A Short Introduction to Databases

6th Edition

Viji Kumar

Copyright © Viji Kumar 2020

Book 1 – Modelling Data

Chapter 1 – Explanatory Framework

The aim of this book is to explain how discrete items of information ***(data)*** e.g. names of individuals or organisations can be linked to other items of information e.g. addresses, dates of birth, or amounts invoiced. How that linked data can be used to drive other processes or produce reports is illustrated here by creating and populating a database of football *(soccer)* league results, a world-wide phenomenon and I hope, explicable to most adults *(young and old)*.

Here the word ***database*** will refer to any collection of linked data in any structured format e.g. in rows and columns. The structure of a database, its form, will depend on its function e.g. the types of reports that the users of the database have specified. The purposes for which data are collected are myriad and it is prudent to bear in mind that there almost certainly will be data protection legislation to regulate the use of some types of data.

To report on football games, you compare the number of goals scored by each team and deduce the result. In this model the two participating teams are labelled *team one* and *team two*. The starting point is the cataloguing of the different types of entities required to model the progress of a league season. This model only uses the following types of entities; competitions, league seasons, teams, venues, league games and league goals but it is an extensible model.

It is also necessary to allow for the possibility that these teams may be relegated or promoted at the end of a season and that they may also play in other types of competitions e.g. knock-out tournaments. The entity types additionally required to extend the model would then include qualification paths, group games, knock-out games, group goals and penalty shoot-out goals inter alia. These additional entity types indicate that there will need to be variants of some types of entities to model other types of competitions, e.g. the different types of games, qualification paths, and goals. Carolus *(Father of Taxonomy)* Linnaeus's words below may seem a trifle overwrought in this context, but they nevertheless are an extremely useful insight.

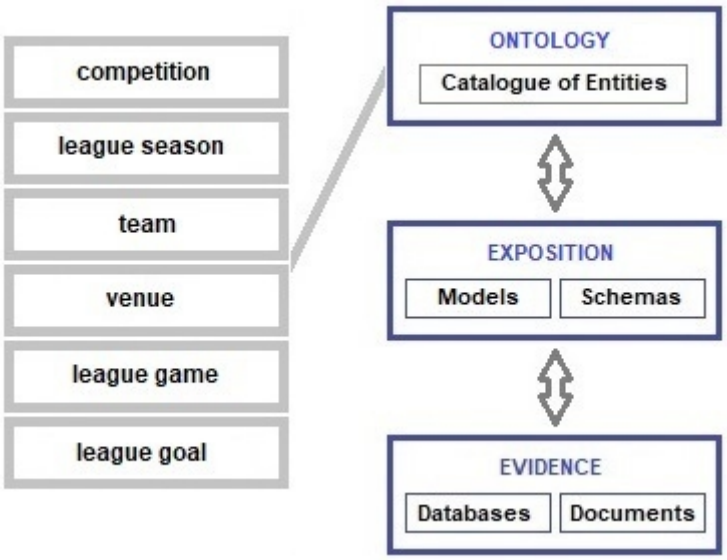
The first step in wisdom is to know the things themselves; this notion consists in having a true idea of the objects; objects are distinguished and known by classifying them methodically and giving them appropriate names. Therefore, classification and name-giving will be the foundation of our science.

Systema Naturae (1735), trans. M. S. J. Engel-Ledeboer and H. Engel (1964)

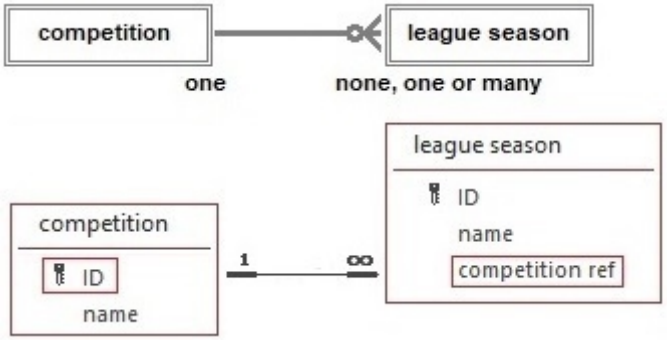
The initial analysis should produce data models and/or schemas, also known in the demotic as the ***metadata***. The outcome of that analysis is the specification of a catalogue of linked entities about which data will have to be collected. Entities are constructs that can be defined as distinct and with a separate existence e.g. an individual, a credit card or a league game. Data are the attributes of different types of entities, e.g. an individual's name, details of the transactions carried out using a particular debit card, the start time and venue of a game of football or the time a goal is scored. Entities can also refer to separate, definable, and distinct constructs that have no material existence

e.g. competitions, families, or league seasons. In this model all abstract entities have links, either directly or indirectly, to material entities e.g. all league games are associated with a particular league season of a particular competition. The types of data used in this exercise will be strings of text *(e.g. names)*, dates/times and integers.

