INFO-6210 DATA MANAGEMENT AND DATABASE DESIGN

UNIVERSITY DATABASE

ABSTRACT:

University Database is gathered with an intention to make well-informed decisions while choosing a university. Not just for students but also for their parents. Large amount of data is gathered, cleaned and structured in multiple tables providing different aspect of university and the essential factors to be considered while choosing university, for example : tuition fees, entrance exam scores cutoffs, University Rankings by different agencies, location of university, number of student etc.

OBJECTIVES:

- Selection of appropriate data
- Delineate the right commands to ensure their correct use and reduce the likelihood of error occurring.
- Draw significant conclusion from the data to ensure efficient choice making

Requirements/Task(s):

Task 1: Data Gathering

API

Firstly using the API provided by U.S Department of Education API to extract information of all the US universities in the US. The gathered data was then cleaned, audited and normalized. The

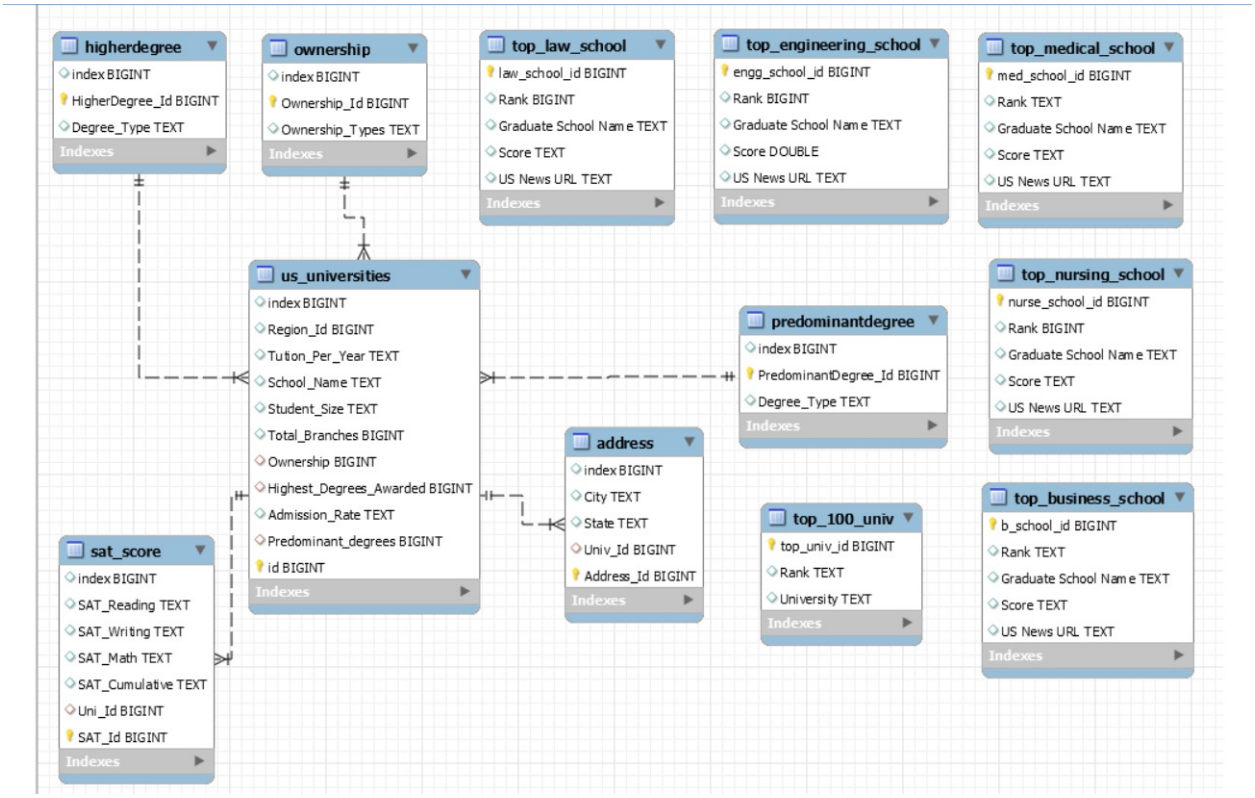
Webscraping

From Top universities the top 100 univeristy ranking data was obtained.

Also from US news the top 100 universities by category was obtained. Now the ese information were also cleaned,audited and normalized to combine it to the data from API in MYSQL.

Task 2 : Auditing and Normalizing Data

ER DIAGRAM



Audit Validity/Acurracy :

In this fast moving digital world, having accurate data is one of the most important aspects of data collection. Incorrect data may result from migration of data from one database to another, presence of incorrect values, or even time-bound data changes.Reviewing is an efficient way to check the correctness of the data.To get the correct accurate data we have used trending hashtags from twitter to get the tweet and data related to the hashtags. The documents in collections are created with specific filesd like tweet_id,hashtags,username,followers,etc to remove unwanted metadata and other data associated and which is not required.

Audit Completeness :

Data completeness refers to whether there are any gaps in the data from what was expected to be collected, and what was actually collected.The problem of

incomplete data can be resolved by ensuring that the data cannot be submitted, unless all expected data is present. Having a mandatory feild of PlayerID, Game ID and Team ID,Tweet_Id etc has made sure there is completeness and has resulted in less time consumption for audting completeness.

