数据库模型与设计

一．所涉问题分析

本数据库后端模型用于储存并分析数据集” A Global Database of COVID-19 Vaccinations”,该数据集收集了不同国家与地区跟踪全球疫苗推广的规模和速度。该数据集定期更新，包括所有有数据的国家（截至 2021 年 4 月 7 日，有 169 个国家）的疫苗接种总数、第一和第二剂量接种总数、每日疫苗接种率和人口调整覆盖率等数据。为了完成数据库的概念模型设计，我们需要对所涉及的数据集进行分析：

(1) . country_data

本数据集类型是以下四个数据集的集合：

Wale.csv	Canada.csv	United States.csv	Denmark.csv

这四个数据集具有相同的属性列：

location	date	vaccine	source_url	total_vaccinations	people_vaccinated	people_fully_vaccinated	total_boosters

通过对数据的分析，发现存在关系：

(**location,date**)->vaccine,source_url,total_vaccinations,people_vaccinated,people_fully_vaccinated,total_boosters

(**location**,**date)**之间不存在依赖关系,同时其他属性直接依赖于(**location,date**)因此这个关系本身属于 2NF。

推测：

total_vaccinations,people_vaccinated,people_fully_vaccinated,total_boosters 之间存在依赖关系，根据数据推测公式：

total_vaccinations=people_vaccinated+people_fully_vaccinated+total_booster

证明：

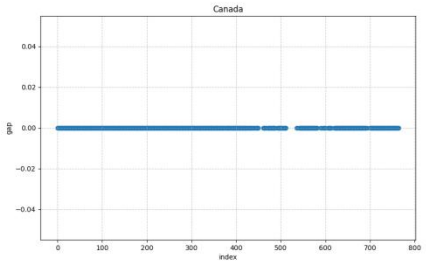
利用 python 的 pandas 和 numpy 库进行数据分析，并使用 matplotlib 库进行可视化。

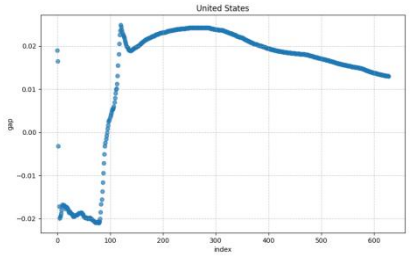
1.定义：

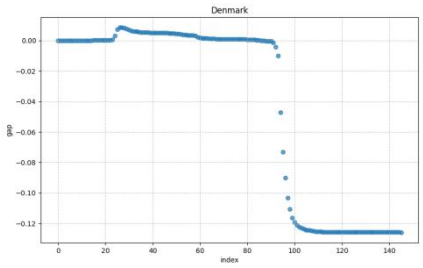
gap=(people_vaccinated+people_fully_vaccinated+total_booster-total_vaccinations)/total_vaccinations

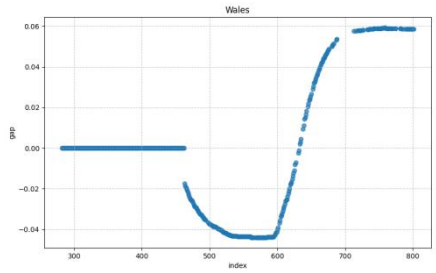
2.计算并可视化，结果如下：

	Canada	Denmark	United States	Wales
Average_gap	0.0	-0.041773288159879 67	0.0137814454205157 54	-0.0020181700985952 44









结论：对于这四个数据集，这个结论的平均误差约为 2%，在可接受范围内，因此在数据库关系设计时不包含需要将 total_vaccinations 与构成它的三个属性分离，消除数据冗余。

若在关系中消去 total_vaccinated，关系为：

(**location,date**)->vaccine,source_url,people_vaccinated,people_fully_vaccinated,total_boosters

所有非主属性直接依赖于复合主属性，非主属性之间不存在直接依赖与传递依赖，本关系是 3NF 的。

(2) . vaccinations-by-age-group.csv

本数据集包括如下属性:

location	date	age_group	people_vaccinated_per_hundred	people_fully_vaccinated_per_hundred	people_with_booster_per_hundred

由于可能存在在同一(**location,date**)对多个 age_group 的观测结果，但是对于(**location,date,age_group**)这一复合属性，对其他属性是唯一决定的，因此本表可归纳为关系：

