## 数据库模型与设计

### 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：**

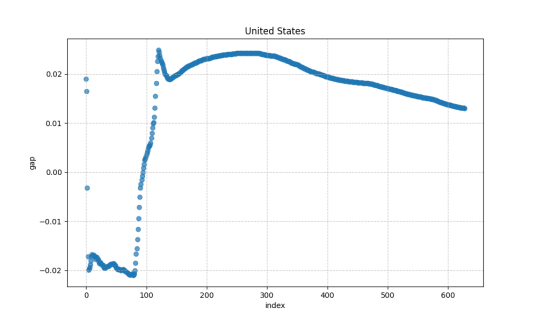
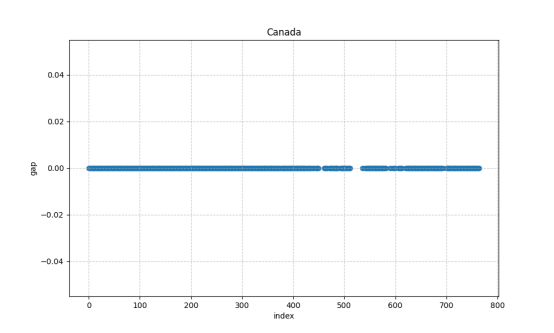
利用python的pandas和numpy库进行数据分析，并使用matplotlib库进行可视化。

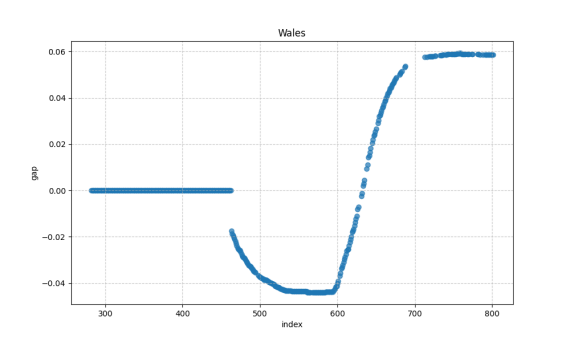
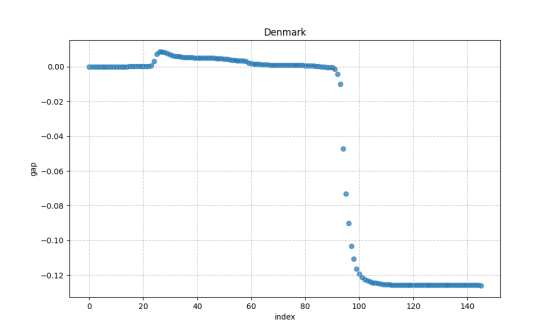
1.定义：

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 | -0.04177328815987967 | 0.013781445420515754 | -0.002018170098595244 |





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。

1. . 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。

1. .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。。

1. .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。

1. .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。

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。

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**)->,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

1. 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：

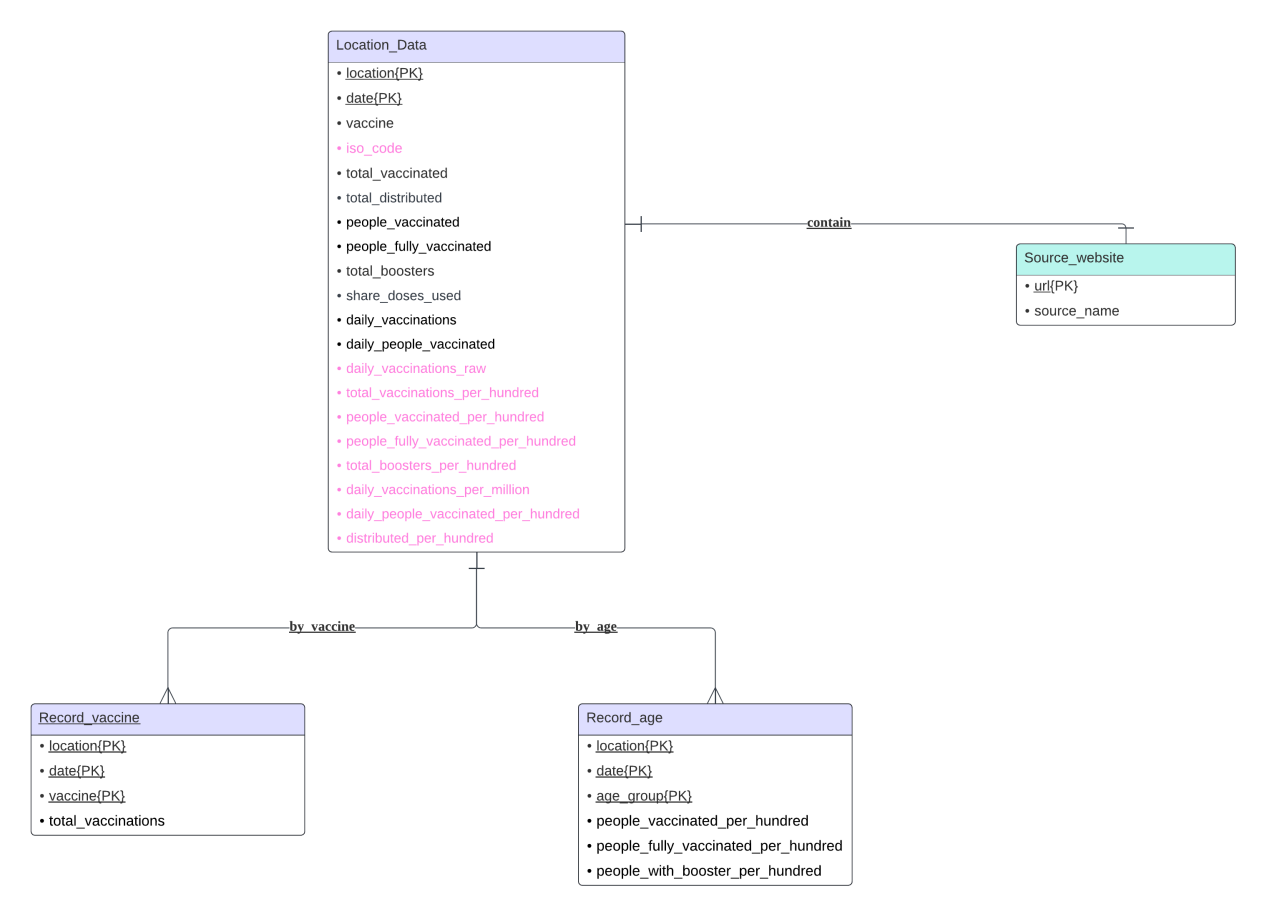
Average\_gap=0.023235314511047416

因此可提取出关系:

(**location,date**)->,daily\_vaccinations,share\_doses\_used,daily\_people\_vaccinated,people\_vaccinated\_per\_hundred,distributed\_per\_hundred

### 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 |

1. 识别弱实体：设计中未涉及弱实体
2. 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）

1. 处理M:N关系：ERD图中未涉及M:N关系
2. 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
3. 处理子类：未出现
4. 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。

1. 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。
2. 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。

1. 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

### 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。