League competitions are decided by the accumulation of points awarded to teams during the course of a season e.g. 3 points for a victory and a point for a draw. To effectively report on a league, a table summarising the results and ranking the teams based on the points total is required. All the developmental stages of a database, from creating the data model to testing the model with sample data, will be discussed.



The relationship or the link between any two entities is represented by a line starting at entity type 1 (e.g. competition) and ending with a nought and a crow's foot at entity type 2 (e.g. league season). This type of link specifies that there may be none, one or many type 2 entities associated with a single type 1 entity in a specific role and that the type 2 entities must be linked to one and only one type 1 entity e.g. there may be none, one or many league seasons related to a particular competition. An alternate depiction of the link *(MS Access)* is also provided below showing the mechanism that enforces the relationship.



Only 4 of the relationships shown below are necessary to deduce the result of a league game and generate the league table. The venues for league games are recorded only because such information

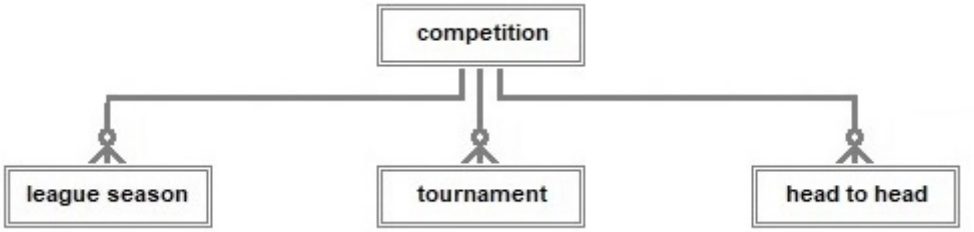
may be of interest to some sports fans and to enable statistics to be generated about a team's performance at a particular venue over different league seasons. These relationships are implemented in a database by linking the entity types e.g. either of the two participating teams may score a goal during the course of a league game.

The data model and the sample reports should enable a non-database specialist to understand what data are being collected and how that data will be managed. The data model showing the relationships between the entity types is also specified as a machine-readable schema using a markup language, XML. The XML schema can be safely ignored if the data model makes sense. The XML schema is additionally provided in the same helpful spirit that the Rosetta Stone was used to publish a decree about Ptolemy V's reign in 3 different scripts *(Egyptian hieroglyphs, Demotic and Ancient Greek)*. That thoughtful act subsequently helped Thomas Young and Jean-Francois Champollion decipher Egyptian hieroglyphs.

There is a large amount of historical data available to test models purporting to explain league competitions. The raw data necessary to test the models can be found in existing databases or in documents such as newspapers and books. Constructing and populating databases in accordance with this model should produce databases amenable to being queried to obtain the results of league games and to enable the subsequent generation of league tables. The detailed model will define each type of entity by listing the attributes that are of interest e.g. the names of competitions, teams and venues or the date, time, participating teams, and venue of a game.

This data model can be extended to include two other types of competitions, multi-phase tournaments and head to head competitions *(i.e. between just two teams)*. Head to head competitions include Test series and competitions where the two participants qualify by winning other competitions e.g. the English Community Shield and the Spanish Super Cup. Furthermore, a competition may have both, a league season followed by a tournament to decide the prizes for the season. The English Football League Championship has had in recent years, a *(play-off)* tournament subsequent to the league

season to decide the third team to be promoted to the Premier League. The top tier of professional Rugby League in England also employs play-offs.



At a minimum, the events of interest during a soccer league game are the events required for the deduction of the result, i.e. the goals. For every goal scored, the type of goal *(open play, penalty, or own goal)* is noted in addition to the number of elapsed minutes *(rounded up)* from the start of the game. Each goal is treated as a separate entity to allow extensions to record the scorers if required. For the sake of simplicity, the venerable convention of recording all injury time goals as being scored in the last minute of the relevant half is observed, enabling half time, full time, and extra time scores to be generated separately. The disadvantage of observing this convention is that the duration of injury time and the accurate timings of injury time goals are not recorded. Other data not considered here include the starting line-up for games, disciplinary events, goal scorers, missed penalties and substitutions but the model can be extended. That sums up the remit of this league model.

