College selection platform

The goal of this project is to design a relational database from raw data available on the government website of the United States(link). The aim is to integrate the available information of the universities of United states so as to provide students a database and platform to compare and choose appropriate university for their studies.

Contents of this document:

1. Information about the relations and corresponding attributes

Relation Name	Attribute Name	Attribute data description
Institution	institutionID	ID of the institution
	institutionName	name of the Institution
	city	city of institution
	state	state of institution
	zipCode	City of institution
AccredatedBy	institutionID	ID of the institution
	accredCode	code of accredating agency
AccredatingAgency	accredCode	code of accredating agency
	accredAgency	name of accedating agency
Relation Name	**Attribute Name**	**Attribute data description**
InstitutionInformation	institutionID	ID of the institution
	url	url of institution website
	npcUrl	url of netprice calculator of institution
	mainCampus	True/False stating if the information belongs to the main campus of the institute
	numBranch	Number of branches of institution
	governanceStructure	indicating if the institute is public, private(nonprofit) or private(profit)
	affiliation	indicating if the institutions is identified as minority-serving institutions
	admissionRate	rate of admission in the institute
	totalAdmissions	total number of degree/certificate-seeking students enrolled
	pctPartTimeAdmissions	proportion in percentage of the students enrolled for part-time education
	completionRate	students who complete within 100 or 150 percent of the expected time to completion
	avgFacultySalary	Average salary of faculty
	rating	rating of the institute given by the users of this application
	ranking	ranking of the institute
	onCampusHousing	indicating if OnCampus housing facility is available or not
	employeeSatisfaction	feedback of employee satisfaction at the institute
	transportFacility	indicating if transport facility is available or not
	numReviews	number of reviews of the institute given by the users of this application
Expenses	institutionID	ID of the institution
	totalFee	total program fee
	tuitionFeeInState	tuition fee for in-state students
	tuitionFeeOutState	tuition fee for Out-state students
	bookSupplies	bookSupplies expenses
	housing	housing expenses
	miscellaneous	miscellaneous expenses
InstitutionType	institutionID	ID of the institution
	menOnly	indicating if the institute contains only male students
	womenOnly	indicating if the institute contains only female students
	distanceOnly	indicating if the institute provides only distant education

