CSI 4142 Fundamentals of Data Science

Deliverable 2

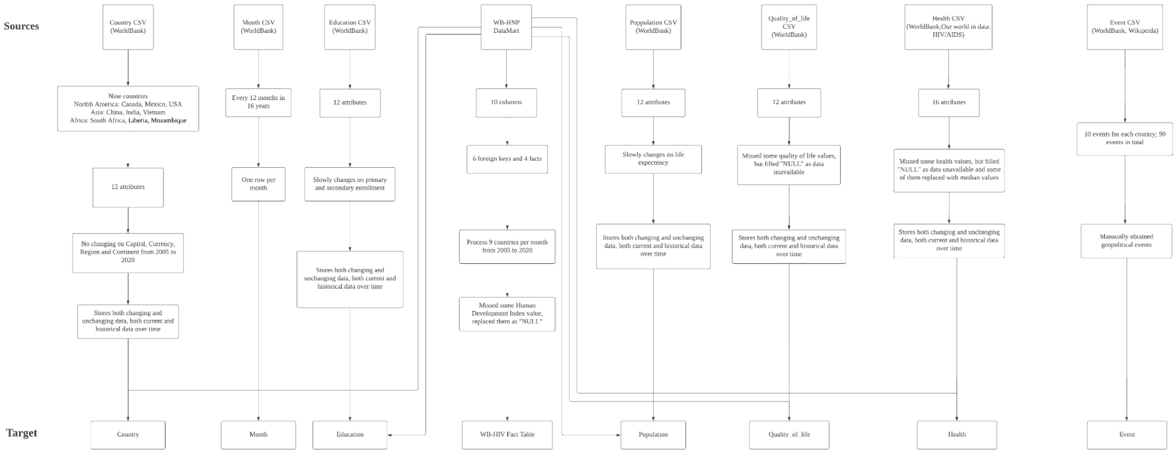
Group 15:

Langqing Zou (300035036)

Kangwei Liao (8568800)

James Chu (300052548)

1. The high-level data staging plan


Sources
Country CSV (WorldBank)
Month CSV (WorldBank)
Education CSV (WorldBank)
WB-HNP DataMart
Poppulation CSV (WorldBank)
Quality_of_life CSV (WorldBank)
Health CSV (WorldBank;Our world in data: HIV/AIDS)
Event CSV (WorldBank; Wikipeida)
Nine countries Northh America: Canada, Mexico, USA Asia: China, India, Vietnam Africa: South Africa, Liberia, Mozambique
Every 12 months in 16 years
12 attributes
10 columns
12 attributes
12 attributes
16 attributes
10 events for each country; 90 events in total
6 foreign keys and 4 facts
12 attributes
One row per month
Slowly changes on primary and secondary enrollment
Slowly changes on life expectency
Missed some quality of life values, but filled "NULL" as data unavailable
Missed some health values, but filled "NULL" as data unavailable and some of them replaced with median values
No changing on Capital, Currency, Region and Continent from 2005 to 2020
Process 9 countries per month from 2005 to 2020
Stores both changing and unchanging data, both current and historical data over time
Stores both changing and unchanging data, both current and historical data over time
Stores both changing and unchanging data, both current and historical data over time
Stores both changing and unchanging data, both current and historical data over time
Manaually obtained geopolitical events
Stores both changing and unchanging data, both current and historical data over time
Missed some Human Development Index value, replaced them as "NULL"
Target
Country
Month
Education
WB-HIV Fact Table
Population
Quality_of_life
Health
Event

PS: Clearer version in [High-level data staging plan.pdf]

2. List of data quality issues

Data missing

When we are collecting data, we encounter data missing a couple of times. Here we demonstrate how we handle two cases of data missing:

- Half of the total data are missing
  - Drop the attribute
    - We consider that attribute is no longer helpful in our data mart
    - Examples during data collection:
      - *Maternal Leave benefits*
      - *Female/male age at first marriage*
  - Find the replacement data source
    - When the data of an attribute that we consider it very important to study in, we decided to find the replacement data source.
    - Examples during data collection:
      - *Human Development Index*
- Less than half of the total data are missing
  - Missing data for consecutive years
    - When the data are missing for consecutive years, for example, data are missing from 2005 to 2010, we consider it not helpful to replace missing data with a median value, so we decided to drop them.
      - Examples during data collection
        - *Maternal Leave benefits*
  - Missing data for a couple of years
    - When the data are missing for a couple of years, for example, data are missing in 2008 and 2015, we decided to replace them with a median value in a specific range.
      - Examples during data collection:
        - *Children With Diabetes*, we replaced the missing in 2008 with the median value of 2005-2010 and we

replaced the missing data in 2015 with the median value of 2010-2015.