(location,date,age_group)->people_vaccinated_per_hundred,people_fully_vaccinated_per_hundred,people_with_booster_per_hundred

未发现非主属性之间存在直接依赖或传递依赖于主属性，非主属性直接依赖于复合的主属性，该关系是 3NF 的。同时主属性之间不存在直接依赖和传递依赖关系。

(3) .vaccinations-by-manufacturer.csv

本数据集包括如下属性:

location	date	vaccine	total_vaccinations

由于可能存在在同一(**location,date**)对多个 vaccine 的结果，但是对于(**location,date,vaccine**)这一复合属性，对其他属性是唯一决定的，因此本表可归纳为关系：

(location,date,vaccine)->total_vaccinations

未发现非主属性之间存在直接依赖或传递依赖于主属性，非主属性直接依赖于复合的主属性，同时非主属性之间不存在直接依赖和传递依赖关系，该关系是 3NF 的。。

(4) .locations.csv

本数据集包括如下属性:

location	Iso_code	vaccines	last_observation_date	source_name	source_website

在分析中，可以发现以下的关系是存在的：**(location->iso_code) /(iso_code->location)**即出现了双向依赖，因此为了数据库消除冗余，需要选择一个作为数据库中多次出现的 location，

从语义上考虑，选择 location。
同时虽然 source_website 似乎与 source_name 也是双向依赖的，但是其实不是，同一个 source_name 名下存在多个 source_website,考虑到数据库的规范化，需要单独识别关系:
(**source_website**)->source_name
因此本数据集可识别为两个关系:
1. (**location,last_observation_date**)->vaccines,source_website
2. (**source_website**)->source_name
在这两个关系中，未发现非主属性之间存在直接依赖或传递依赖于主属性，非主属性直接依赖于复合的主属性，这两个关系都是 3NF 的。
(5) .vaccinations.csv
本数据集包含以下属性：

location	Iso_code	total_vaccinations	people_vaccinated
people_fully_vaccinated	total_boosters	daily_vaccinations_raw	daily_vaccinations
date	total_vaccinations_per_hundred	people_vaccinated_per_hundred	people_fully_vaccinated_per_hundred
total_boosters_per_hundred	daily_vaccinations_per_million	daily_people_vaccinated	daily_people_vaccinated_per_hundred

在分析中，location 与 iso_code 的双向依赖消除已经在上文解释，采用与 country_data 相同的方式检测假设是否成立:
total_vaccinations=people_vaccinated+people_fully_vaccinated+total_booster
结果如下:
average_gap=-0.024983135695691626
在合理的误差允许范围内，可以优化掉 total_vaccinations。
从语义的角度推测：
total_vaccinations_per_hundred=people_vaccinated_per_hundred+people_fully_vaccinated_per_hundred+total_booster_per_hundred
采取相同的方式进行验证：
Average_gap=-0.024623301531360664
在误差允许范围内，可以优化掉 total_vaccinations_per_hundred。
同时考虑到实际的语义和 daily_vaccinations_raw 明显存在较多缺失项 ，而 daily_vaccinations 代表实际的数据，在数据库存储中可以主要存储 dailt_vaccinations。
因此本数据集提取到的关系为：
(**location,date**)->people_vaccinated, people_fully_vaccinated, total_booster,daily_vaccinations,people_vaccinated_per_hundred,people_fully_vaccinated_per_hundred,total_booster_per_hundred,daily_vaccinations_per_million,daily_people_vaccinated,daily_people_vaccinated_per_hundred
(6) us_state_vaccinations.csv
本数据集包含以下属性：

location	date	total_vaccinations	people_vaccinated
people_fully_vaccinated	total_boosters	daily_vaccinations_raw	daily_vaccinations
total_distributed	total_vaccinations_per_hundred	people_vaccinated_per_hundred	people_fully_vaccinated_per_hundred

total_boosters_per_hundred	share_doses_used	daily_people_vaccinated	distributed_per_hundred

与之前的数据集处理一致，存在推测：

total_vaccinations=people_vaccinated+people_fully_vaccinated+total_booster

进行验证：

Average_gap=0.02328948156341341

total_vaccinations_per_hundred=people_vaccinated_per_hundred+people_fully_vaccinated_per_hundred+total_booster_per_hundred

