

## Report

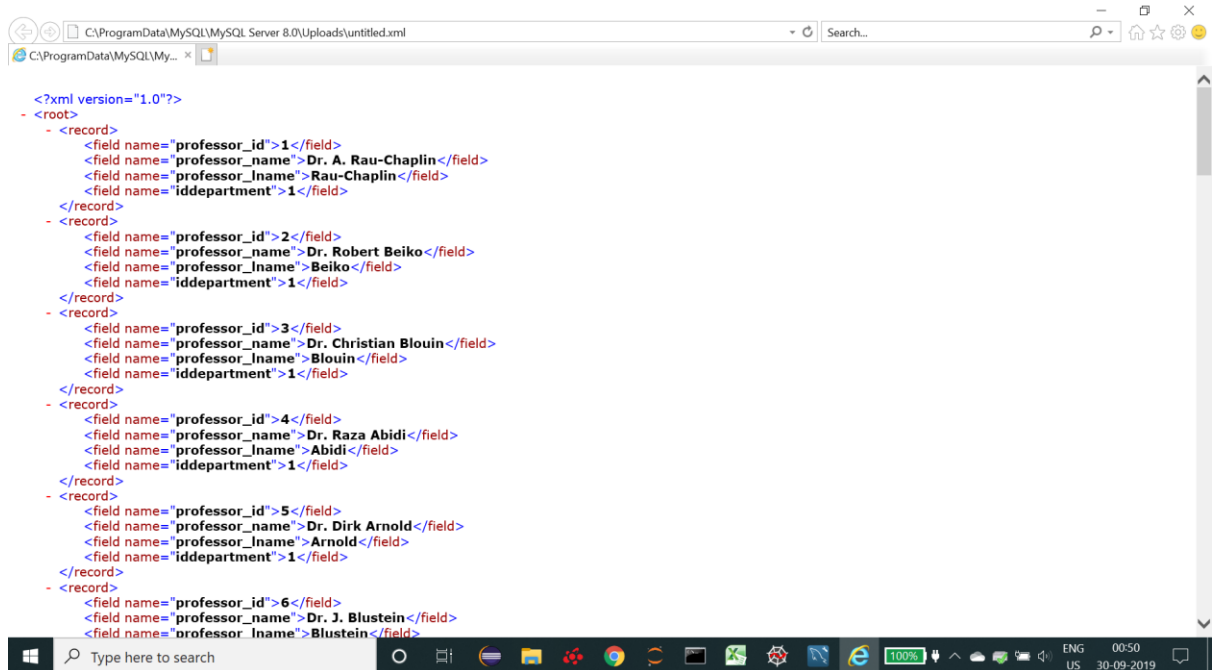
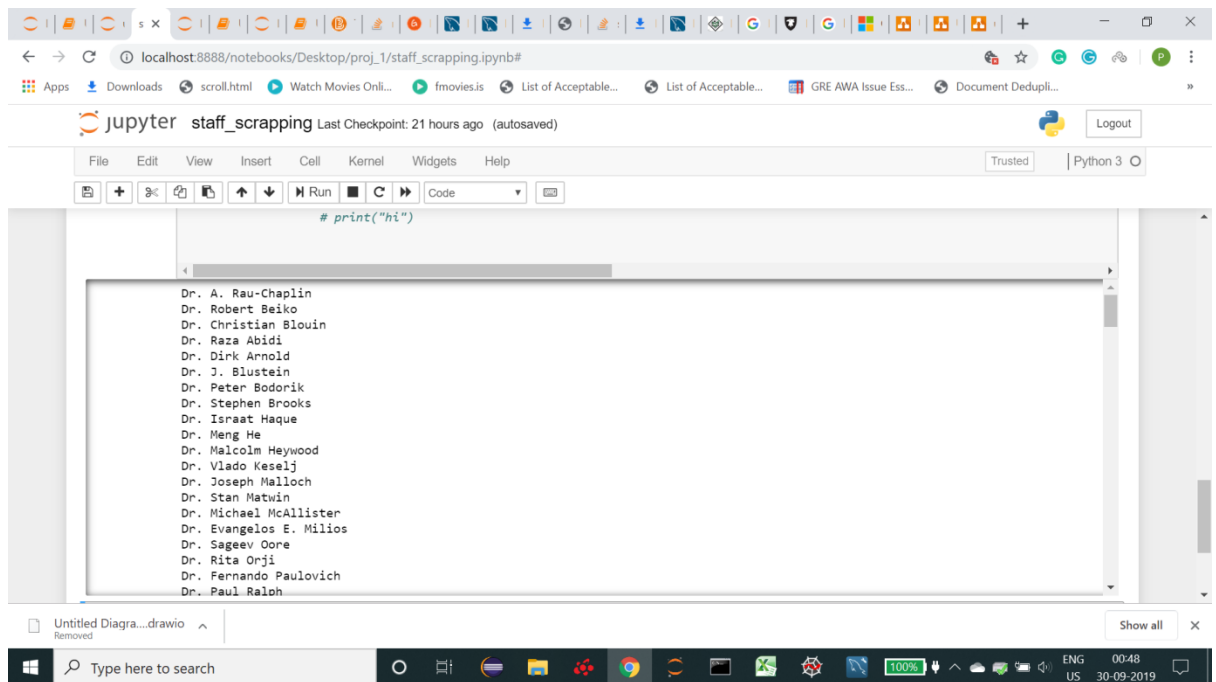
### Disclaimer

