SDG 4: Quality Education

Datasets integrated:
1. Completion Rate
2. Dropout Rate

Source:

Entities (Datasets) and their attributes

Entity 1 / Dataset 1

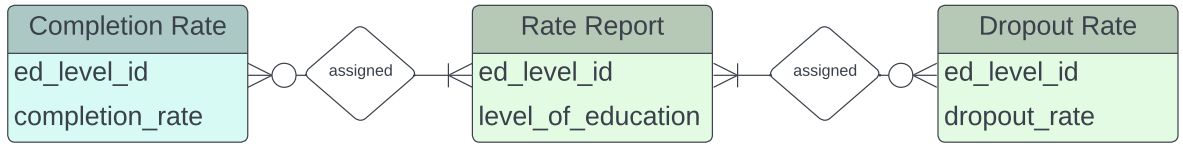
Completion Rate		
	geolocation	varchar
	level_of_education	varchar
	sex	varchar
	year	int
	completion_rate	text

Entity 2 / Dataset 2

Dropout Rate		
	geolocation	varchar
	level_of_education	varchar
	sex	varchar
	year	int
	dropout_rate	text

ERD for Dataset Integration

Simplified ERD using few atrributes,



Integrated Dataset

Rate Report		
PK,FK	ed_level_id	int
	geolocation	varchar
	level_of_education	varchar
	sex	varchar
	year	int
PK,FK	rate_type	int

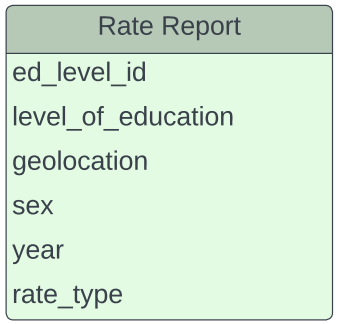
Completion Rate		
FK	ed_level_id	int
	geolocation	varchar
	level_of_education	varchar
	sex	varchar
	year	int
	completion_rate	text
FK	rate_type	int

Dropout Rate		
FK	ed_level_id	int
	geolocation	varchar
	level_of_education	varchar
	sex	varchar
	year	int
	dropout_rate	text
FK	rate_type	int

Note:
There are no PKs in each entity therefore I decided to create a primary key (ed_level_id) in the new dataset.
Completion rate and Dropout rate columns are initially assigned with text data type to avoid loading errors due to expected null values represented by undetermined notation.
The chosen datasets contains entries for Senior High School, completion rate and dropout rate was measured in 2018 onwards 2 years after the program implementation hence the the relationship between Reported Rate to Completion Rate AND Dropout Rate in this ERD is 0 or many unlike usual case of mandatory many.

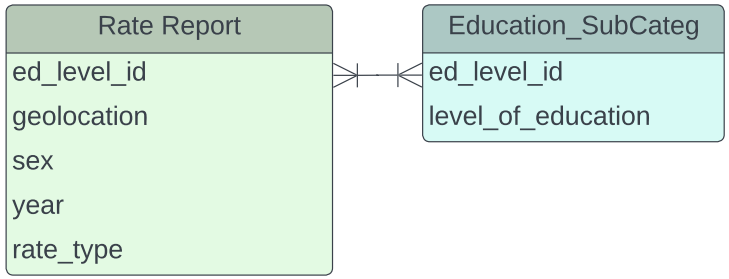
Normalized Model

1NF



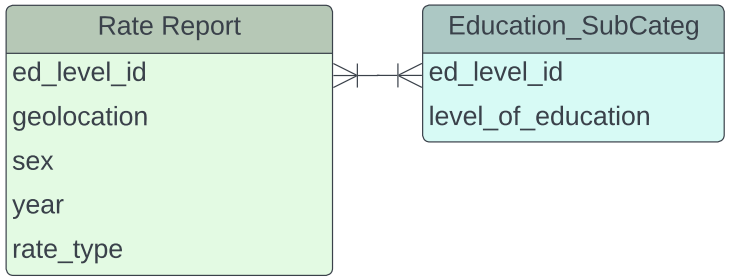
The integrated dataset does not contain any repeating values in a column hence already in first normal form

2NF



The level_of_education is dependent on ed_level_id and is partially dependent on the composite key, hence a new table is created in this case.

