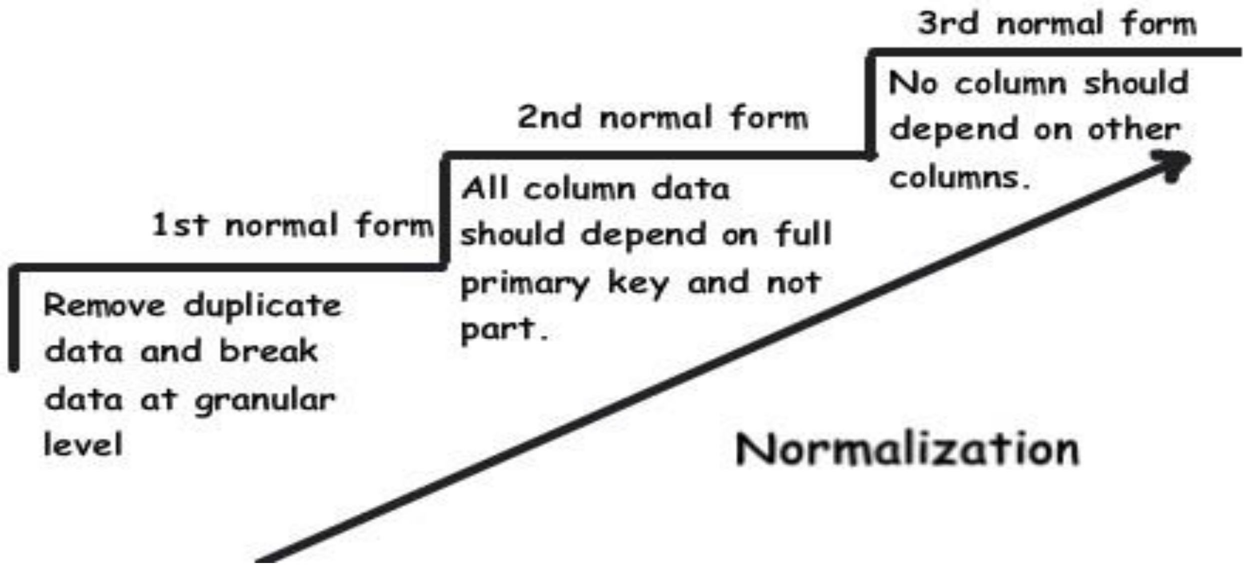
Data Normalization

Data normalization is generally considered the development of clean data. this process includes eliminating unstructured data and redundancy (duplicates) in order to ensure logical data storage.

Normalization rules divides larger tables into smaller tables and links them using relationships. The purpose of Normalization in SQL is to eliminate redundant (repetitive) data and ensure data is stored logically.


3rd normal form
No column should depend on other columns.
2nd normal form
All column data should depend on full primary key and not part.
1st normal form
Remove duplicate data and break data at granular level
Normalization

Database normalization is typically a refinement process after the initial exercise of identifying the data objects that should be in the relational database, identifying their relationships and defining the tables required and the columns within each table.

In SportsStats Dataset there are 2 csvs present namely noc_region & athlete_events. Columns present in noc_regions csv are: NOC, region & notes

Columns in noc_regions	Columns in athlete_events	
NOC	id	NOC
regions	name	games
notes	sex	year
	age	season
	height	city
	weight	sports
	team	events
	Medal	

First normal form (1NF): In 1NF I am dividing columns into 2 tables as T1 and T2

T1	T2	
Sports	Id	NOC
Events	Name	Games
City	Sex	Weight
Year	Age	Teams
Season	Height	Medal

Second normal form (2NF): In 2NF I am diving T1 into 3 tables and T2 into 3 tables

T1.1 Event	T1.2 Sport	T1.3 Game
Events	Sports	Year
City		Season
Sports		

We have eliminated game column because year and season column are enough for us to know which game it belongs to.

T2.1 Athlete		T2.2 NOC	T2.3 Team
Id	Weight	NOC	Teams
Name	Height	Region	NOC
Sex, Age	Events		
Medal	Teams		

Third normal form (3NF):

EventDetail	
Ed_id	Primary Key
A_id	F Key
E_id	F Key

GameDetail	
Gamed_id	Primary Key
Game_id	
E_id	

Team	
Tead_id	Primary Key
NOC_id	F key
Teams	

NOC	
NOC_id	Primary Key
NOC	
Region	

Athlete	
A_id	Primary Key
Name	
Sex	
Age	
Height	
Weight	
Medal	
Team_id	F Key

Event	
E_id	Primary Key
Events	
City	
Sport_id	F Key

Sport	
Sport_id	Primary Key
Sports	

Game	
Game_id	Primary Key
Year	
Season	

In third normalization form, I have divided columns in such a way that there is no transitive dependency. In above tables, all the non-key attributes are now fully functional, dependent only on the primary key.