bengreen66@gmail.com','08534234739','phone','longgg56','password555')	#1062 – Duplicate entry '1' for key 'PRIMARY'	the database was unable to insert the data into the database, giving an error since the primary key column already has the value of 1.
Foreign key	The foreign key is to establish and enforce a link between the tables in a database. The foreign key ensures data integrity by enforcing referential integrity constraint.	To test the foreign key integrity constraint, a SQL statement has been used to delete the row that contains the item_id equal to 1. The item_id attribute in the item table is connected to item_id in the editor_item table through a foreign key.	DELETE FROM item WHERE item_id = 1	#1451 – Cannot delete or update a parent row: a foreign key constraint fails (`science archives`.`editor_item`, CONSTRAINT `editor_item_ibfk_1` FOREIGN KEY (`item_id`) REFERENCES `item` (`item_id`))	The database was unable to delete the row, since a foreign has been established on the item id 1.
Not null	The not null constraint ensures that the cell must contain a value and cannot be null.	In order to examine its functionality. a SQL insert statement has been used on the table subject which contains the NOT NULL constraint on the column subject.	INSERT INTO `subject`(`subject_id`, `subject`) VALUES ('7',NULL)	#1048 – Column 'subject' cannot be null	An error has occurred as a result. The database does not allow a null value within the cell in the column subject.
Unique	The unique constraint ensures that all values within a column are different.	The table nuser contains a unique column, username. In order to test its functionality, a SQL insert statement has been inserted into the table nuser. This insert statement will contain a username value with the same value already in the table.	INSERT INTO `nuser`(`user_id`, `first_name`, `middle_name`, `last_name`, `email`, `phone_no`, `contact_preference`, `username`, `password`) VALUES ('7','Paul','Stone','Gilmore','stone99@gmail.com','07869123567','email','stone99','rock678');	#1062 – Duplicate entry 'stone99' for key 'username	As a result, the database was unable to allow the insertion, due to the unique constraint. An error was shown.
Check	The check constraint is used to limit the value range that is placed within a column. The table journal_article contains an attribute called article_first_page_number. A check value is used for this attribute so the number does not become higher the value of another attribute called article_last page_number within the same row.	An insertion SQL statement is created to test the check constraint. This is done by inserting an article_first_page_number number value higher than the article_last_page_number number value.	INSERT INTO `journal_article`(`item_id`, `journal_article_title`, `journal_title`, `article_first_page_number`, `article_last_page_number`, `volume_number`,`journal_article_abtract`) VALUES ('14','title','200','100','1','abstract');	integer out of range	Database was unable to allow insertion, due to a check constraint. The value was not within range.

Figure 2 - Table containing Integrity constraints

SQL statements used to determine constraints effectiveness are shown within the SQL file.

Requirement 2

Business Processes

The user has multiple options of interacting with the system. A list of business processes is shown below, showing numerous ways of how users can interact with the system and view certain aspects of the database.

- Users can search for items based on publication type (thesis, conference paper, articles and book) which is shown on figure 3.
- Users can view more detailed information on items (thesis, conference paper, articles and book), such as year of publication, average score, hyperlink on item, item abstract, associated subject and publication type which is shown on figure 4.
- Users can choose to view only certain information by filter.
- Users can see items based on their scientific interest which is shown on figure 5.
- Users can store items already read and input a review or score which is shown on figure 6.
- Users can view items and their score rating. Score rating is revealed through descending which is shown in figure 7.
- Users can view the author ratings of authors in descending order, done by other users. In addition, the table shows the names of users who rated and reviewed authors. This is shown in figure 8.

thesis_title	type_of_thesis	Publication_type	hyperlink	publication_year
The implementation of artificial intelligence and ...	Research thesis	thesis	https://era.ed.ac.uk/handle/1842/4757	2015
Developmental pathways of suicidality and self-har...	Analytical thesis	thesis	https://era.ed.ac.uk/handle/1842/45656	2009
Discovery of genetic factors for reading ability a...	Research thesis	thesis	https://era.ed.ac.uk/handle/1842/46767	2019