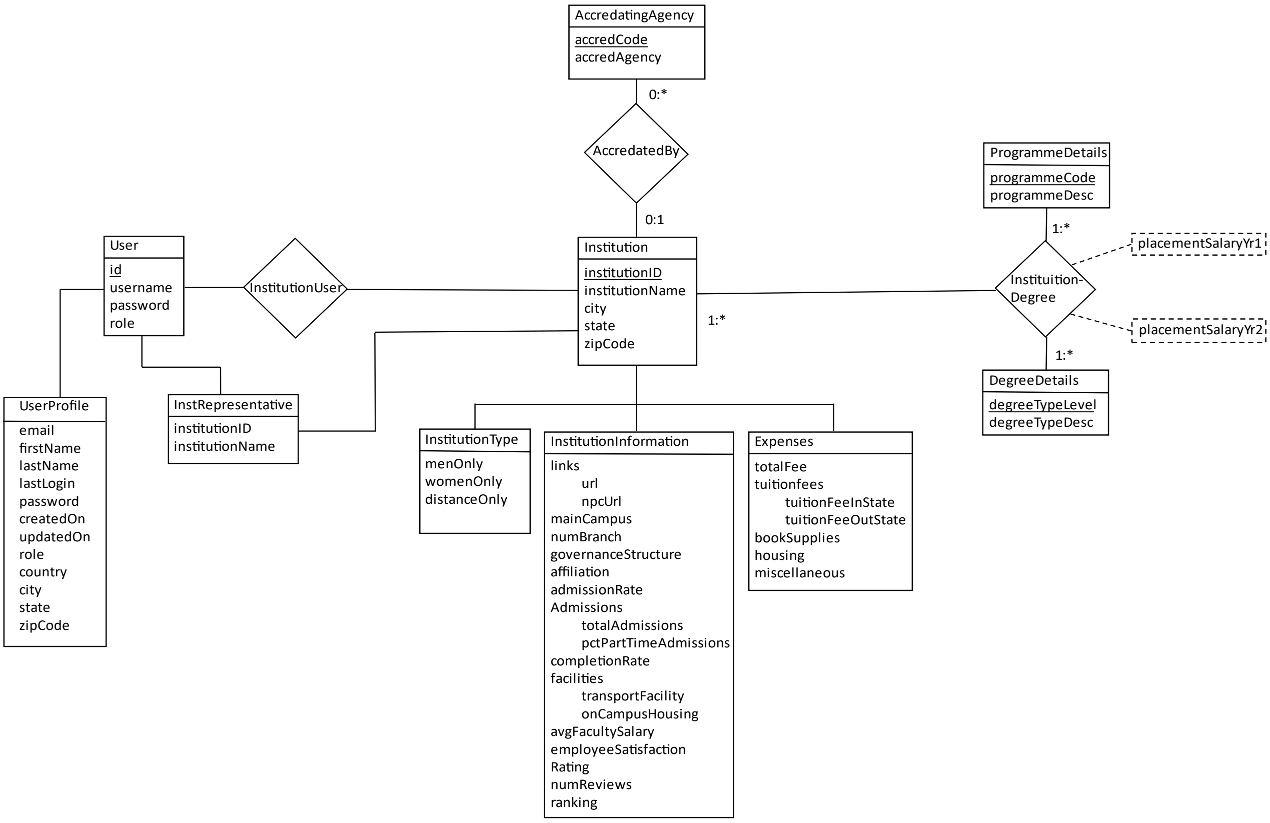
Relation Name	**Attribute Name**	**Attribute data description**
InstitutionDegree	institutionID	ID of the institution
	programmeCode	Programme code
	degreeTypeLevel	Level of the type of degree
	placementSalaryYr1	Average salary of students after 1 year of graduation
	placementSalaryYr2	Average salary of students after 2 years of graduation
ProgrammeDetails	programmeCode	Programme code
	programmeDesc	Programme description
DegreeDetails	degreeTypeLevel	Level of the type of degree
	degreeTypeDesc	description of the type of degree
User	id	
	username	
	password	
	role	
Relation Name	**Attribute Name**	**Attribute data description**
InstitutionUser	institutionID	ID of the institution
	id	user's IDs who have applied to the institutions in the above institutionID
Country	countryId	
	countryName	
userProfile	id	
	email	
	firstName	
	lastName	
	createdOn	
	updatedOn	
	country	
	city	
	state	
	zipCode	
representative	id	
	institutionID	
	institutionName	

Note on data format modification:

The column Affiliation contains information indicating if the institutions is identified as minority-serving institutions. It has values containing 6-digit integer, each indicating a type of university.
For example, if the value is 110000, it means that the corresponding institution falls in first two categories of types. The categories are as follows:

- HBCU=Historically Black Colleges and Universities
- PBI=Predominantly Black Institutions
- ANNHI=Alaska Native-/Native Hawaiian-serving Institutions
- TRIBAL=Tribal Colleges and Universities
- AANAPII=Asian American-/Native American-Pacific Islander-serving Institutions
- HSI=Hispanic-serving Institutions
- NANTI=Native American Non-Tribal Institutions

2. ER Diagram



3. Explanation of any non-obvious translations from ER model to relational schema:

1) The relationship InstitutionDegree connects the institution to the Program Details and the Degree Details. An institution can have multiple programs. A program can be offered in multiple levels of degrees. Thus, the relationship from institution to programs is one to many, and the that from program to levels of degrees is also one to many.
   Similarly, a degree level can be associated with many programs and a program can be associated with many institutions. Hence the relationship from degree level to programs is one to many and that from program to institutions is also one to many.
   The relationship also has two more attributes placementSalaryYr1 and placementSalaryYr2.
2) Institution has information, expenses and type. So, the attribute institutionID of the entity Institution is inherited in these three entities.
3) An institution can be accredited by a single accrediting agency. However, an accrediting agency can accredit multiple institutions. The relationship AccredatedBy connects institution with accrediting agencies. Hence. For institution to agency, it is zero or one to one, and for agency to institution, it is zero or one to many.
4) The user and the institution are connected through a relationship InstitutionUser. If a user applies for admission to a particular institution, then that user’s id and the corresponding institutionID will be populated in this relationship. A user can apply to multiple institutions and an institution can have multiple applicant users.
   The user has a user profile. And institution representative is a type of user. So InstRepresentative is a partial specialization of user.

