University of Ottawa

School of Electrical Engineering and Computer Science

CSI4142 Fundamentals of Data Science

Project Phase 2: Physical Design and Data Staging

Instructions:

A. This is a team assignment.
B. Submit your documentation via BrightSpace using your team locker.
C. For your source code, you may either submit a zipped file or provide a link to a GitHub repository.
D. Demonstrate your work during a Zoom meeting with the TA, in the timeslot allocated to you. Note that all team members are required to attend this demonstration and you will be asked to turn your cameras on.

Project Description - World Bank Nutrition, Health and Population data mart

Data science and Artificial Intelligence (AI) have been very successful to discover important trends in data over time. Increasingly, to organizations such as the World Bank provide access to open-source repositories for data analytics and data mining, in order to enable data scientists to use these resources in their individual projects.

Specifically, the World Bank Health Nutrition and Population Statistics (WB-HNP) database "provides key health, nutrition and population statistics gathered from a variety of international and national sources. Themes include global surgery, health financing, HIV/AIDS, immunization, infectious diseases, medical resources and usage, noncommunicable diseases, nutrition, population dynamics, reproductive health, universal health coverage, and water and sanitation" [1].

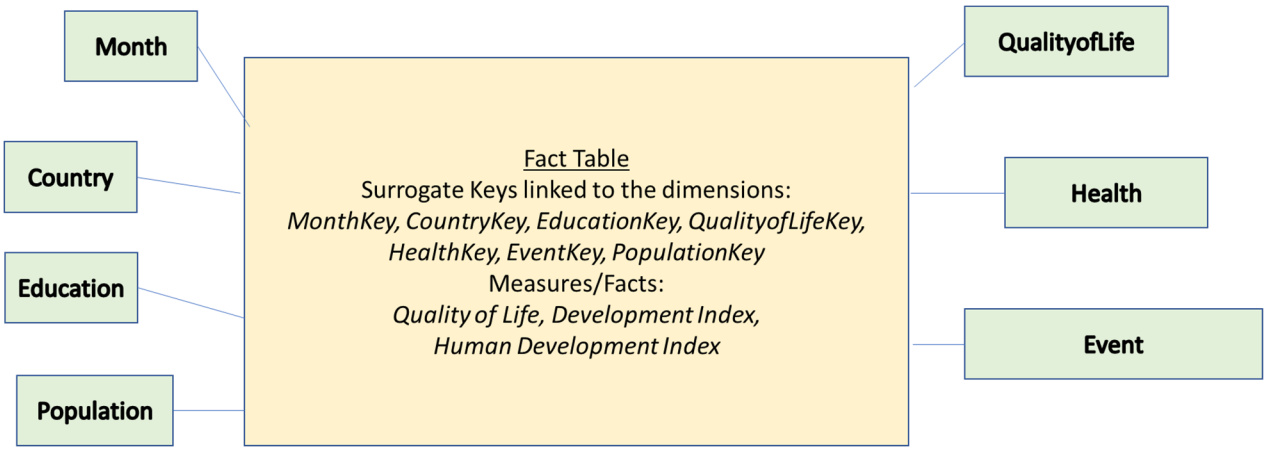
As a first step, you completed a conceptual design for the WB-HNP data mart, for the entire world population.

As a proof of concept, you now decide to focus on only nine (9) countries, for the period spanning 2005 to 2020.

Your choice should include the three (3) countries in North America (Canada, the USA and Mexico) as well as at least six (6) countries from two (2) other subregions of the world.

Note that three (3) of these six (6) countries should be classified as developing [3], according to the world development index, while the remaining three (3) should be classified as underdeveloped nations [4].

Below a suggested high level conceptual model, where the dimensional attributes are not shown.



You will notice that the data mart tracks the monthly data of countries, in terms of aspects such as education, quality of life, and health. In addition, information about environmental, economical, and geopolitical events are also included. Finally, the three (3) measures included are quality of life, development index, and human development index. It follows that the values stored in these measures may change over the timeline of the data mart, e.g., a country moving from a developing nation to a developed nation, and vice versa.

Please note the following details.

1. Three measures – *Quality of Life, Development Index* and *Human Development Index* – are used to assess and to contrast countries. These measures may be obtained from external sources, such as [2] to [9].
    - ~ The *Quality of Life Index* [2] is a holistic measure used to determine the "satisfaction with life" of populations. It attempts to measure which countries will provide the best opportunities for a healthy, safe, and prosperous life in the years ahead. As such, it refers to an individual's perception of their

position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards, and concerns. The ratings of individual countries may be mapped into {excellent = 1, high = 2, medium = 3, low = 4, very poor = 5}.

- ~ The well-known *Development Index* [3] is often used to determine the level of economic development and may be mapped to values in {developed = 1, developing = 2, underdeveloped = 3}.
- ~ The *Human Development Index* [8, 9] is used to encompass topics such as education, poverty, and access to healthcare [5-7].
  The *Human Development Index* takes a value in {Very high = 1, high = 2, medium = 3, low = 4, data unavailable = 5}.

Below some more details about the dimensions and their attributes.