Figure 3 - Shows list of thesis available on database

thesis_title	type_of_thesis	academic_institution	avg_score	publisher_name	publication_year	Publication_type	hyperlink
The implementation of artificial intelligence and ...	Research thesis	University of Nottingham	13.50	Economy of Power	2015	thesis	https://era.ed.ac.uk/handle/1842/4757
Developmental pathways of suicidality and self-har...	Analytical thesis	Harvard	65.50	Fallguy society	2009	thesis	https://era.ed.ac.uk/handle/1842/45656
Discovery of genetic factors for reading ability a...	Research thesis	University of Edinburgh	34.50	Constellation	2019	thesis	https://era.ed.ac.uk/handle/1842/46767

Figure 4 - Shows list of thesis with additional information

first_name	middle_name	last_name	subject	avg_score	publication_year	hyperlink	journal_article_title
Lin	Kin	Wong	artificial intelligence	95.50	2023	https://www.scielosp.org/pdf/bwho/v82n11/v82n11a09...	Causability and explainability of artificial intel...
Joshua	Arm	Powers	artificial intelligence	95.50	2023	https://www.scielosp.org/pdf/bwho/v82n11/v82n11a09...	Causability and explainability of artificial intel...
Paul	Stone	Gilmore	visual impairment	50.45	2017	https://link.springer.com/article/10.1007/s12193-0...	General data on the visually impaired
Paul	Stone	Gilmore	medicine	45.90	2020	https://link.springer.com/article/10.1007/s10209-0...	Various medicine for the visually impaired
Lydia	Maria	Rockwell	medicine	45.90	2020	https://link.springer.com/article/10.1007/s10209-0...	Various medicine for the visually impaired
Mathew	Kin	water	medicine	45.90	2020	https://link.springer.com/article/10.1007/s10209-0...	Various medicine for the visually impaired
Lin	Kin	Wong	artificial intelligence	45.90	2020	https://link.springer.com/article/10.1007/s10209-0...	Various medicine for the visually impaired
Joshua	Arm	Powers	artificial intelligence	45.90	2020	https://link.springer.com/article/10.1007/s10209-0...	Various medicine for the visually impaired

Figure 5 - Recommending Journal articles to users based on their scientific interest.

first_name	middle_name	last_name	avg_score	Publication_type	thesis_title	hyperlink	read_date
Lydia	Maria	Rockwell	65.50	thesis	Developmental pathways of suicidality and self-har...	https://era.ed.ac.uk/handle/1842/45656	2019-05-23 00:00:00
Paul	Stone	Gilmore	34.50	thesis	Discovery of genetic factors for reading ability a...	https://era.ed.ac.uk/handle/1842/46767	2021-08-23 00:00:00
Mathew	Kin	water	34.50	thesis	Discovery of genetic factors for reading ability a...	https://era.ed.ac.uk/handle/1842/46767	2021-02-10 00:00:00
Lin	Kin	Wong	13.50	thesis	The implementation of artificial intelligence and ...	https://era.ed.ac.uk/handle/1842/4757	2020-01-10 00:00:00
Joshua	Arm	Powers	13.50	thesis	The implementation of artificial intelligence and ...	https://era.ed.ac.uk/handle/1842/4757	2020-12-10 00:00:00
George	May	Wall	13.50	thesis	The implementation of artificial intelligence and ...	https://era.ed.ac.uk/handle/1842/4757	2022-05-10 00:00:00

Figure 6 - Listing thesis that the user has already read, including hyperlink and review/score.

conference_paper_title	avg_score ▾ 1	hyperlink
Identifying people with cancer	95.50	https://era.ed.ac.uk/handle/1842/45656
Bayesian Gaussian Processes for Identifying the De...	67.50	https://era.ed.ac.uk/handle/1842/767
Youth understanding of online privacy and security	60.45	https://era.ed.ac.uk/handle/1842/41396

Figure 7 - Showing the score rating of conference papers in ascending order.