Punarva vyas

BannerId: B00841566

In assignment 1 of CSCI 5408 course, data scraping is done manually or programmatically from Dalhousie University's website, and it is used only for educational purpose. Sensitive information, such as personal Email, personal contact numbers are not extracted. However, names of instructors, professors, or other staff members available on the Dalhousie University websites are extracted for course (CSCI 5408) related analysis, such as "find how many employees have similar first name etc." The scope of the extracted data usage is limited to the course CSCI 5408 only. The course instructor and the Faculty of Computer Science cannot be held responsible for any misuse of the extracted data.

- A) I have used draw.io website for drawing the data model for google app and for website of the university. Initially I analysed the data set provided and thought of various entities that can be made. After rigorous analysis I managed to get 4 entities Genre, application, Category and Reviews. Many apps can be included in one category, one application has many reviews, one application can belong to many genres and one genre has many applications so I have shown in my ER diagram. Similarly I have opened the dal website and have drawn the ER diagram. The diagrams can be seen in drawio folder.
- B) I have used Beautiful soup for extracting the desired data stored it in a list. I did some processing on the data extracted to take the data that is only required. Then I converted the data into xml format and stored it in a xml file. Here are the images of the extracted data and Xml converted data.



C) Now xml file which was created in data extraction process was taken and data was inserted in mysql database with the above query as shown in figure.