4. Relational Schema:

Schemas derived from the entity sets:

Institution (institutionID, institutionName, city, state, zipCode)
AccredatingAgency(accredCode, accredAgency)
InstitutionInformation (institutionID, url, npcUrl, mainCampus, numBranch, governanceStructure, affiliation, admissionRate, totalAdmissions,pctPartTimeAdmissions, completionRate, avgFacultySalary, rating,ranking, onCampusHousing, employeeSatisfaction, transportFacility, numReviews)
Expenses (institutionID, totalFee,tuitionFeeInState, tuitionFeeOutState, bookSupplies, housing, miscellaneous)
InstitutionType (institutionID, menOnly, womenOnly, distanceOnly)
ProgrammeDetails(programmeCode, programmeDesc)
DegreeDetails (degreeTypeLevel,degreeTypeDesc)
User (id, username, password, role)
UserProfile(id, email, firstName, lastName, lastLogin, createdOn, updatedOn, country, city, state, zipCode)
InstRepresentative (id, institutionID, institutionName)
Country (countryId, countryName)

5. The relationship sets in our design are listed below:

AccredatedBy: relating institution with the accrediting agencies
InstitutionDegree: relating institution with the programs and degree levels
InstitutionUser: relating institution with the user

6. Schemas derived from the relationship sets:

AccredatedBy (institutionID, accredCode)
InstitutionDegree (institutionID, programmeCode, degreeTypeLevel, placementSalaryYr1, placementSalaryYr2)
InstitutionUser (institutionID, id)

7. Design Choices:

1) Institute is identified uniquely by its ID and its address. So, the Institution relation is the main relation. There are three design choices made as follows:
   i) InstitutionType: This relation inherits the ID attribute from Institution. It has been made as a separate table because generally, very less is the frequency that a user wants to search if the institution is of type having only men, only women, or if it provides only distance education.
   ii) InstitutionInformation: This relation contains detailed information about the universities. While retrieving the list of universities, only a basic information is necessary to be displayed on the home page of an application (such as name, ranking, etc.). Hence, the detailed information is kept separate from the main Institution relation even though the number of rows of both might be comparable.
   iii) Expenses: The number of institutions having/providing this information is around half of the total number of institutions. Hence, this relation is being made separately.
2) Institution is connected to the ProgramDetails and DegreeDetails through relationship InstitutionDegree. The design choices are justified below:
   i) DegreeDetails: A single institution has multiple degree levels. In this database, there are 8-degree levels. If there were to be added in institution relation, there would be a data redundancy.
   ii) ProgramDetails: Similar to above, the distinct values of program codes and their corresponding names are very less as compared to number of institutions. Secondly, the programdetails would be multivalued attributes.

   Hence to prevent the database from data redundancy, these two relations are prepared and are connected to institution via InstitutionDegree. Further, there are two more attributes in the relationship. This data is available comparatively for very a smaller number of schools. So as the primary key required for these two is already present in the relationship. Hence there two attributes are added to this relation.

8. Constraints checks:

1) In the InstitutionInformation relation, the totalAdmissions represents total number of students that were admitted. The attribute pctPartTimeAdmissions represents the percentage of students amongst those who were part time students. So, if the totalAdmissions data is available, only then the pctPartTimeAdmissions can be available.
2) For the relation InstitutionType, if the institution is of the type menonly, then the value for womenOnly should not be true. And vice-a-versa.

9. Data Loading operation details:

1) The dataset is taken from a website containing America's education data. (link)
2) After finalizing the columns, we extracted the selected data and performed data cleaning operations.
3) We prepared base tables alldata, fieldata and dataforuser. The data for all the tables was inserted from these tables.

10. References

[1] https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6-2020/resources
[2] https://dev.mysql.com/doc/refman/8.0/en/
[3] https://www.techonthenet.com/mysql/index.php
[4] https://stackoverflow.com/
[5] A. Silberschatz, HF. Korth, S. Sudarshan, Database System Concepts (6)