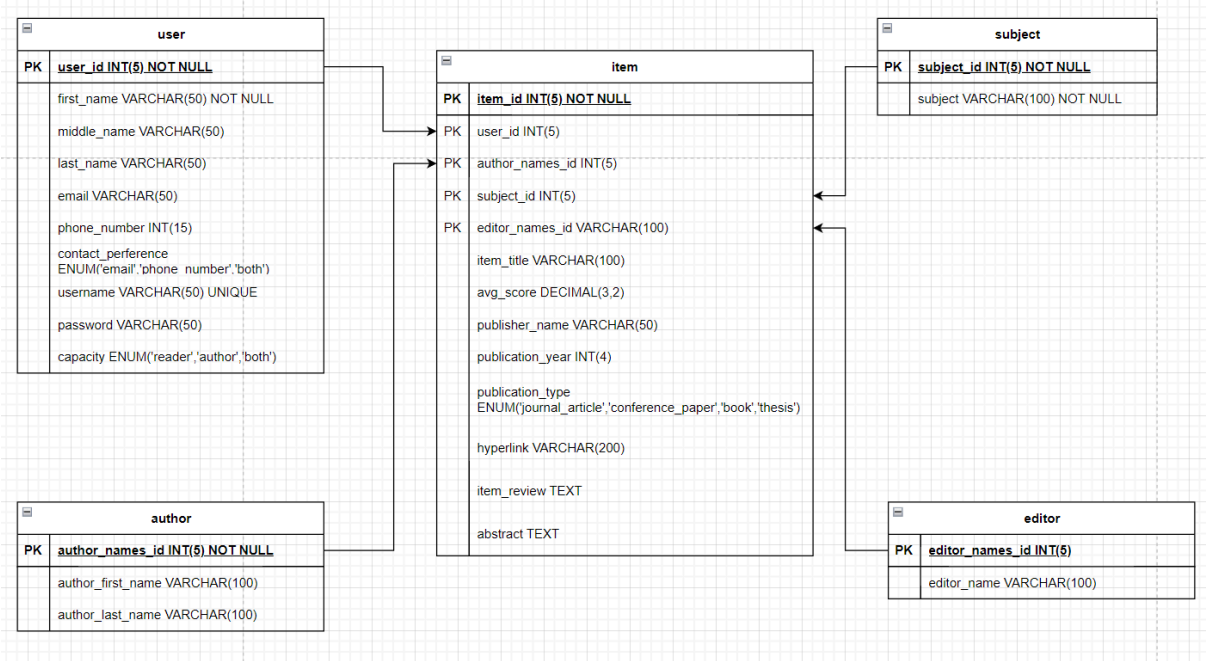
first_name	middle_name	last_name	author_rating ▾ 1	author_review_text	author_first_name	author_last_name
Mathew	Kin	water	90.00	almost perfect	Shekhar	Saxena
Lin	Kin	Wong	50.00	average	Akif	Khan
Paul	Stone	Gilmore	13.00	not good	Vitor	Filipe
Paul	Stone	Gilmore	12.00	bad	Hugo	Fernandes
Paul	Stone	Gilmore	12.00	not great	Akif	Khan
Lydia	Maria	Rockwell	11.00	awful	Vitor	Filipe

Figure 8 - Shows author rating in descending order. In addition, the table shows the names of users who rated and reviewed authors.

Requirement 3

Converting the relational schema into Kimball's Dimensional Model

The kimball's Dimensional model allows users to construct several star schemas to achieve numerous reporting needs. There are numerous benefits of using Kimball's Dimensional Model instead of using the proposed relational schema. These advantages include, its simplicity to design, general being faster to design and easier to understand. However, the model generally has its issues. This being that it is normally vulnerable to data redundancy, less flexibility and does not offer a complete view of the business processes of the database.



Kimball's Dimensional Model of the relational schema (figure 9).

The diagram figure 5 shows the kimball's dimensional model of the relational schema. This section will discuss the advantages and disadvantages of using this model in more detail.

Advantages

The kimball's Dimensional Model represents a less time consuming approach of designing a schema. Since the model is generally simpler due to having less entities
which can be seen as you compare the two models (figure 1, figure 9) In addition, the relationships between entities are less complex as most of the entities are connected to a single table. This reduces the time needed to understand relationships between entities and offers a simpler view of how the business process of the database functions. As a result, Kimball's Dimensional Model offers an abstractive view of the model, which benefits individuals who do not have a complex understanding of traditional relational schemas.

Disadvantages

Kimball's Dimensional Model presents numerous disadvantages as well. The effort to denormalise the data into Kimball's Dimensional format can be a time consuming process. In addition, the model is vulnerable to duplicate data, as it can be used across several data marts. As a result, this may introduce conflicting data compared to the traditional relational schema, leading to confusion and increased storage space. An example of this is if two items with different title names containing the same user were inserted into the database. The name attribute will show duplicate data.