Entities must not be multiplied beyond necessity

In keeping with the imperative expressed by Occam’s razor, a principle advocating parsimony, an attempt has been made to keep to a minimum the number of entity types required for the explanation of this model.

Chapter 2 – Definition of Entity Types

Two attributes, the ***ID*** and the ***name*** are sufficient to identify or refer to any competition, team, or venue in this model. While these entities have many other attributes that may be of interest to sports fans, e.g. the founding date of a team or the current manager or coach; such attributes have not been included here for the sake of simplicity.

ID	name
1001	English Premier League
1002	FIFA World Cup
1003	FA Challenge Cup
1004	UEFA European Championship
1005	UEFA Champions League
1006	Football League Championship
1007	Charles Darwin League

competition

The ID is a number that has been assigned to each of these entities for use as a unique and permanent identifier because the name may change *(humans are capricious)* e.g. what was the European Champions Clubs' Cup or European Cup is now the UEFA Champions League. The Charles Darwin League is a fictitious example as are most of the teams and all of the venues.

ID	name
1001	Lions
1002	Tigers
1003	Eagles
1004	Falcons
1005	Dolphins
1006	Sharks
1007	Iceland
1008	Sheffield FC

team

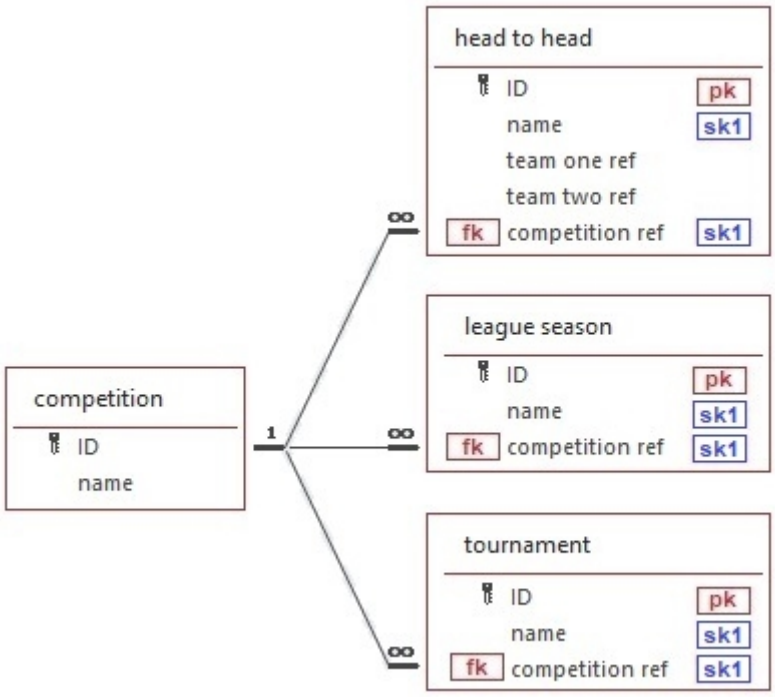
ID	name
1001	Richard Dawkins Stadium
1002	James Clerk Maxwell Stadium
1003	Rosalind Franklin Stadium
1004	Maurice Wilkins Stadium
1005	Greta Thunberg Stadium
1006	Alfred Russel Wallace Stadium
1007	Jocelyn Bell Burnell Stadium

venue

Competitions, teams, and venues are defined with both their attributes as keys denoting that they must be unique. The ID is the designated unique identifier and in database jargon is called the ***primary key***. It is used by data management systems to uniquely identify a particular entity. The name, which may change, also has to be unique in this context because users of the databases may wish to search for results and other statistics about a particular competition, team or venue using the name. To prevent duplicate names, the name attribute is designated as a ***secondary key***. This enables database designers to know that they must check for uniqueness before recording a name or an error may occur.

To cater for searches using any previous names an entity may have had, the model will need to be extended. To allow searches using an entity's previous name will require the maintenance of records of previous names and a link to the entity via the associated unique identifier *(in 2020, searching for "UEFA Cup" on the web will return "UEFA Europa League" related links at the top of the list)*.

All league seasons, tournaments and head to heads must be linked to a competition. This is where the diagrams translate into linked tables. Therefore, in addition to an ID and a name, these tables have a reference to a competition using the competition's ID.



These attributes are called ***foreign keys*** in the tables of league seasons, tournaments, and head to heads. Unlike the other 2 types of competitions, head to head competitions are defined allowing for the two participating teams to be recorded but the identities of the teams may not be known when a particular head to head tournament is organised and can be left empty. The names of the league seasons, tournaments and head to heads may not be unique except when coupled with the competition reference. The secondary key specified is now a composite key that comprises both, the competition foreign key in addition to the name, to ensure that the names of the league seasons, tournaments and head to heads are unique for each competition. The two versions of the league season table below illustrate the replacement of the reference ID with the name to produce more meaningful reports.

