



Linnéuniversitetet
Kalmar Väst

Linnaeus University
2DV513 - Database Theory

Assignment 2



Author: Rashed QAZIZADA
Area: Computer Science
Supervisor: Ilir Jusufi
Course code: 2DV513
Semester: Autumn 2020

Abstract

This page was intentionally left unchanged.

Keywords

Place your keywords here

Contents

1	Task : Relational algebra	1
2	Task : FDs and Normalization	1
2.1	Finding functional dependencies.	1
2.2	Finding the keys of the relation	1
2.3	Showing that the relation is in 3NF but not in BCNF	1
2.4	Decomposing the relation in relations that are in BCNF.	2
2.5	E/R diagram that describes the system with incorporation of all dependencies	2
3	Task : Setting up the Reddit database	3
3.1	An E/R diagram for my design as well as schemeas with their types . . .	3
4	Task: Importing data	4
4.1	Conclusion	4
5	Task: Queries	5

1 Task : Relational algebra

1.1. $\Pi_{\text{name}}(\sigma \text{ code} = '2dv513' (\text{student} \bowtie \text{enrolledIn}))$

1.2. $\Pi_{\text{name}}(\sigma \text{ code} = '1dv513' \cap \text{code} '2dv513' (\text{student} \bowtie \text{enrolledIn}))$

1.3. $\Pi_{\text{lecturer}}(\sigma \text{ code} = '2dv610' (\text{enrolledIn} \bowtie \text{subject}))$

1.4. $\Pi_{\text{lecturer}}(\sigma \text{ code} = '1dv513' \cap \text{code} '2dv513' (\text{enrolledIn} \bowtie \text{subject}))$

1.5. $\Pi_{\text{names}}(\sigma \text{ lecturer} \neq 'Iilr' (\text{student} \bowtie \text{subject}))$

2 Task : FDs and Normalization

2.1 Finding functional dependencies.

I can assume that this database will check every week. Therefore each week is a unique.

Day time \mapsto applicant manager room.

Applicant day \mapsto time room manager.

2.2 Finding the keys of the relation

Day and time are to be the keys of the relation.

2.3 Showing that the relation is in 3NF but not in BCNF

The relation should not have partial dependency neither transitive dependency. This means the relation will be in 3NF if it satisfies 2NF.

- Checking for partial dependency:

Since we have day and time as the composite keys. Therefore, I cannot know who is the applicant only from knowing the day value or the time value. Furthermore, the same rule applies to the manager, and still, I cannot know who is having the time or the day. Therefore, there is no partial dependency. So, it is in 2NF.

- Now, let's check for transitive dependency:

From the transitive dependency perspective, I could not determine who is the applicant only from knowing the manager or the room. Therefore, there is no transitive dependency. This means the relation has 3NF normalization.

- Now, showing that it does not satisfy BCNF.

Since, we have a non-prime attribute manager which determines a prime attribute which is the time then this relation is not in BCNF. For this I assume that the Manager can have just one meeting for each day.

2.4 Decomposing the relation in relations that are in BCNF.

In order to compose the relation to make relations in BCNF, the primary key of the relations needs to be changed. So, the relations will be as following:

People (M_ID, Manager, applicant)

date (day, time, M_ID, room)

Now we can see that the above relations satisfy 3NF normalization, and it does not have a non-prime attribute which determines a prime attribute, Therefore, it satisfies BCNF as well.

