DALHOUSIE UNIVERSITY

COURSE ID: CSCI5408

DATA MANAGEMENT, WAREHOUSING AND ANALYTICS

Assignment: 1

Type of the document: Assignment Report

INITIAL DESIGN:

Data model for Google data:

The initial design was made based on the information provided in the CSV files, i.e., for the google app reviews. I have created two Entities named ApplicationUserReviews, GooglePlayStoreApps, and the attributes for GooglePlayStoreApps are App, Category, Rating, Reviews, Size, installs, Type, Price, Content Rating, Genres, Last Updated, Current Version, Android Version. Attributes for ApplicationUserReviews are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity. There are few Null values for several attributes, and the rows were manually deleted in the CSV file before importing into the Tables. Also, the Applications with same name and different Genres were not deleted because of the genre difference ad they were imported to the tables.

The initial ERD structure for this data model can be found in Figure:1.

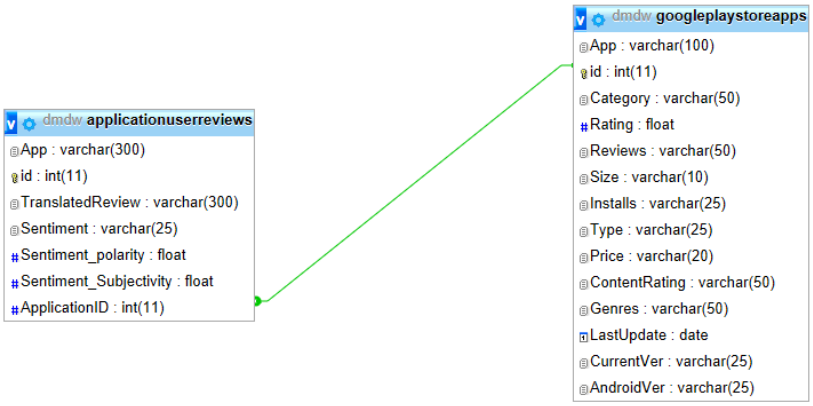

Figure: 1

I have added ID as primary attribute to the both entities and related them with the help of App name.

The initial design has One to Many relationships for Apps, Reviews entities respectively, because an app can have multiple number of reviews.

Data model for business model:

I have analyzed the Dalhousie University website and found 14 entities. They are Academic Programs, Building Directory, Campus development projects, Course, Current student service, Events, Exam, Faculty, Important Dates, It help desk, Libraries, Online tutorials, Residence, Sub programs.

Academic Programs: The entity contains the names of the programs like Graduate, Undergraduate.]

Building Directory: This includes details regarding Id, Building name.

Campus Development Projects: Details of the project location and status of the project.

Course: Contain Details of the course, sub course.

Current Student services: This entity has the details like service provided, web link to the site.

Events: This entity contains the details of the event like Date of the event, time, name.

Exam: Details regarding faculty, course, section, date of exam, time, location.

Faculty: Details regarding name of the faculty, Department, Official Mil Id, office.

Important dates: Details of the notice.

It help desk: Details regarding help type, place.

Libraries: Details include name, Id.

Online tutorials: Details include Tutorial name, Tutorial type.

Residence: Details include Type of residence, name.

Sub course: Details include faculties provided in programs like Graduate, professional.

Initial ERD for the business model can be found in Figure 2.

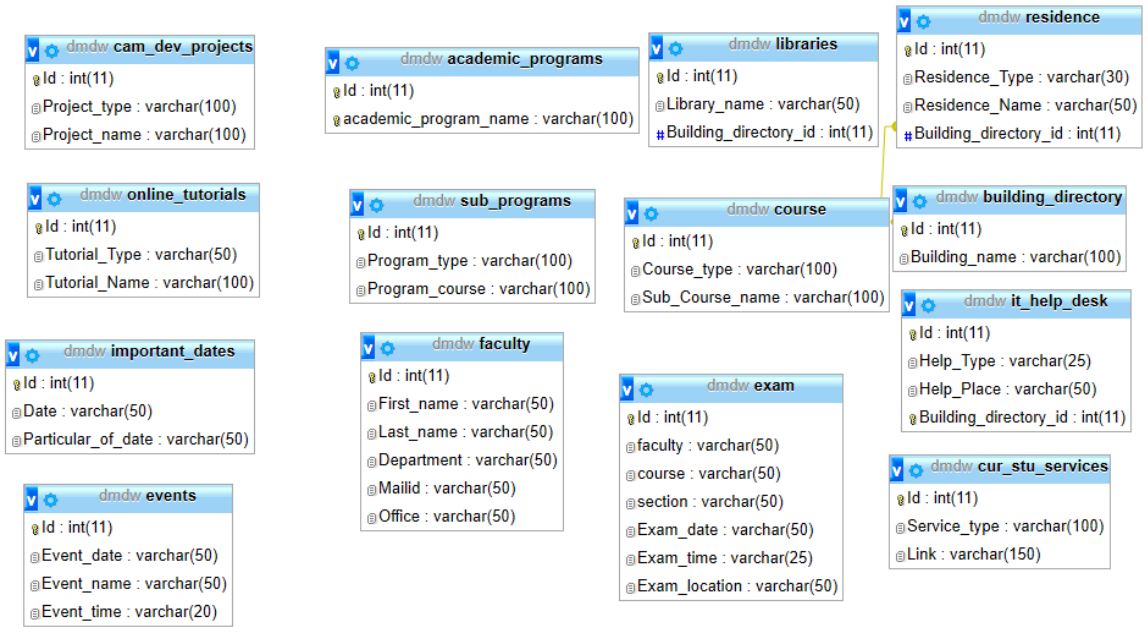

Figure 2.