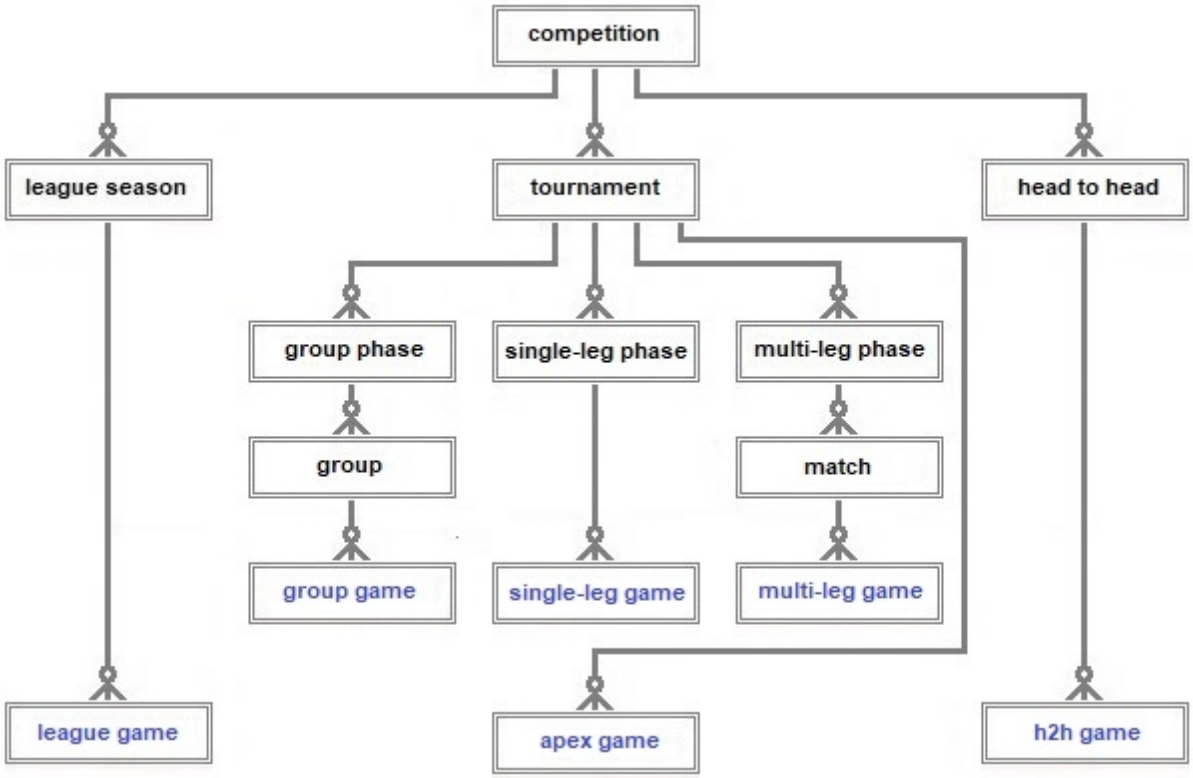
ID	name	competition ref
1001	2020/21	1001
1002	2020/21	1006
1003	2020	1007
1004	2021	1007

league season

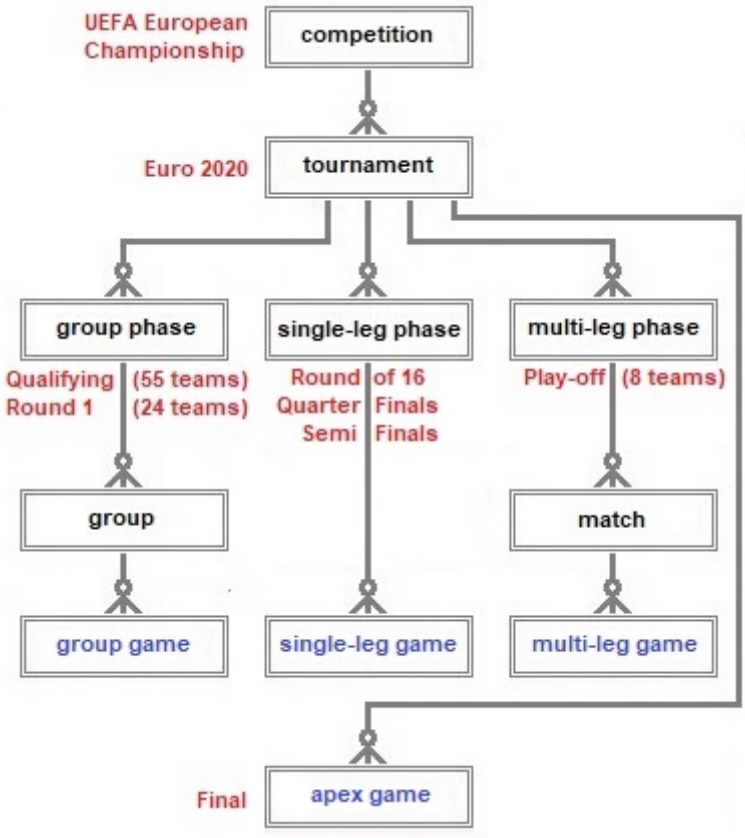
ID	name	competition
1001	2020/21	English Premier League
1002	2020/21	Football League Championship
1003	2020	Charles Darwin League
1004	2021	Charles Darwin League