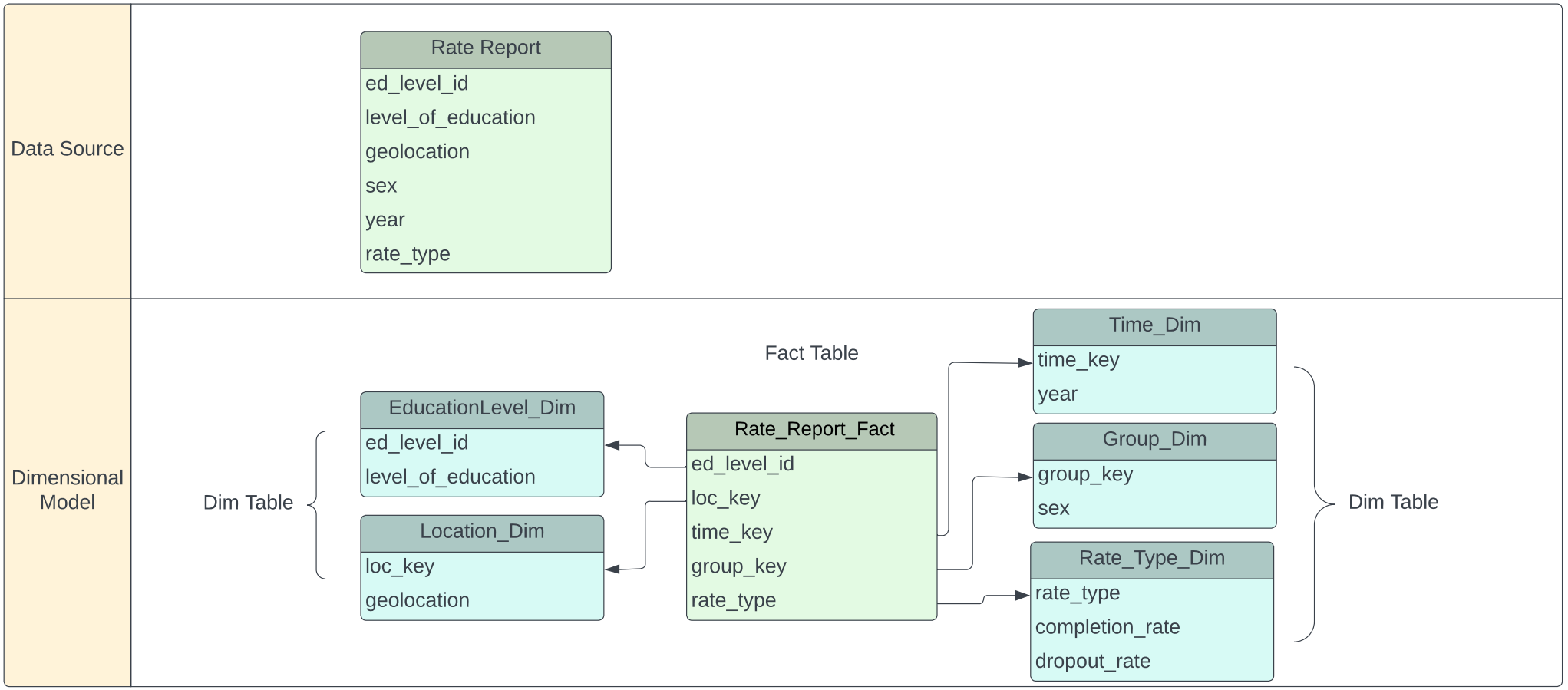
3NF



There are no transitive dependencies found in the table, hence the model is already in 3NF.

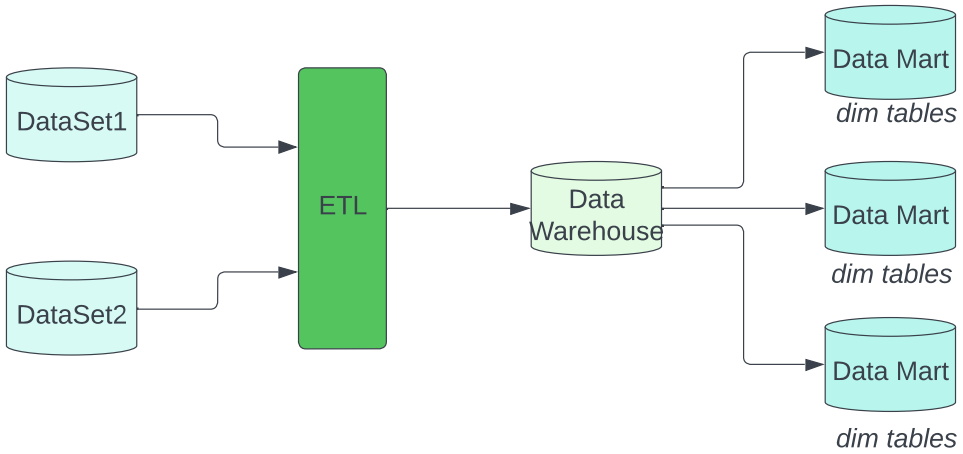
Note: Usually, the 3NF model would be much more complex if there are several columns in the dataset (i.e., business models).
Since the datasets from openstat under SDG are simplified and does not have several attributes, the generated 3NF looks simple.
For the later part and succeeding codes, dim tables will be treated as distinct queries as if to mimic data marts in business setting.

Dimensional Model



Note: Usually, the Dimensional model would be much more complex if there are several columns in the dataset (i.e., business models).
Since the datasets from openstat under SDG are simplified and does not have several attributes, the generated dimensional model looks simple due to lack of dimensional table families.
For the later part and succeeding codes, dim tables will be treated as distinct queries as if to mimic data marts in business setting.

General Illustration:



Codes:

DateSet1: Completion Rate
File Name: comp_rate.sql
Description: contains code for staging area, includes data cleaning and column insertion

DateSet2: Dropout Rate
File Name: drp_rate.sql
Description: contains code for staging area, includes data cleaning and column insertion

Integrated Dataset: Rate Report
File Name: rate_report.sql
Description: contains code used to create pre-structured table. contains combined values from dataset1 and dataset2

Normalized Model
File Name: normalized_model_extra_table.sql
Description: contains codes used to create the extra table housing the level_of_education (*refer to the normalized model*)

Dimensional Model
File Name: dimensional_model_dim_and_fact_table.sql
Description: contains codes used to create fact table and dimensional tables from the integrated dataset.