2.5 E/R diagram that describes the system with incorporation of all dependencies

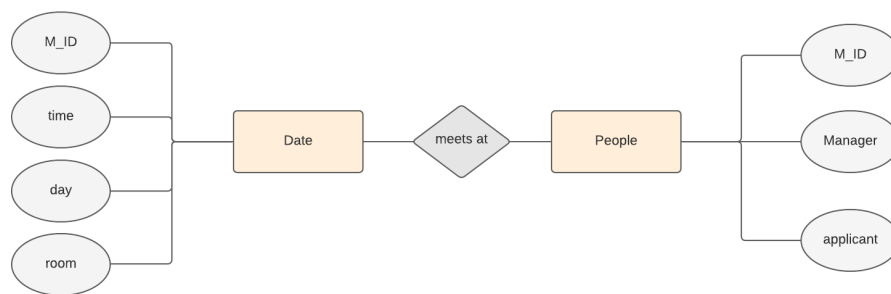


Figure 1: E/R diagram

3 Task : Setting up the Reddit database

3.1 An E/R diagram for my design as well as schemeas with their types

The E/R diagram shows the relations between a reddit comment with a subreddit and links.

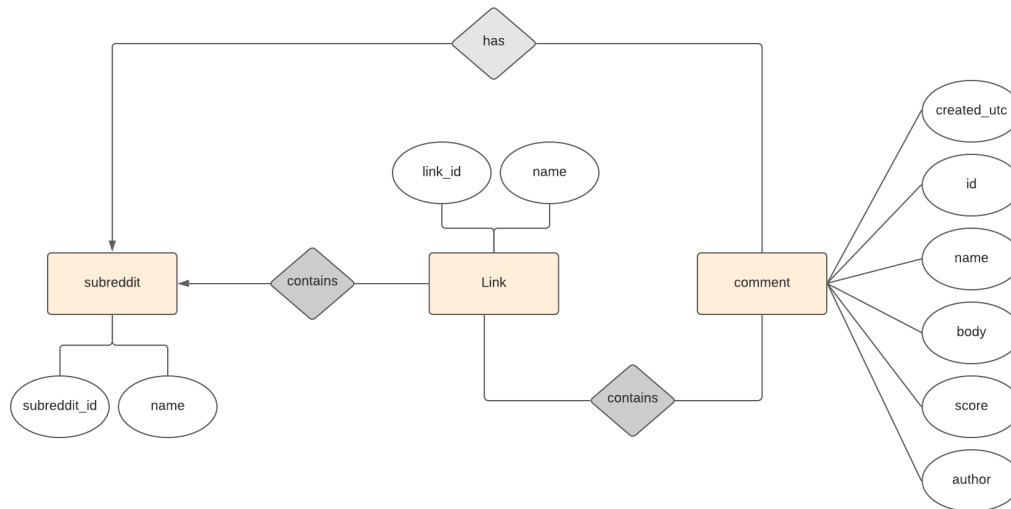


Figure 2: E/R diagram with their types

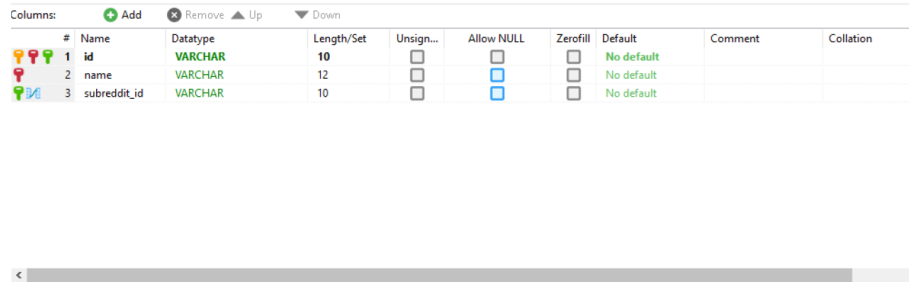
Schemas with types:

Comment: has subreddit_id and link_id as foreign keys

Columns: + Add - Remove ▲ Up ▼ Down										
#	Name	Datatype	Length/Set	Unsign...	Allow NULL	Zero fill	Default	Comment	Collation	Expre
1	id	VARCHAR	10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			
2	name	VARCHAR	45	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			
3	author	VARCHAR	45	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			
4	score	INT	11	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			
5	body	VARCHAR	10000	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			
6	parent_id	VARCHAR	20	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			
7	created_utc	TIMESTAMP	6	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			
8	subreddit_id	VARCHAR	10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			
9	link_id	VARCHAR	10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			

Figure 3: subreddit_id and link_id

Link: has subreddit_id as foreign keys.