A league season *(introduced by William McGregor of Aston Villa in 1888 to English football)* has a specific number of games related to it depending on the number of meetings between any pair of the participating teams. The English Premier League has two games *(home and away)* between each pair of the 20 participating teams with a total of 380 games per season whereas the rugby union Six Nations Championship only has a single game between each pair, a total of 15 games per season. Another round robin format, tournament groups, often have four teams playing a total of 6 games when the teams meet once *(FIFA World Cup)* and 12 games when they meet twice *(UEFA Champions League)*.

There are some competitions that apply rules that require amendments to the usual way outcomes are calculated. The Scottish Premier League has employed an unorthodox mechanism for ensuring that 12 teams play 38 games each during a league season. After three meetings between all of the 12 participating teams, the SPL is treated as two separate leagues of 6 teams based on their league positions. There is then a single game between each of the six teams in each of the two leagues, a total of 228 league games for an SPL season. After the split, the SPL season becomes 2 separate leagues of 6 teams playing each other once, but with the teams retaining their points from before the split. The decision as to which of the teams play at home after the split is decided, presumably, from reviewing the 3 previous meetings between any two teams.

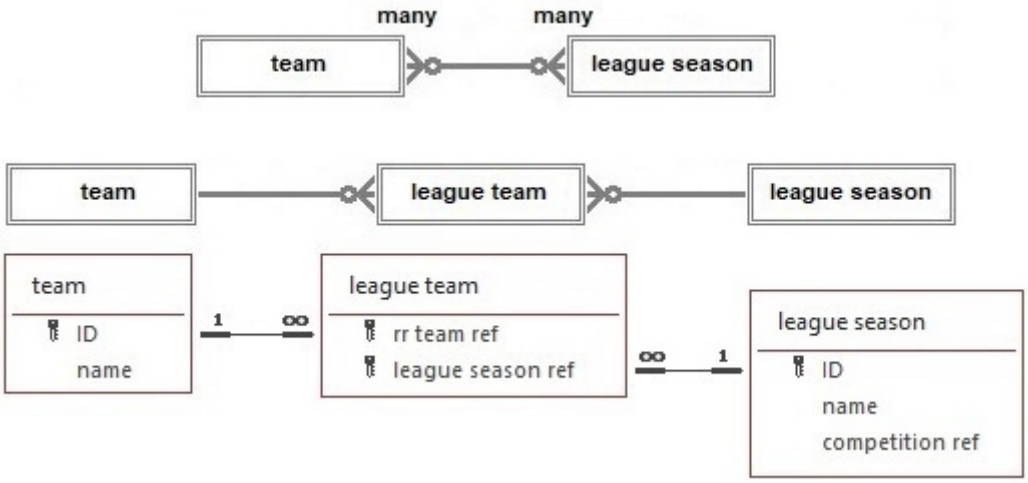


A tournament, as defined here, starts with four or more teams and then whittles the teams down with each successive phase, e.g. the 2018 FIFA World Cup in Russia started with a group phase *(32 teams)*, followed by 3 single-leg knock-out phases *(16, 8 and 4 teams)* ending with the final and third-place games. Tournament prizes are decided by the final and sometimes, the third-place game and/or the plate final, collectively known here as apex games. *(A plate tournament is a parallel tournament for teams eliminated at an early phase.)* Apex games are different from other types of games because there can only be one final, third-place decider, or plate final for a tournament. There is no option modelled here for tournaments to be decided by multi-leg apex matches as they are rare these days. As far as I know, the last time a professional European football tournament was decided by a multi-leg apex match was the 1997 UEFA Cup final between Schalke and Inter Milan over two games. If multi-leg apex matches are required, the model must be further extended. The extended model is used below to describe the 2018 FIFA World Cup and the 2020 UEFA European Championships.