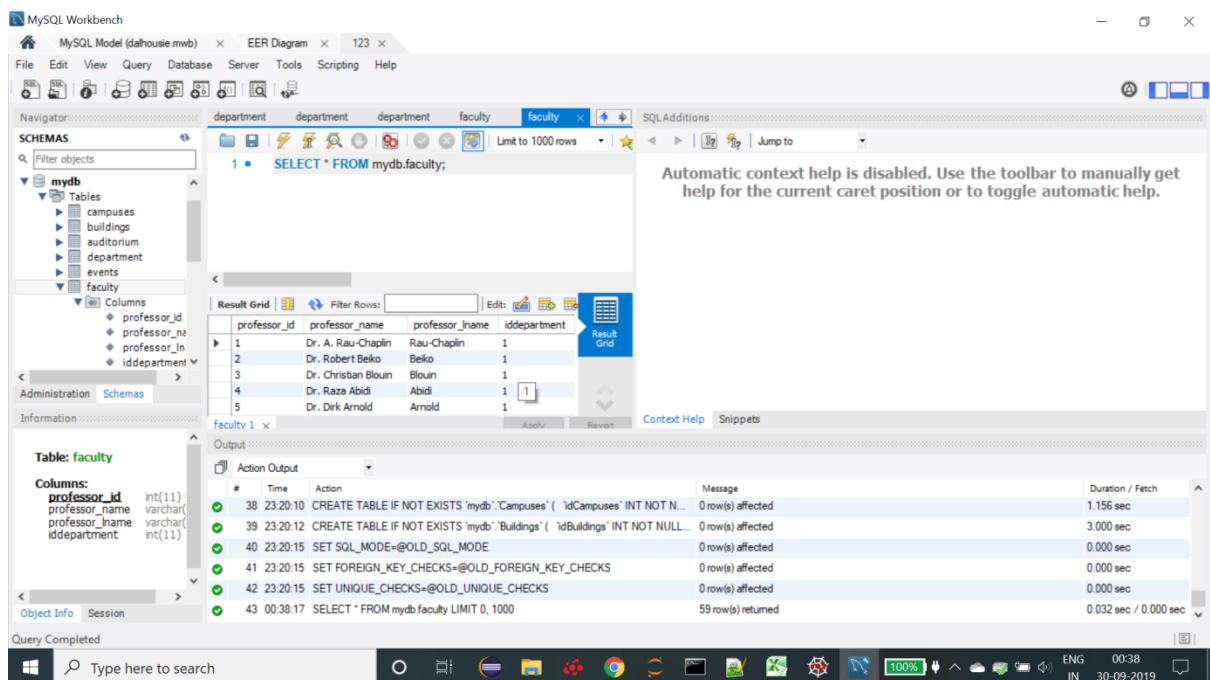
```

Select Command Prompt - mysql --local-infile=1 -uroot -p
mysql> LOAD XML INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/untitled.xml'
-> INTO TABLE faculty
-> ROWS IDENTIFIED BY '<record>';
Query OK, 36 rows affected (0.12 sec)
Records: 36 Deleted: 0 Skipped: 0 Warnings: 0

mysql> LOAD XML INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/untitled1.xml'
-> INTO TABLE faculty
-> ROWS IDENTIFIED BY '<record>';
Query OK, 23 rows affected (0.14 sec)
Records: 23 Deleted: 0 Skipped: 0 Warnings: 0

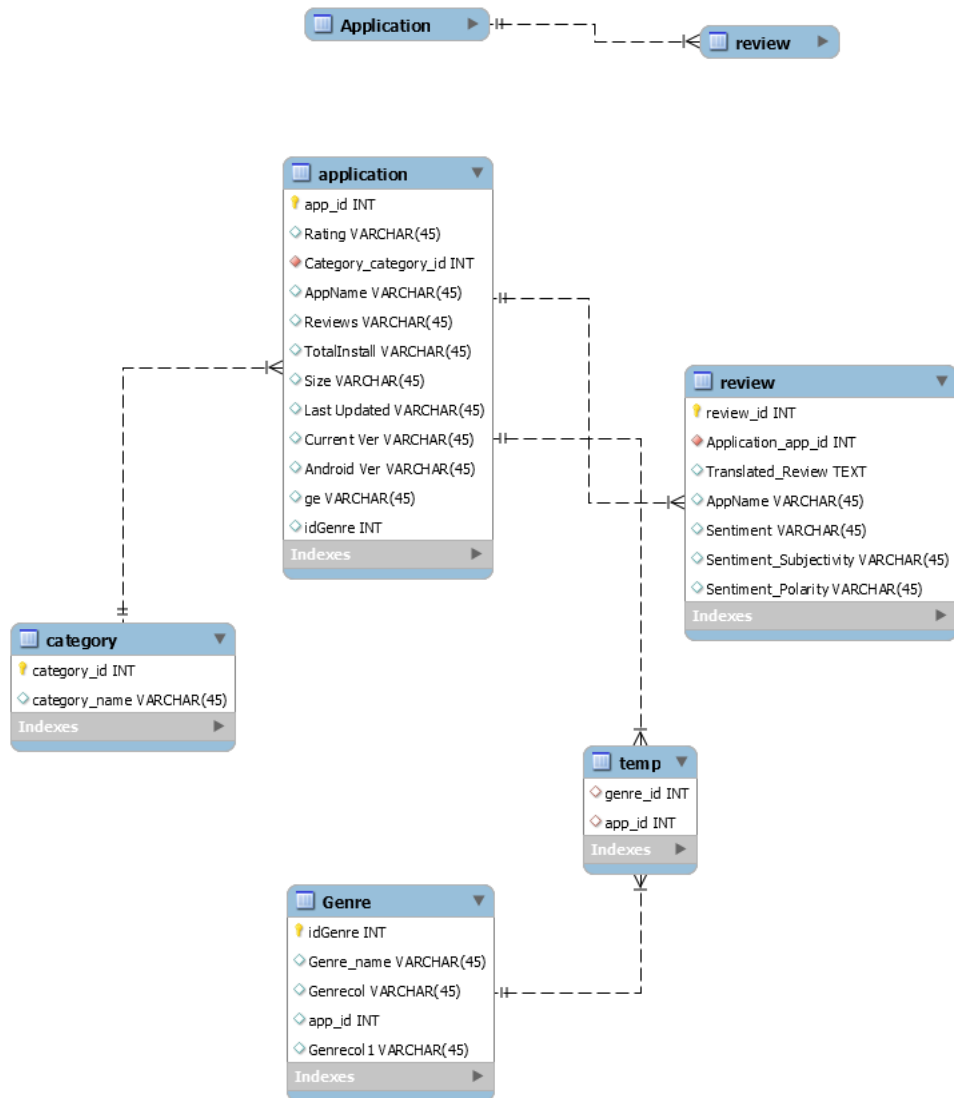
mysql> set foreign_key_check=0;
ERROR 1193 (HY000): Unknown system variable 'foreign_key_check'
mysql>