进行验证：

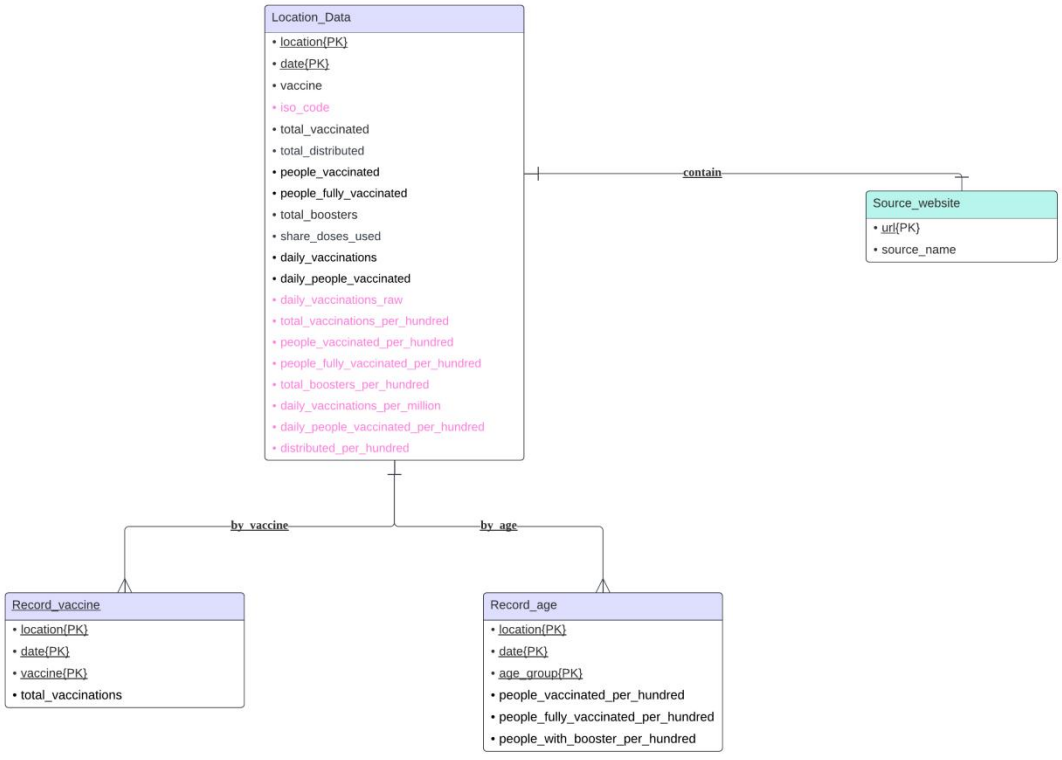
Average_gap=0.023235314511047416

因此可提取出关系:

(**location,date**)->people_vaccinated, people_fully_vaccinated, total_booster,daily_vaccinations, total_distributed, people_fully_vaccinated_per_hundred, total_boosters_per_hundred,share_doses_used,daily_people_vaccinated,people_vaccinated_per_hundred,distributed_per_hundred

二．数据库概念模型设计

使用实体关系(E-R)图进行概念模型设计：



考虑到数据残缺程度与数据重要性，在 ERD 图中将数据重要性与残缺程度较好的属性用黑色表示，其他属性存在相当程度的残缺以及存在较大的分析困难度，用浅色表示。

将 ERD 图映射为数据库模式：

使用七步映射法(7-step mapping process)进行 ER 图到数据库的映射：

1. 识别强实体：

强实体有：Record_normal,Record_age,Record_vaccine,Source_website。为每一个强

实体创建一个关系（表）：

表 1：Source_Website

属性	字段名	数据类型	是否为空/约束条件
url	URL	VARCHAR(50)	主键
source_name	NAME	VARCHAR(30)	

表 2：Location_Data

属性	字段名	数据类型	是否为空/约束条件
location	LOCATION	VARCHAR(50)	主键
date	DATE	DATETIME	主键
vaccine	VACCINES	VARCHAR(200)	可以为空
source_website	SOURCE_WEBSITE	VARCHAR(50)	非空
People_vaccinated	PEOPLE_VACCINATED	INT	可以为空
people_fully_vaccinated	PEPLE_FULLY_VACCINATED	INT	可以为空
total_boosters	TOTAL_BOOSTERS	INT	可以为空
daily_vaccinations	DAILY_VACCINATIONS	INT	可以为空
Iso_code	ISO_CODE	VARCHAR(30)	非空