Returning to the detail of the league model, the participation of teams in league seasons is an example of a many-to-many linkage. A league season may have none, one or many participating teams depending on whether it is in the future, present or past and a team may have participated in, be participating in or be eligible to participate in none, one or many league seasons.

The decomposition of that single many-to-many link into 2 one-to-many links is achieved by the creation of an intermediate entity, the ***league team***. It has just two attributes and they are also foreign keys. The primary key of the league team type of entity is composed of the two foreign keys referencing the league season and team entities.



rr team	*league season*
Tigers	**2020**
Eagles	**2020**
Eagles	**2021**
Falcons	**2020**
Falcons	**2021**
Dolphins	**2020**
Dolphins	**2021**
Sharks	**2021**

league team

The participants for the Charles Darwin League's 2020 and 2021 seasons are shown above. In August 2020 the Tigers were relegated, and the Sharks replaced them. Examples of how to specify the qualification criteria for a subsequent league season, e.g. league to league qualifiers *(Football League Championship to the English Premier League)* or apex games to league qualifiers *(Championship play-off to the Premier League)* will be provided. In the top two tiers of English professional football, the first seventeen teams from a Premier League season and the first two from a Championship season play in the next Premier League season. The qualification path for the twentieth team is via the Championship play-off. The 3 teams relegated from the Premier League play in the subsequent Championship season as does the losing play-off finalist.

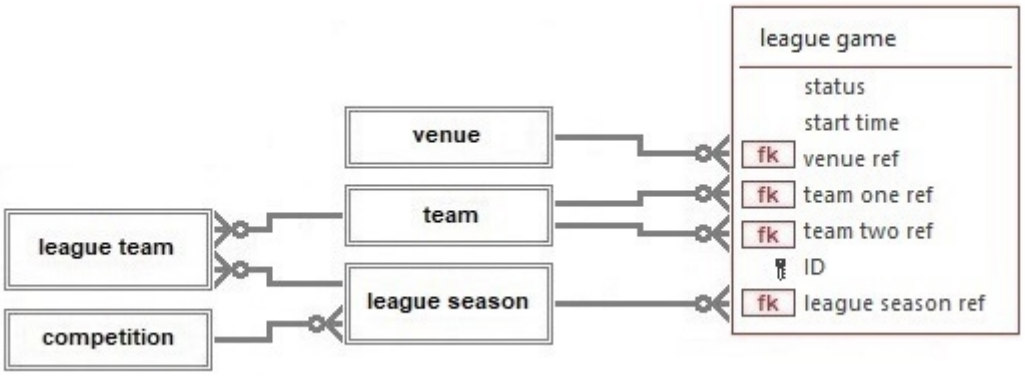
Chapter 3 – Round Robin Model

This chapter aims to explain how to construct a round robin competition model using the 2020 season of the Charles Darwin League *(CDL)* for illustration. The CDL is a four-team league where each team plays each of the other teams twice *(home and away)* during the season. The game results and the league standings at the completion of the 2020 season are shown below.

team one	*tally one*	*tally two*	*team two*
Eagles	1	1	Dolphins
Tigers	0	2	Falcons
Dolphins	2	1	Tigers
Falcons	1	1	Eagles
Eagles	0	0	Falcons
Tigers	3	2	Dolphins
Dolphins	3	1	Eagles
Falcons	2	0	Tigers
Eagles	1	0	Tigers
Dolphins	1	2	Falcons
Tigers	1	2	Eagles
Falcons	0	0	Dolphins

name	*pld*	*win*	*drw*	*los*	*pts*	*GD*	*for*	*agn*
Falcons	6	3	3	0	12	5	7	2
Eagles	6	2	3	1	9	0	6	6
Dolphins	6	2	2	2	8	1	9	8
Tigers	6	1	0	5	3	-6	5	11

For this exercise, the collection of the goal scoring data for any game is assumed to be in one of the following three mutually exclusive states; **not applicable** (*to be played or the score is not known*), **incomplete** (*game in progress with all known scores*) and **complete** (*independently verifiable score at the end of the game*). This rudimentary control mechanism *(the **status** attribute)* for specifying the extent of the known scoring data about a game should suffice for this explanation but included among its many shortcomings are that it neither allows the recording of abandoned games nor can it record the fact that there may have been temporary breaks in play due to unforeseen events during a game.