```

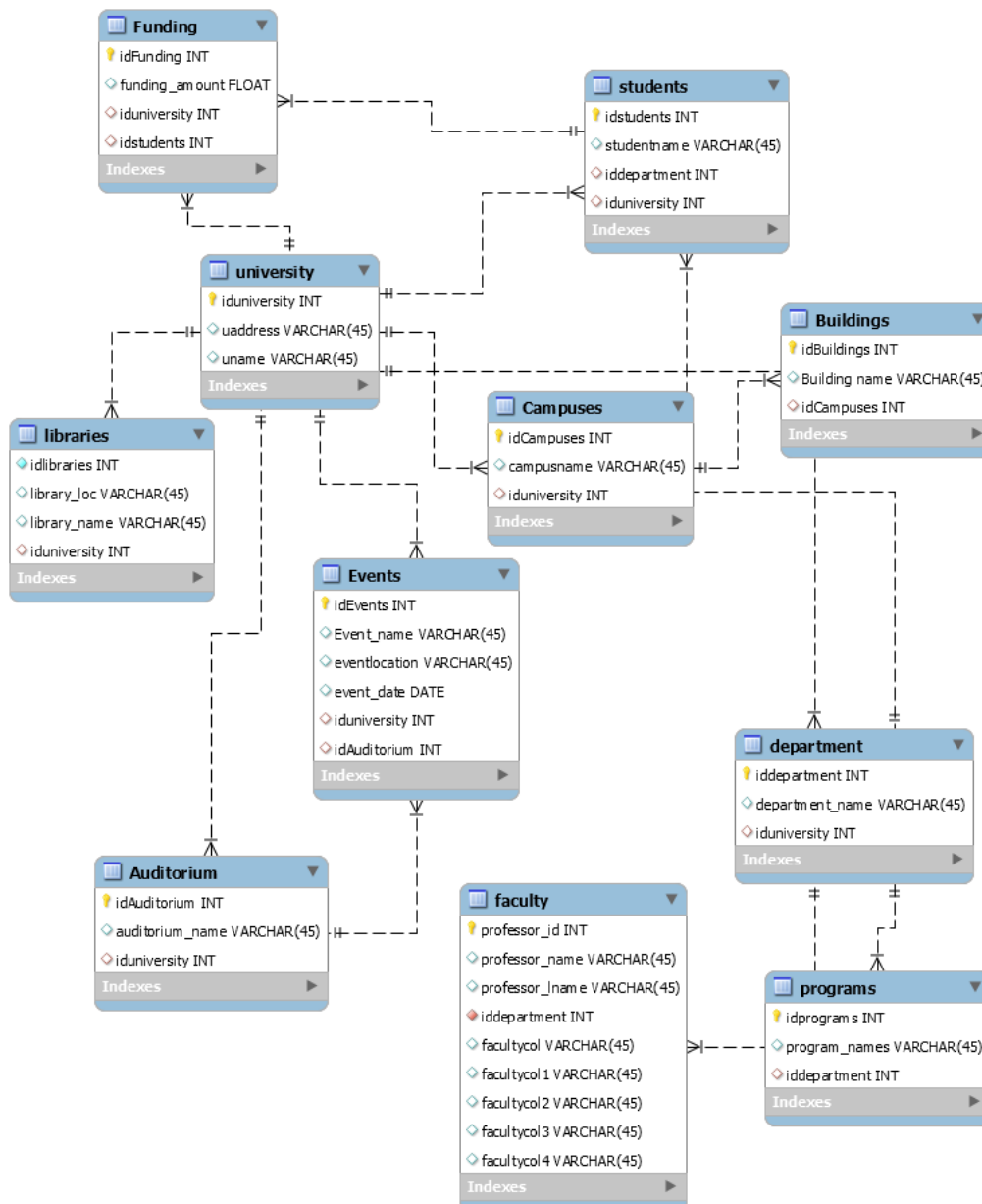


D) Yes my initial sketch had design issues it wasn't in the normalized form. Initially there were only two entities named reviews and app having repetitive values. The searching through them could also have consumed a lot of time. But then I normalized with two more entities named genre and category.