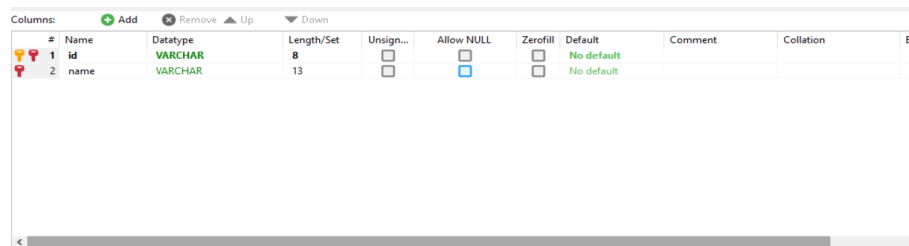


The screenshot shows a table definition interface with columns: #, Name, Datatype, Length/Set, Unsigned, Allow NULL, Zerotill, Default, Comment, and Collation. The table has three columns: 1. id (VARCHAR, 10, Unsigned, Allow NULL, Zerotill, Default: No default), 2. name (VARCHAR, 12, Unsigned, Allow NULL, Zerotill, Default: No default), and 3. subreddit_id (VARCHAR, 10, Unsigned, Allow NULL, Zerotill, Default: No default).

#	Name	Datatype	Length/Set	Unsigned	Allow NULL	Zerotill	Default	Comment	Collation
1	id	VARCHAR	10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default		
2	name	VARCHAR	12	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default		
3	subreddit_id	VARCHAR	10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default		

Figure 4: subreddit_id

Subreddit:



The screenshot shows a table definition interface with columns: #, Name, Datatype, Length/Set, Unsigned, Allow NULL, Zerotill, Default, Comment, Collation, and Exp. The table has two columns: 1. id (VARCHAR, 8, Unsigned, Allow NULL, Zerotill, Default: No default) and 2. name (VARCHAR, 13, Unsigned, Allow NULL, Zerotill, Default: No default).

#	Name	Datatype	Length/Set	Unsigned	Allow NULL	Zerotill	Default	Comment	Collation	Exp
1	id	VARCHAR	8	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			
2	name	VARCHAR	13	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No default			

4 Task: Importing data

- First, I populated tables with constraints the process almost ended in 29 minutes just to migrate the entire data to the database.

```
150428 errorrrrrrr Subreddit ER_DUP_ENTRY Duplicate entry 't5_2cneq' for key 'PRIMARY'
150428 errorrrrrrr Link ER_DUP_ENTRY Duplicate entry '5zep2' for key 'PRIMARY'
150428 added Comment
150429 errorrrrrrr Subreddit ER_DUP_ENTRY Duplicate entry 't5_6' for key 'PRIMARY'
150429 errorrrrrrr Link ER_DUP_ENTRY Duplicate entry '5zik0' for key 'PRIMARY'
150429 added Comment
```

- Second, I tested without constraints which means no primary keys nor unique or foreign keys were assigned and the process took almost 22 minutes just to get finished. Furthermore, the table was full of duplicates data.

4.1 Conclusion

Comparing the results of populating the data with and without constraints, I can see that the insertion without constraints is faster. However, to migrate the data without constraints I achieved faster results on the other hand we should also consider the duplication of the data, Furthermore, we should avoid adding constraints after importing it or I could choose some rows even if I have duplication with data just by using some specific keyword. However, it is a good approach to use constraints before populating or migrating the data to the database.

5 Task: Queries

How many comments have a specific user posted?