daily_people_vaccinated	DAILY_PEOPLE_VACCINATED	INT	可以为空
share_doses_used	SHARE_DOSES_USED	FLOAT	可以为空
total_distributed	TOTAL_DISTRIBUTED	INT	可以为空
total_vaccinated	TOTAL_VACCINATED	INT	可以为空
Daily_vaccinations_raw	DAILY_VACCINATIONS_RAW	INT	可以为空
Total_vaccinations_per_hundred	TOTAL_VACCINATIONS_PER_HUNDRED	FLOAT	可以为空
People_vaccinations_per_hundred	PEOPLE_VACCINATIONS_PER_HUNDRED	FLOAT	可以为空
People_fully_vaccinations_per_hundred	PEOPLE_FULLY_VACCINATIONS_PERHUNDRED	FLOAT	可以为空
Total_boosters_per_hundred	TOTAL_BOOSTERS_PER_HUNDRED	FLOAT	可以为空
Daily_vaccinations_per_million	DAILY_VACCINATIONS_PER_MILION	INT	可以为空
Daily_people_vaccinated_per_hundred	DAILY_PEOPLE_VACCINATED_PER_HUNDRED	FLOAT	可以为空

Distributed_per_hundred	DISTRIBUTED_PER_HUNDRED	FLOAT	可以为空

表 3：Record_age

属性	字段名	数据类型	是否为空/约束条件
location	LOCATION	VARCHAR(50)	主键
date	DATE	DATETIME	主键
age_group	AGE_GROUP	VARCHAR(50)	主键
People_vaccinated_per_hundred	PEOPLE_VACCINATED_HUND	INT	可以为空
people_fully_vaccinated_per_hundred	PEPLE_FULLY_VACCINATED_HUND	INT	可以为空
total_boosters_per_hundred	TOTAL_BOOSTERS_HUND	INT	可以为空

表 4：Record_vaccine

属性	字段名	数据类型	是否为空/约束条件
location	LOCATION	VARCHAR(50)	主

			键
date	DATE	DATETIME	主键
vaccine	VACCINE	VARCHAR(50)	主键
total_vaccines	TOTAL_VACCINES	INT	非空

2. 识别弱实体：设计中未涉及弱实体
3. 转换关系：识别关系的基数如下：

   Location_Datal **contain** source_website 1:1

   Location_Data **by_vaccinations** Record_vaccine 1:N

   Location_Data **by_age_group** Record_vaccine 1:N

   将 Location_Datal 中的 source_website 标记为外键（Source_website 的主键）
4. 处理 M:N 关系：ERD 图中未涉及 M:N 关系
5. 处理 1：N 关系：在 Location_Data 中若设置外键 vaccine 与 age_group 缺乏数据支持，暂不涉及
6. 处理子类：未出现
7. 处理约束：已在数据分析阶段对数据约束进行了分析并消除了部分冗余

数据库标准化过程：

1. 对于表 Source_website 对应的关系:**(Source_website)**->Source_name:所有非主属性直接依赖于主属性，并且不存在对主属性的传递依赖，这个关系是 3NF 的。
2. 对于表 Location_Data:

(location,iso_code)在语义上是同一意义，存在双向依赖，违反了 2NF，因此单独建立表用于存储 loctaion 到 code 的映射，同时根据数据分析,存在依赖：

(people_vaccinated,people_fully_vaccinated,total_boosters)->total_vaccinated

(people_vaccinatedper_hundred,people_fully_vaccinated_per_hundred,total_boosters_per_hundred)->total_vaccinated_hundred

违反了 3NF，考虑到 total 数据在统计学上的重要性，另建表专门进行存储含有 total 意义的属性。同时另外建表处理重要性欠缺的属性，如含有'per'字段的属性。

此时所有表所有非主属性直接依赖于主属性，并且不存在对主属性的传递依赖，这个关系是 3NF 的。

3. 从方便查询的角度来说，可以单独维护一个 location->url 的映射来存储当前的数据来源，但是从可拓展性的角度来说，无法保证不同数据源提供的数据相同，因此未进行分表操作，就当前数据库的存储方案来说，此处确实存在一定的数据冗余，但是如果不能确定数据来源对每个国家都是 unique 的，不能进行分割操作。
4. 对于表 Record_age：

(location,date,age_group)->people_vaccinated_per_hundred,people_fully_vaccinated_per_hundred,people_with_booster_per_hundred

