# Lab Report 1 – Introduction to MySQL

## Databases - CSC 3320 121

### Hudson Arney

### 23 January 2024

What have you learned from the lab?

* The most important part I learned is to never forget the initial password (just kidding).
* I learned how to set up an SQL Server from scratch.
* How to create a table and import data into SQL.
* How the tables can work together to create an optimally functional database.

What caused issues?

* The biggest issue I had was with creating the primary keys for the indices in the data.
* For this I was stuck for awhile but after talking with you, you made it straightforward what I needed to do to fix the issue.
* I also was confused on the ERD diagram for a while. I did the best I could but think a bit more clarification in the weeks to come will help.

What went well or surprised you?

* Even though there were a lot of firsts in this lab it went mostly smooth (although plugging and chugging isn’t the hardest activity)
* Even though we didn’t have to come up with a lot of our own ideas for this lab, I was still able to learn quite a bit and was a good intro to using SQL for storing data.

Observations and query results for each major step:

1. Opening gene\_info1000 in excel

A screen shot of a computer

Description automatically generated

From this data there aren’t any labels for the columns which makes it a bit harder to interpret. It seems like columns I+J are descriptive. I really don’t know what any of the other columns would mean but they seem to have some sort of trend where similar numbers are used over in a specific column.

1. Table Creation

A screenshot of a computer

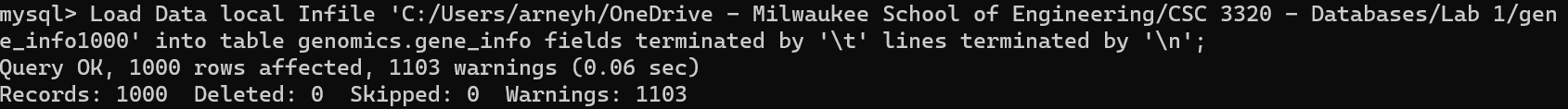
Description automatically generated

I failed before because I didn’t realize the genomics schema wasn’t selected.

A screen shot of a computer program

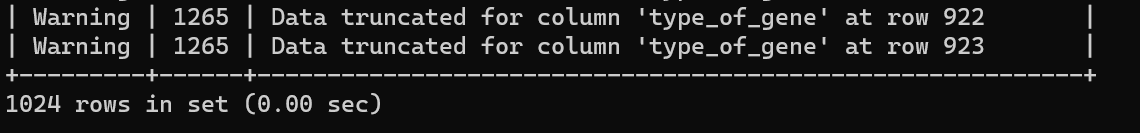
Description automatically generated

1. Read in the gene\_info1000 file by command line



Once I changed the quotes and my file path it worked first try.

1. Looking at Warnings



Quite a few warnings as to be expected from the instructions.

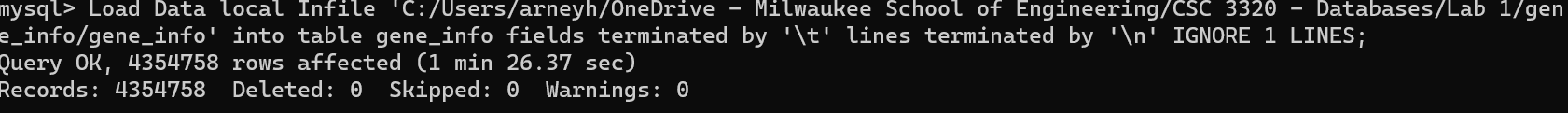
1. Create a new table with larger fields for type of gene and Other designations

A screenshot of a computer

Description automatically generated

The creation of the empty table with column headers took almost no time, it’s the insertion which will take longer.

1. Load the full dataset



Took two attempts to get the file path correct. Table insertion took a while because we had to create it from scratch with a lot of data from gene\_info (Over 400 MB).

1. Check for warnings and verify the length of the database table.

A black screen with white text

Description automatically generated

Worked first attempt, made sure I had the IGNORE 1 LINES so the first line wouldn’t be added. This takes a bit more time than after we give the table indexes in the ending steps.

1. Issue the following select statement, which limits the results to the first 100 records, and inspect the table.

A screenshot of a computer

Description automatically generated

Now it is essentially looking the same as the table in excel, except now with helpful column labels.

1. Let’s create a new table with just the information we need.

A screenshot of a computer

Description automatically generated

Had to change the timeout setting to get this to work. Table creation took a while because we had to create it from scratch with a lot of data from gene\_info (Over 400 MB).

1. Retrieve the record for a specific gene.

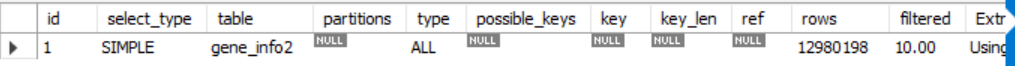
A screenshot of a computer

Description automatically generated



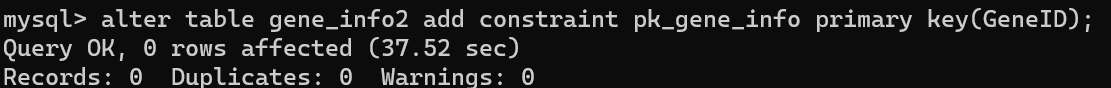
Shows some of the duplicate gene records.

1. Re-execute the command with the explain option.





Helped my understanding of the data in the table

1. Alter the gene info2 table to add a primary key.

This step was the biggest hurdle for me. Had to drop the table, remake it, and reinsert the data. Before I was getting duplicate indexes errors.

1. Run query again.

A screen shot of a computer

Description automatically generated

Easy to follow, worked first attempt. This index was very fast as it has a key now.

1. Create the table and import the data

A black screen with white text

Description automatically generated

Took quite awhile but was successful on the first attempt! I believe it took so long because it created a table from scratch (no indexes) and it had a lot of data (over 200 MBs) coming in from gene2pubmed

Final Schema from Workbench

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

ERD Diagram

* I believe gene2pubmed is the merging block to gene\_info2 and pubmed, therefore making it a relationship. However, gene2pubmed is still a table with attributes.
* I was a bit unsure of the relationships however in my head it makes sense for a publication to have one or many genes; however, a gene doesn’t not need a publication. I’m sure how to display this logic in the diagram.

A diagram of a data flow

Description automatically generated