The starting time and the venue reference are also common attributes for all types of games though they are unnecessary for the deduction of the result. The two participating team references are attributes of league games but are not common to all types of games in the extended model e.g. head to head games. The league season reference is the only attribute specific to a league game. A unique identifier, ID, is the prime key and the two team references, the venue reference and the league season reference are all foreign keys. The tables below show two versions of the league game table at the conclusion of the CDL 2020 season. The first shows the data recorded and the second is a report for the benefit of humans.

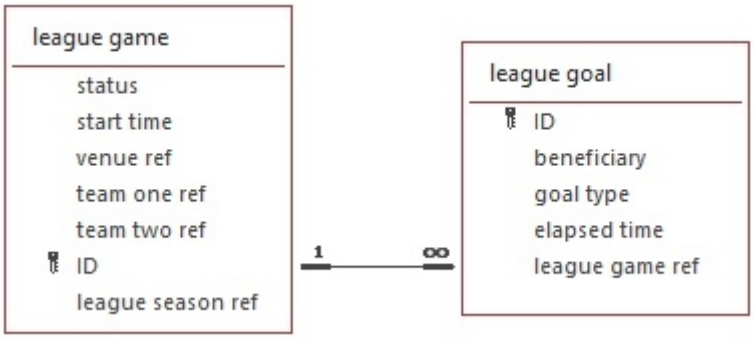
status	*start time*	*venue ref*	*team one ref*	*team two ref*	*ID*	*league season ref*
complete	**04/07/2020 15:00**	**1001**	**1003**	**1005**	**1001**	**1003**
complete	**04/07/2020 15:00**	**1002**	**1002**	**1004**	**1002**	**1003**
complete	**11/07/2020 15:00**	**1003**	**1005**	**1002**	**1003**	**1003**
complete	**11/07/2020 15:00**	**1004**	**1004**	**1003**	**1004**	**1003**
complete	**18/07/2020 15:00**	**1001**	**1003**	**1004**	**1005**	**1003**
complete	**18/07/2020 15:00**	**1002**	**1002**	**1005**	**1006**	**1003**
complete	**25/07/2020 15:00**	**1003**	**1005**	**1003**	**1007**	**1003**
complete	**25/07/2020 15:00**	**1004**	**1004**	**1002**	**1008**	**1003**
complete	**01/08/2020 15:00**	**1001**	**1003**	**1002**	**1009**	**1003**
complete	**01/08/2020 15:00**	**1003**	**1005**	**1004**	**1010**	**1003**
complete	**08/08/2020 15:00**	**1002**	**1002**	**1003**	**1011**	**1003**
complete	**08/08/2020 15:00**	**1004**	**1004**	**1005**	**1012**	**1003**

league game

start time	venue	team one	team two	ID	league season
04/07/2020 15:00	Richard Dawkins Stadium	Eagles	Dolphins	1001	2020
04/07/2020 15:00	James Clerk Maxwell Stadium	Tigers	Falcons	1002	2020
11/07/2020 15:00	Rosalind Franklin Stadium	Dolphins	Tigers	1003	2020
11/07/2020 15:00	Maurice Wilkins Stadium	Falcons	Eagles	1004	2020
18/07/2020 15:00	Richard Dawkins Stadium	Eagles	Falcons	1005	2020
18/07/2020 15:00	James Clerk Maxwell Stadium	Tigers	Dolphins	1006	2020
25/07/2020 15:00	Rosalind Franklin Stadium	Dolphins	Eagles	1007	2020
25/07/2020 15:00	Maurice Wilkins Stadium	Falcons	Tigers	1008	2020
01/08/2020 15:00	Richard Dawkins Stadium	Eagles	Tigers	1009	2020
01/08/2020 15:00	Rosalind Franklin Stadium	Dolphins	Falcons	1010	2020
08/08/2020 15:00	James Clerk Maxwell Stadium	Tigers	Eagles	1011	2020
08/08/2020 15:00	Maurice Wilkins Stadium	Falcons	Dolphins	1012	2020

league game

For league games, the score will only involve goals during the 90 minutes as penalty shoot-outs are not utilised. A reference to the league game is the only attribute specific to a league goal. The attributes common to all goals in the extended model are an ID, the elapsed time, the beneficiary *(i.e. either team one or team two)* and the type of goal *(open play, penalty, and own goal)*. League goals scored during stoppage time are recorded as being scored in the 45th or 90th minute.