Audit Consistency/Uniformity:

Data consistency becomes a challenge to check while using No-SQL,but since the data that we have used is converted from SQL,and the SQL data is checked for consistency using primary key,foreign keys which were created while normalization the data is consistent as the keys are still present and can be used to check for the consistency.
Similarly the data from Twitter is consistent as it is created using specific fields and removing all that is not required(meta data,timestamps,etc).

Normalization:

By limiting a table to one purpose you reduce the number of duplicate data contained within your database. This eliminates some issues stemming from database modifications.

To achieve these objectives, we'll use some established rules. As you apply these rules, new tables are formed. The progression from unruly to optimized passes through several normal forms.

1.First normal form (1NF) • Each table has a primary key: minimal set of attributes which can uniquely identify a record • The values in each column of a table are atomic (No multi-value attributes allowed). • There are no repeating groups: two columns do not store similar information in the same table.

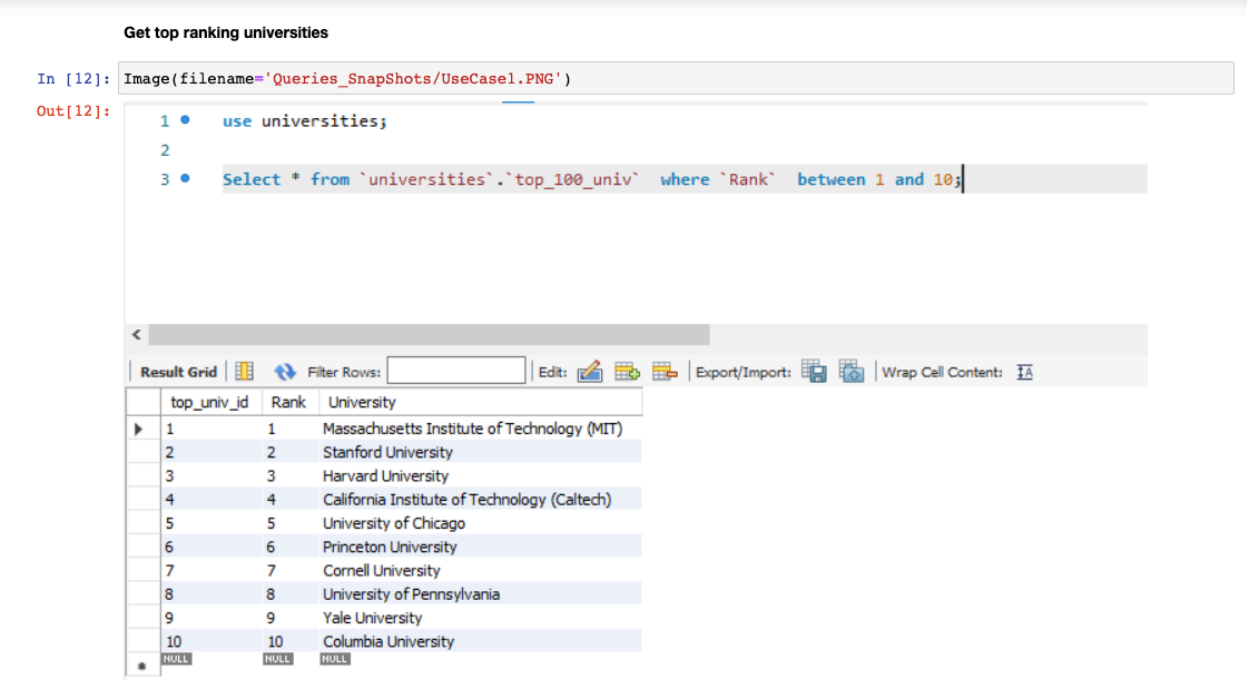
2.Second normal form (2NF) • All requirements for 1st NF must be met. • No partial dependencies. • No calculated data

3.Third normal form (3NF) • All requirements for 2nd NF must be met. • Eliminate fields that do not directly depend on the primary key; that is no transitive dependencies.

Task 3 : Performing Analysis

Analysis was performed using SQL queries and tableau was used to visualize the results.

For example,



Summary:

- The API data is latest which is 2019 Fall.
- Only 18.7% of the colleges have listed their cumulative SAT scores displayed.
- The highest number of Universities are in the states California and New York and in the cities New York and Chicago.
- The highest and the most predominant awards are Certificates and then Undergradutes. It is rarely a graduate degree.

- The country has more of Private- For-Profit Universities

Link to access project:

Github :

- Ankita : https://github.com/fx2044/tiwari_ank_dmdd
- Vasuki Manoharan

References:

https://github.com/dhavalpotdar/College-Scorecard-Data-Analysis

https://github.com/RTICWDT/open-data-maker/blob/master/API.md

http://brunokoba.com/blog/Webscraper/

https://www.pythonforbeginners.com/api/python-api-and-json

https://learning.oreilly.com/library/view/web-scraping-with/9781491910283/ch04.html

https://www.youtube.com/watch?v=SPtEh9c5Xf4

https://stackoverflow.com/questions/7884567/python-web-scraping-beautiful-soup

https://www.geeksforgeeks.org/working-with-missing-data-in-pandas/

https://www.dataquest.io/blog/sql-insert-tutorial/

License :

Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.