DATA EXTRACTION AND DATA INSERTION:

Data extraction and collection is done for business model and not for existing data model.

Disclaimer (Data Scraping)

Name: Menni Prem Kumar

Id: B00843422

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 exacted data usage is limited to the course CSCI 5408 only. The course instructor and the Faculty of Computer Science cannot be held responsibility for any misuse of the extracted data.

Please find the attached scripts within the folder containing the python scrip files and xml data files for all the entities.

We have used Python and Beautiful soup package to scrap the data from Dalhousie University website and created XML files to load data into tables.

Data Insertion:

I have used XAMPP server (PhpMyAdmin) to create the tables and to populate them with the datasets.

XML files were attached for each entity.

FINAL DATA MODELING AND NORMALIZATION:

1. **Google review data:**

   The relationship between the entities GooglePlayStoreApps, ApplicationUserReviews are one to many. I have added ID as primary column and then normalized the GooglePlayStoreApps entity by removing columns related to the version details. Thus, by forming a new entity avdetails.
   The relationship between avdetails and GooglePlayStoreApps is one to one and the relationship between GooglePlayStoreApps and ApplicationUserReviews is one to many. The ERD for the Google review data can be found in the figure 3.
   In order to improve the processing time, I have divided the details of the application into two categories and loaded into 2 tables. So that the details can be found in one table thus by reducing the processing time. In the picture there is one to many relationships between avdetails entity and Googleplaystore apps. I tried to edit the relationship and it is not possible in XAMPP which I am using. So, the correct relation for that is one to one. Primary key for the entities is ID's, I have related details with the help of application name.

Figure 3

2. **Business Model:**

   Is my Initial design free from any design issues: No.

   I have tried relating the tables and could able to get fan trap problems. The entities which I have selected are 14 and I can able to successfully relate 3 entities and while relating other 3 the relationship is auto populating with one to many relationships which causing fan trap problem.

The remaining 8 entities which are isolated to each other and they can't be related to each other in any way with the help of current available one. Even related, Fan trap issues are occurring. So I would like to take few other entities in the upcoming assignment and then relates what are the entities which are left. The final ERD after normalization is in the figure 4.

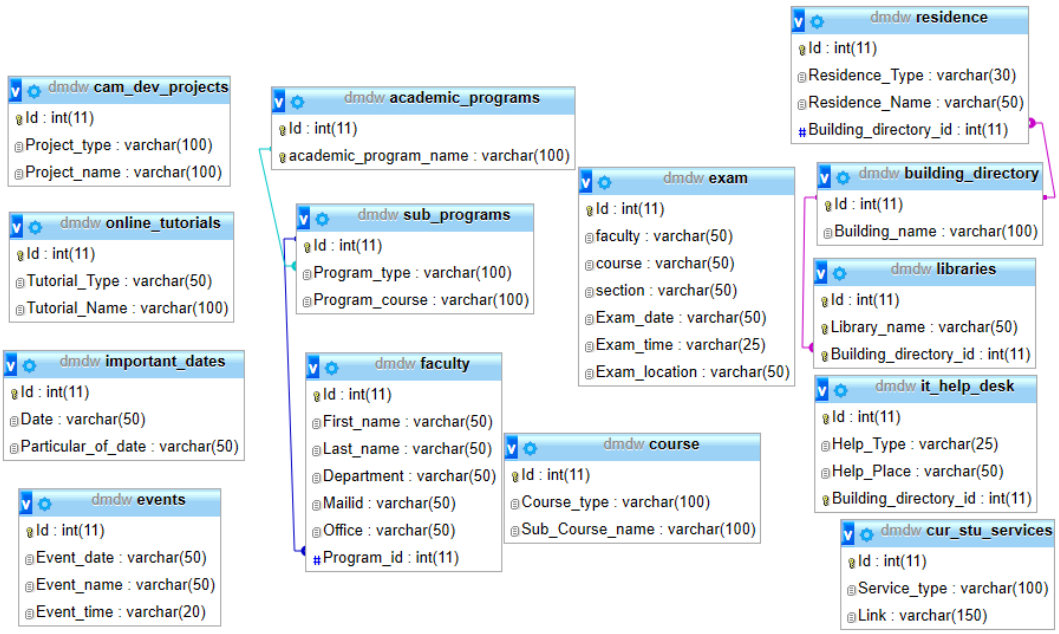

Figure 4

The relationship between academic programs and sub programs is one to many and the relationship between sub programs to faculty is one to many. Regarding important dates and events entities, I don't have an attribute which can relate those tables at present. I would like to add few more attributes and entities for the ERD and will try to relate all the entities in future assignments. The relationship between Building directories, libraries, residence is one to one. I couldn't able to edit the relationships and they are represented as one to many. The correct relationship is one to one for those three. In all the entities, Id's are used as primary keys, used to relate the other tables.

SQL QUERY:

1. SELECT Department FROM `faculty` WHERE Last_name like ("A%") GROUP by Department ORDER by count(*) desc LIMIT 1
   Note: I have taken the details of Computer science, Health Administration.

2. SELECT Program_course FROM `sub_programs` where Program_type = "undergraduate" GROUP BY Program_type, Program_course order by Program_course desc LIMIT 1

REFERENCES:

Internet Site:

[1] Beautiful Soup, http://zetcode.com/python/beautifulsoup/ [Accessed on Sept 23rd 2019]

[2] Data Scraping, site: https://www.dal.ca/ [Accessed on Sept 20th 2019]