所有非主属性直接依赖于主属性，并且不存在对主属性的传递依赖，这个关系是 3NF 的。

5. 对于表 Record_vaccine:

(location,date,vaccine)->total_vaccinations

所有非主属性直接依赖于主属性，并且不存在对主属性的传递依赖，这个关系是 3NF 的。

考虑到 location 存在是否为 us 的情况，也需要另建表来存储为 us 的地区。

因此这个数据库涉及的所有关系都是 3NF 的。

考虑到 sql 的关键字，将 URL,NAME,LOCATION,DATE 后面加上’_’进行存储

三． Database Schema

数据库中包含以下表项：

表 1：Source_Website

属性	字段名	数据类型	是否为空/约束条件
url	URL_	VARCHAR(50)	主键
source_name	NAME_	VARCHAR(30)	

表 2：Record_vaccine

属性	字段名	数据类型	是否为空/约束条件
location	LOCATION_	VARCHAR(50)	主键
date	DATE_	DATETIME	主键
vaccine	VACCINE	VARCHAR(50)	主键
total_vaccines	TOTAL_VACCINES	INT	非空

表 3：Record_age

属性	字段名	数据类型	是否为空/约束条件
location	LOCATION_	VARCHAR(50)	主键
date	DATE_	DATETIME	主键
age_group	AGE_GROUP	VARCHAR(50)	主键

People_vaccinated_per_hundred	PEOPLE_VACCINATED_HUND	INT	可以为空
people_fully_vaccinated_per_hundred	PEPLE_FULLY_VACCINATED_HUND	INT	可以为空
total_boosters_per_hundred	TOTAL_BOOSTERS_HUND	INT	可以为空

表 4：Location_Data_Total

属性	字段名	数据类型	是否为空/约束条件
location	LOCATION_	VARCHAR(50)	主键
date	DATE_	DATETIME	主键
vaccines	VACCINES	VARCHAR(200)	可以为空
source_website	SOURCE_WEBSITE	VARCHAR(50)	非空，外键
total_vaccination	TOTAL_VACCINATION	INT	可以为空
total_distributed	TOTAL_DISTRIBUTED	INT	可以为空
Share_doses_used	SHARE_DOSES_USED	INT	可以为空

表 5：Location_Data_non_total

属性	字段名	数据类型	是否为空/约束条件
location	LOCATION_	VARCHAR(50)	主键
date	DATE_	DATETIME	主

			键
source_website	SOURCE_WEBSITE	VARCHAR(50)	非空/外键
People_vaccinated	PEOPLE_VACCINATED	INT	可以为空
people_fully_vaccinated	PEPLE_FULLY_VACCINATED	INT	可以为空
total_boosters	TOTAL_BOOSTERS	INT	可以为空
daily_people_vaccinated	DAILY_PEOPLE_VACCINATED	INT	可以为空
daily_vaccinations	DAILY_VACCINATIONS	INT	可以为空

表 6：US_Location

属性	字段名	数据类型	是否为空/约束条件
location	LOCATION_	VARCHAR(50)	主键

表 7：ISO_LOCTION

属性	字段名	数据类型	是否为空

			/约束条件
location	LOCATION_	VARCHAR(50)	主键
Iso_code	ISO_CODE	VARCHAR(50)	非空

表 8：Location_Data_Per

属性	字段名	数据类型	是否为空/约束条件
location	LOCATION_	VARCHAR(50)	主键
date	DATE_	DATETIME	主键
Daily_vaccinations_raw	DAILY_VACCINATIONS_RAW	INT	可以为空
Total_vaccinations_per_hundred	TOTAL_VACCINATIONS_PER_HUNDRED	FLOAT	可以为空
People_vaccinated_per_hundred	PEOPLE_VACCINATED_PER_HUNDRED	FLOAT	可以为空
People_fully_vaccinated_per_hundred	PEOPLE_FULLY_VACCINATED_PERHUNDRED	FLOAT	可以为空
Total_boosters_per_hundred	TOTAL_BOOSTERS_PER_HUNDRED	FLOAT	可以为空
Daily_vaccinations_per_million	DAILY_VACCINATIONS_PER_MILION	INT	可

			以为空
Daily_people_vaccinated_per_hundred	DAILY_PEOPLE_VACCINATED_PER_HUNDRED	FLOAT	可以为空
Distributed_per_hundred	DISTRIBUTED_PER_HUNDRED	FLOAT	可以为空

完成了对数据库的设计。