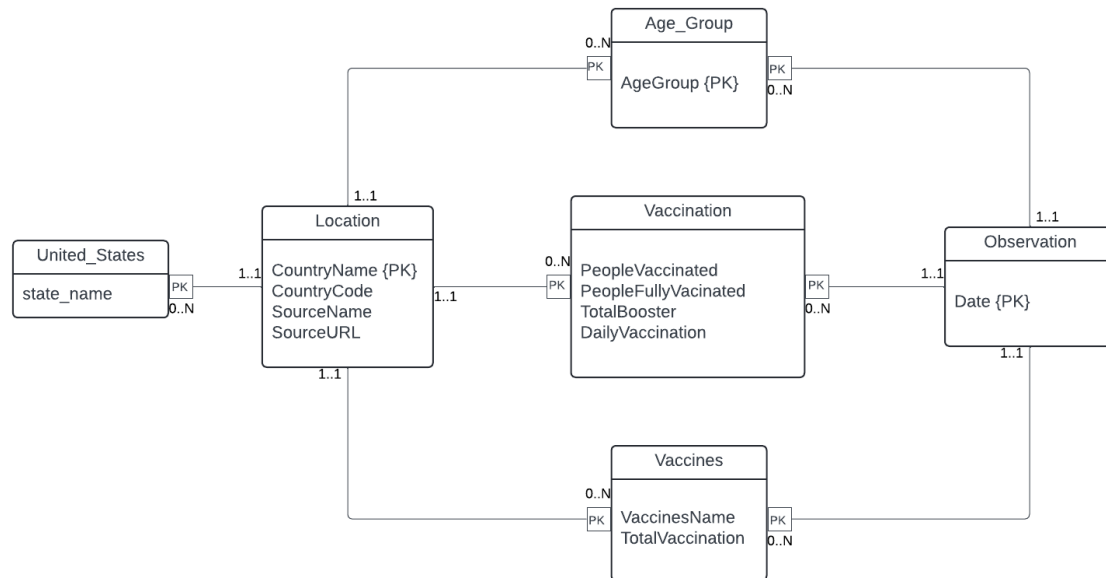# Assignment 4:

DataBase Design

ISYS1055

Part B: Designing the Database

1. ER Diagram

According to the files given, it can be seen that location name, code, and date are crucial attributes that act as keys to define other attributes such as people who got vaccinated for the first dose or second dose. Therefore, "Location" and "Observation" are two owner entities with the primary key "country_name" and "Date" respectively. Additionally, the source name and url attributes can be NULL due to some missing records for several locations.

First of all, the relation "Vaccination" includes "people_vaccinated", "people_fully_vaccinated", and "total_booster" as those are the three original attributes that are used to calculate other columns within the file named "Vaccination". It is a weak entity with the one-many cardinality as with a location there might be many records for vaccination, similar to the relationship with "Date".

Secondly, the relation "Age_Group" is a list of specific age groups separated to easily analyze the demographic. It is also a weak entity of "Location" and "Observation", with one-to-many cardinality for both relationships, indicating which age group was provided the vaccination per a location at a specific date recorded.

Similarly, the "Vaccines" dataset contains the "vaccines_name" such as Moderna or Pfizer along with the "total_vaccination", hence, the number of each vaccine types administered within that location at a specific date.

Last but not least, there is a special case that the United States of America is the only country that has the record divided into several states. Thus, the United States is assigned as a separate weak entity related to the owner entity "Location". Especially, besides the cardinality of one-to-many, the participation [0..1]

shows that not all "country_name" within "Location" will have a state, but all the states within entity "United States" are assigned with a "country_name" which is The United State in this case.

2. Mapping Schema

Step 1: Mapping strong entities

- o **Location** (<u>CountryName</u>, CountryCode, SourceName, SourceUrl)
- o **Observation** (<u>Date</u>)

Step 2: Mapping weak entities

- o **Vaccination** (<u>CountryName*</u>, CountryCode*, <u>Date*</u>, PeopleVaccinated, PeopleFullyVaccinated, TotalBoosted, DailyVaccination)
- o **Age_Group** (<u>CountryName*</u>, <u>Date*</u>, <u>AgeGroup</u>)
- o **Vaccine** (<u>CountryName*</u>, <u>Date*</u>, <u>VaccineName</u>, TotalVaccination)
- o **United_States** (<u>CountryName</u>, <u>StateName</u>)

Step 3: Map 1:1 relationships

(None)

Step 4: Map 1:N relationships

(None)

Step 5: Map M:N relationships

(None)

Step 6: Map multi-valued attributes

(None)

Step 7: Map higher-degree relationships

(None)

Step 8: Final Schema

- o **Location** (<u>CountryName</u>, CountryCode, SourceName, SourceUrl)
- o **Observation** (<u>Date</u>)
- o **Vaccination** (<u>CountryName*</u>, CountryCode*, <u>Date*</u>, PeopleVaccinated, PeopleFullyVaccinated, TotalBoosted, DailyVaccination)
- o **Age_Group** (<u>CountryName*</u>, <u>Date*</u>, <u>AgeGroup</u>)
- o **Vaccine** (<u>CountryName*</u>, <u>Date*</u>, <u>VaccineName</u>, TotalVaccination)
- o **United_States** (<u>CountryName</u>, <u>StateName</u>)

3. Normalisation

During the normalisation process, there are several challenges encountered that affect the final outcomes: redundancy, data integrity, and also anomalies in the event there are updates to the owner entities. Redundancy refers to the situation when there is a redundant functional dependency within a relation that must be removed or decomposed into a separate relation. Data integrity ensures the consistency of the dataset imported to the database, for example, any changes in the "Location" table must be propagated to all other tables that have a foreign key referencing an attribute of "Location" and reflect in the according table, otherwise, there are anomalies occur due to constraint violation.

In this circumstance, the above relations satisfy the requirement for the first normal form (1NF) as there are no repeating groups witnessed while atomic values are included for each column.

Moving onto the second normal form (2NF), it can be seen that almost all attributes depend on the whole key of each relation. For instance, within the "Location" relation, "CountryName" defines "CountryCode", "SourceName" and also "SourceUrl". Especially, those relations "Age_Group", "Observation" or "United_States" only contains primary keys and also in 2NF. However, there is one relation need to be decomposed to achieve 2NF, which is the "Vaccination" relation as "CountryCode" is not dependent on the whole key combined between "CountryName" and "Date" but the "CountryName" only.

- ~~**Vaccination** (CountryName*, CountryCode*, Date*, PeopleVaccinated, PeopleFullyVaccinated, TotalBoosted)~~
- **Vaccination** (CountryName*, Date*, PeopleVaccinated, PeopleFullyVaccinated, TotalBoosted)

According to the definition of the third normal form (3NF), non-key attributes must be dependent on nothing but the key. As a result, all relations are now in 3NF.

Final schema:

- **Location** (CountryName, CountryCode, SourceName, SourceUrl)
- **Observation** (Date)
- **Vaccination** (CountryName*, Date*, PeopleVaccinated, PeopleFullyVaccinated, TotalBoosted, DailyVaccination)
- **Age_Group** (CountryName*, Date*, AgeGroup)
- **Vaccine** (CountryName*, Date*, VaccineName, TotalVaccination)
- **United_States** (CountryName, StateName)