E) These are the two images of before normalization and after normalization.



## Normalised university diagram



F) select dep.department\_name from department dep

inner join faculty fac

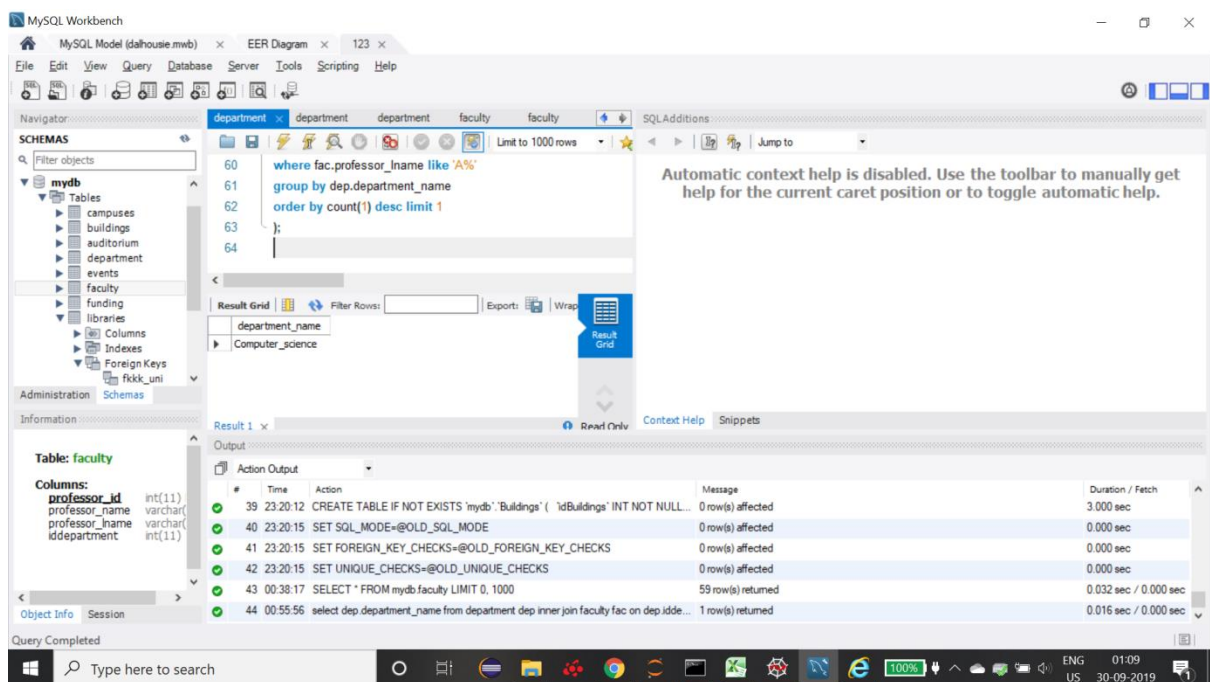
on dep.iddepartment = fac.iddepartment

where fac.professor\_name like 'A%'

group by dep.department\_name

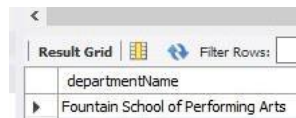
having count(1)=

```
(
select count(1) from department dep
inner join faculty fac
on dep.iddepartment = fac.iddepartment
where fac.professor_lname like 'A%'
group by dep.department_name
order by count(1) desc limit 1
);
```



```
2) select dep.department_name from department dep
inner join programs pp
on dep.iddepartment = pp.iddepartment
where program.programType = 'Undergraduate'
group by dep.department_name
having count(1)=
(
```

```
select dep.department_name from department dep
inner join programs pp
on dep.iddepartment = pp.iddepartment
where program.programType = 'Undergraduate'
group by dep.department_name
having count(1)=
group by dept.department_name
order by count(1) desc limit 1
);
```



The screenshot shows a database query result grid. At the top, there is a header row with the column name 'departmentName'. Below the header, there is a single row of data containing the text 'Fountain School of Performing Arts'. The interface includes a 'Result Grid' tab, a 'Filter Rows' button, and a small icon for the grid view.

departmentName
Fountain School of Performing Arts