The details of the 27 league goals scored during the 2020 CDL season are shown below. If the scorer of a goal is required, a reference to the individual can be added for each goal. If team sheets are also recorded, the beneficiary can then be deduced using the team sheets, the scorer, and the type of goal. The beneficiary attribute is then unnecessary and should be removed. That is all the data required in this model to produce the league game results and league tables for football competitions.

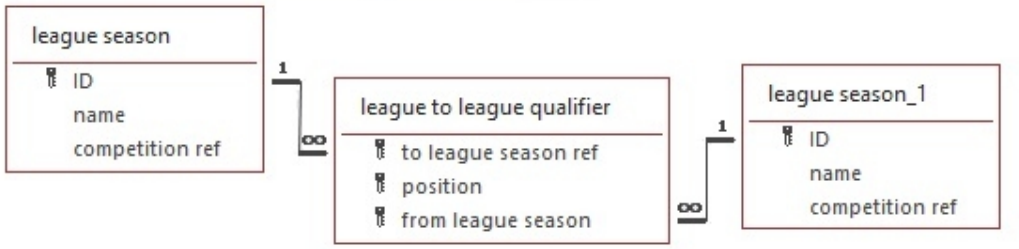
ID	beneficiary	goal type	elapsed time	league game ref
1001	team two	open play	12	1002
1002	team two	penalty	43	1001
1003	team one	own goal	45	1001
1004	team two	penalty	64	1002
1005	team two	own goal	7	1003
1006	team one	open play	31	1004
1007	team two	open play	55	1004
1008	team one	penalty	65	1003
1009	team one	open play	82	1003
1010	team one	open play	9	1006
1011	team two	penalty	10	1006
1012	team one	own goal	34	1006
1013	team one	open play	74	1006
1014	team two	open play	83	1006
1015	team one	own goal	19	1007
1016	team one	penalty	23	1007
1017	team one	open play	41	1008
1018	team two	open play	53	1007
1019	team one	open play	67	1008
1020	team one	open play	85	1007
1021	team two	own goal	8	1010
1022	team one	open play	21	1009
1023	team one	penalty	38	1010
1024	team two	open play	71	1010
1025	team two	open play	7	1011
1026	team two	open play	37	1011
1027	team one	penalty	54	1011

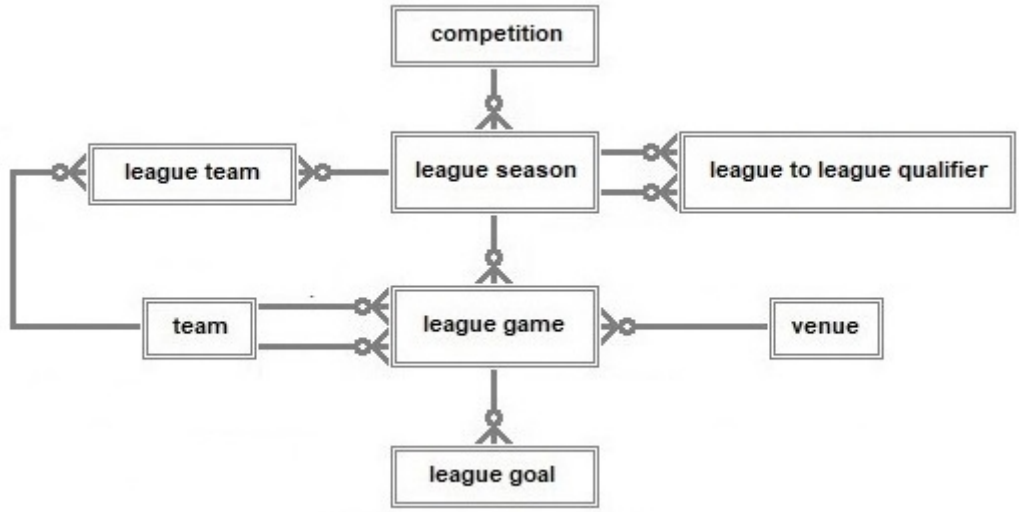
league goal

The final rankings at the end of a league season or the outcome of a knock-out game *(winner/loser)* can be used to specify the criteria for qualification for a subsequent league season or tournament. The specification of qualification paths e.g. from the Football League Championship *(FLC)* play-off final 2019/20) to the English Premier League *(EPL)* 2020/21 season or from the EPL 2019/20 season to the FLC 2020/21 season is discussed in Chapter 5.

to league season	position	from league season
2021	1	2020
2021	2	2020
2021	3	2020

league to league qualifier





Shown below is an extended version of the model with the qualification paths from both the FLC play-off final 2019/20 and the EPL 2019/20 season to the EPL 2020/21 season provided. