CSC343 - Project Phase 2

Xiaochun Tong, Farzaneh Tabandeh

Nov 2021

1.Datasets and Cleaning Process

1.2. Original Datasets

As we had in our phase 1 submission, we used the following data sets:

Countries' Names and Codes
Countries Gross Domestic Product (GDP)
Countries employment rate
Employment based on age group
Employment based on education level
Number of enterprises of specific size

1.2. Data Cleaning

In order to work with these datasets in a meaningful way, we needed to clean the original data files into different CSV files. We created a python script file that reads the original data files and produces a cleaned version of them. That way, we produced one cleaned data file for each specific table. The python file that we used is called **`generate_clean_data.py`**. This file assumes that the cleaned data files are located in the current directory where this python file is run. You will need to hardcode the name of the original files into the python file as specified at the top of the file itself.

2.Design Decisions

The following is our schema for this project.

Country(countryCode, countryName)
CountryGDP(countryCode, year, gdp, gdpPerCapita)
Employment(countryCode, yearQuarter, employmentRate)
EmploymentByAge(countryCode, yearQuarter, ageGroup, employmentByAgeRate)
EmploymentByEducation(countryCode, year, educationLevel, employmentByEducationRate)
EmploymentByEnterprise(countryCode, year, enterpriseSize, numEnterprises)

CountryGDP[countryCode] ⊆ Country[countryCode]
Employment[countryCode] ⊆ Country[countryCode]
EmploymentByAge[countryCode] ⊆ Country[countryCode]
EmploymentByEducation[countryCode] ⊆ Country[countryCode]
EmploymentByEnterprise[countryCode] ⊆ Country[countryCode]

The following is the changes that we made compared to phase 1:

- We added the table “Country” to uniquely define the countryCode and countryName for each country.
- We renamed the table “CountryGDP'' from another name to specifically refer to what it contains.
- We made sure the name of the attributes of each table is self-explanatory and clear.

The following are the design decisions that we specifically made in the Postgres schema in the file **`schema.ddl`**:

1) we have defined 3 custom data types (i.e. domains):
    - AgeGroup
    - EducationLevel
    - EnterpriseSize

   These custom data types make sure that the data inserted into the tables `EmploymentByAge`, `EmploymentByEducation`, and `EmploymentByEnterprise` are according to the definitions and constraints of the original data.

2) We also defined the ”countryCode'' attribute to be of type varchar(3), the “theYear” attribute to be of type varchar(4), and the value of “countryName'' attribute to be of type varchar(100), instead of TEXT which we had previously. Note that we decided to use the name “theYear” for this attribute since the word “year” is a special keyword in SQL and was not allowed.

3) We have enforced the value of “not NULL” for all of the attributes except for the attribute “gdpPerCapita” in the table “CountryGDP”. As in our current dataset, some of the countries in the table “CountryGDP” did not have the value for the attributes “gdpPerCapita”, we allowed this attribute to accept the value of NULL as well. One reason for this choice was that our dataset is like a time series dataset. That means we want to know whether the value of “gdpPerCapita” was reported in a specific year or not. So having a value of NULL will help us to know that.
   Note that if the “time series scenario” was not the case, we had another option as well. Because there could be the case that a country could have the value of “gdp” and not the value for “gdpPerCapita”, we could separate out the two tables of “Country” and “CountryGDP”. That way we could enforce the value of “not NULL” for “gdpPerCapita” as well.