2. Country (CountryKey, Name, Region, Continent, Currency, Capital, Total Population, Birthrate, …).
   The dimension contains the information about each country, with an attribute hierarchy on Name, Region, and Continent. You are required to include at least 12 attributes from the WB-HNP database in this dimension.

3. Month (MonthKey, Name, Quarter, Year, Decade).
   The month dimension is used to be able to have a finer grain, compared to year or quarter, when mapping information about environmental, economical, or geopolitical events. There is an attribute hierarchy on Month, Quarter, Year, Decade.

4. Education (EducationKey, Literacy Rate Attributes, e.g., {Female, Male, Total}, School Enrollment attributes in {primary, secondary, etc.,}, Public Education Spending, …).
   This dimension refers to the attributes relevant for education. You are required to include at least 12 attributes for the WB-HNP database in this dimension.

5. Health (HealthKey, Domestic Health Expenditure, Hospital beds, Immunization attributes, e.g., {Hep, DPT, Measles, Polio}, Number of Surgical Procedures, Number of Death attributes, e.g., {Infant, Stillbirths, Elderly}, Number of health professionals such as {Nurses, Physicians}, Prevalence of health condition such as {stunting, overweight, diabetes, HIV, anemia} in populations {children, male, female, total}, Disease specific attributes, e.g., {Adults with HIV, Adults Newly Infected HIV, Children with HIV, Children Newly Infected HIV,}, …).
   This dimension contains the details relevant for health. You are required to include at least 16 attributes from the WB-HNP database in the dimension.
   (Note: Health may also have been split into "Health" and "Nutrition" dimensions.)

6. Quality of Life (QualityofLifeKey, Quality of Services e.g., {Access to Drinking Water, Access to Sanitation, Access to Basic Handwashing Facilities}, Unemployment rate attributes e.g., {Female, Male, Total}, Maternal Leave benefits, …).
   The dimension contains details about the general quality of life. You are required to include at least 12 attributes from the WB-HNP database in this dimension.

7. Population (PopulationKey, Life Expectancy attributes, e.g., {Female, Male, Total}, Net Migration, Population Statistics attributes by Gender and Age ranges, Rural Population attributes such as {%, Growth, Poverty}, Urban Population attributes such as {%, Growth, Poverty}, …).
   The dimension covers statistics about the population. You are required to include at least 12 attributes from the WB-HNP database in this dimension.
   (Note: Combining Quality of Life and Population is also a possibility, but by separating them into two different dimensions we give an explicit focus.)

8. Event (EventKey, Name, Description, Start-date, End-date, Start-Month, End-month, Outcome, …).
   The event dimension refers to data from external sources about environmental, geopolitical, or economical events. As such, this dimension corresponds to the "promotion" dimension in the Sales data mart used in class. You are required to include at least ten (10) events per country in this dimension.

Hint: Refer to the data dictionary, as obtained from the WBD-HNP open data website, for further details of the source attribute domains and values.

Deliverables:

A. Create the WB-HNP data mart using the PostgreSQL database management system (DBMS).
B. Follow the data staging steps, as discussed during the lectures, to populate the data mart with the data of at least nine (9) countries, from 2005 to 2020. In addition, your data mart should also include the details of at least ten (10) events that took place in, or affected, each country during the stipulated time period.
C. Submit your source code, in a zipped file, or submit a link to a GitHub repository.
D. Demonstrate your work during a Zoom meeting with the TA, in the time slot allocated to you. Note that all team members are required to attend this demonstration and that you will be asked to turn your cameras on.
E. Submit a PDF file with the following details.
   1. A one-page schematic with your high-level data staging plan.

2. A list of data quality issues you encountered and how you handled them. (For instance, how did you detect and handle missing or noisy data (if any); how did you integrate the data from different sources; etc.)
3. A table containing the following information:

Deliverable checklist	**Responsible team member(s)**	**Expected completion date**	**Actual completion date**	**Estimated time (hours) to complete**	**Actual time (hours) to complete**	**Notes (if any)**
Create database instance						
Create <name> dimension						
Create <name> dimension						
…						
Staging of dimension <name>						
Staging of dimension <name>						
…						
Surrogate key pipeline						
Staging of fact table – including FKs and measures						
Data quality handling and reporting						
Others – if any						

Data Sources: Below some sources with current and historic data. In general, https://worldpopulationreview.com and https://ourworldindata.org contain many interesting facts and rankings. You would also need to consult additional news resources, notable for environmental, economical, and geopolitical events in countries.

[1] https://datacatalog.worldbank.org/search/dataset/0037652/Health-Nutrition-and-Population-Statistics
[2] https://www.numbeo.com/quality-of-life/rankings_by_country.jsp
[3] https://worldpopulationreview.com/country-rankings/developing-countries
[4] https://worldpopulationreview.com/country-rankings/underdeveloped-countries

[5] https://worldpopulationreview.com/country-rankings/education-rankings-by-country
[6] https://worldpopulationreview.com/country-rankings/birth-rate-by-country
[7] https://worldpopulationreview.com/country-rankings/poverty-rate-by-country
[8] https://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index
[9] https://ourworldindata.org/grapher/human-development-index-escosura