In order to get the number of comments for a specific user and whereas the author is a specific username then I could achieve this just by `SELECT COUNT(ID) FROM comment WHERE AUTHOR = 'igiveyoumylife'` the result of this specific user that posted is 38.

```
1 SELECT COUNT(ID) FROM comment WHERE AUTHOR = 'igiveyoumylife'
```

Result #1 (1r x 1c)

COUNT(ID)
38

How many comments does a specific subreddit get per day?

I selected the “bugs” subreddit for each day, in total a specific subreddit get 104 comments per day.

```
1 select COUNT(Id) from comment WHERE subreddit_id = 't5_2h84o' GROUP BY created_utc
```

Result #1 (104r x 1c)

COUNT(Id)
1
1
1
1
1
1

How many comments include the word 'lol'?

In this case "SELECT COUNT(*) FROM comments WHERE body LIKE '%lol%' " I am counting the comments where the string 'lol' is contained in that string. The percent signs before and after will ensure that any string that contains lol in it counts it like as following: lollipop, lolypop, lolblolly etc.

```
1 SELECT COUNT(*) FROM comment WHERE body LIKE '%lol%'
```

Result #1 (1r x 1c)	
COUNT(*)	759

Users that commented on a specific link has also posted to which subreddits?

I am using the link to get the author name then I selected all subreddit_id from the same author.. In this case the query will show all the data that fulfils my condition.

```
1 SELECT NAME FROM subreddit WHERE id IN (
2 SELECT subreddit_id FROM comment WHERE author IN (
3 SELECT author FROM comment WHERE link_id = '5zk06')
)
```

subreddit (24r x 1c)	
NAME	
ads	
bugs	
business	
de	
entertainment	
features	
fr	
freeculture	
gadgets	
gaming	
id	
it	
ja	

Which users have the highest and lowest combined scores? (combined as the sum of all scores)

```
CREATE VIEW maxScore AS SELECT author, MIN(score)MinSS,SUM(score) MaxSS
FROM comment GROUP BY author; select author from maxscore where MaxSS =
(select max(MaxSS) from maxscore); select author from maxscore where MinSS =
(select min(MinSS) from maxscore)
```

```
1 CREATE VIEW maxScore AS
2 SELECT author,MIN(score)MinSS,SUM(score) MaxSS FROM comment GROUP BY author;
3 select author from maxscore
4 where MaxSS = (select max(MaxSS) from maxscore);
5 select author from maxscore
6 where MinSS = (select min(MinSS) from maxscore)
```

comment (1x 1c)


comment (1x 1c)

author
[deleted]

Given a specific user, list all the users he or she has potentially interacted with (i.e., everyone who has commented on a link that the specific user has commented on)

In this case duplicated authors are interacted. I had to be careful not to duplicate the authors.

```
SELECT author FROM comment WHERE link_id IN  
(SELECT link_id FROM comment WHERE author = 'aletoledo');
```



author
[deleted]
Kolibri
gizzledos
ProximaC
coldwarrior
[deleted]
srv
garyp714
[deleted]
[deleted]

Studen id: rq222ah

THE END

Which users has only posted to a single subreddit?

Here, I need to select all the authors who commented exactly one subreddit.

```
SELECT COUNT(subreddit_id) AS SubAmount, author  
FROM comment GROUP BY author HAVING COUNT(SubAmount) = 1
```



The screenshot shows a SQL query editor with the following query:

```
1 SELECT COUNT(subreddit_id) AS SubAmount, author  
2 FROM comment GROUP BY author HAVING COUNT(SubAmount) = 1
```

On the right side, there is a sidebar with a search bar and a list of filters:

- Filter ...
- Columns in com
- SQL functions
- SQL keywords
- Snippets
- Query history
- Query profile
- Bind paramete

Below the query editor, there is a table with the following data:

SubAmount	author
1	bostich
1	donh
1	tubelight
1	edf825
1	flosch
1	tinha
1	chiarafax
1	867uht
1	albar129
1	perri