- There are some missing data in Quality_of_life, we filled them as "NULL" as unavailable data

Data inconsistency

When working with multiple data sources, it's likely to have mismatches in the same information across sources. The discrepancies are in formats, units, and spellings. We tried these ways to avoid and handle data inconsistency:

- Make sure every column name is unique and formatting. And for convenience, we only use lowercase for names.
- Make sure the type of value is reasonable

Data duplication

When we integrate the data from different sources, we encounter a data duplication issue. We tried these ways to avoid and handle data duplication:

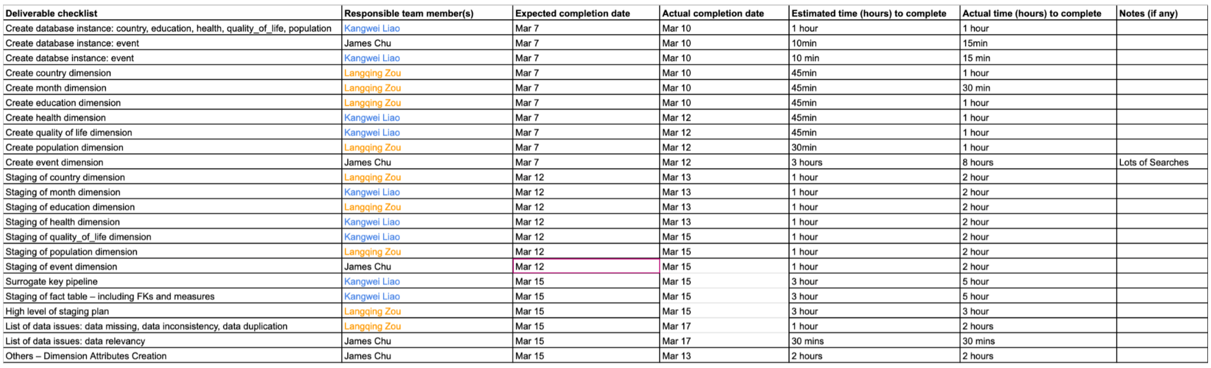
- Redundant or repeated copies of data are removed from a system
- Use UNIQUE keyword in SQL when creating instances

Data relevancy

When obtaining data from Wikipedia, we tried to select anything close enough to environmental, geopolitical and economic, for some countries, e.g Liberia, the data may not strictly be relevant enough since the resources are limited. Method to solve the issue(optional):

- Do more search then update the data if needed, or focus more on recent years' data, or alternative categories, e.g. societal problem

3. Teamwork plan



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: country, education, health, quality_of_life, population	Kangwei Liao	Mar 7	Mar 10	1 hour	1 hour	
Create database instance: event	James Chu	Mar 7	Mar 10	10min	15min	
Create databse instance: event	Kangwei Liao	Mar 7	Mar 10	10 min	15 min	
Create country dimension	Langqing Zou	Mar 7	Mar 10	45min	1 hour	
Create month dimension	Langqing Zou	Mar 7	Mar 10	45min	30 min	
Create education dimension	Langqing Zou	Mar 7	Mar 10	45min	1 hour	
Create health dimension	Kangwei Liao	Mar 7	Mar 12	45min	1 hour	
Create quality of life dimension	Kangwei Liao	Mar 7	Mar 12	45min	1 hour	
Create population dimension	Langqing Zou	Mar 7	Mar 12	30min	1 hour	
Create event dimension	James Chu	Mar 7	Mar 12	3 hours	8 hours	Lots of Searches
Staging of country dimension	Langqing Zou	Mar 12	Mar 13	1 hour	2 hour	
Staging of month dimension	Kangwei Liao	Mar 12	Mar 13	1 hour	2 hour	
Staging of education dimension	Langqing Zou	Mar 12	Mar 13	1 hour	2 hour	
Staging of health dimension	Kangwei Liao	Mar 12	Mar 13	1 hour	2 hour	
Staging of quality_of_life dimension	Kangwei Liao	Mar 12	Mar 15	1 hour	2 hour	
Staging of population dimension	Langqing Zou	Mar 12	Mar 15	1 hour	2 hour	
Staging of event dimension	James Chu	Mar 12	Mar 15	1 hour	2 hour	
Surrogate key pipeline	Kangwei Liao	Mar 15	Mar 15	3 hour	5 hour	
Staging of fact table – including FKs and measures	Kangwei Liao	Mar 15	Mar 15	3 hour	5 hour	
High level of staging plan	Langqing Zou	Mar 15	Mar 15	3 hour	3 hour	
List of data issues: data missing, data inconsistency, data duplication	Langqing Zou	Mar 15	Mar 17	1 hour	2 hours	
List of data issues: data relevancy	James Chu	Mar 15	Mar 17	30 mins	30 mins	
Others – Dimension Attributes Creation	James Chu	Mar 15	Mar 13	2 hours	2 hours	

PS: We also submitted the teamwork plan in a separate file.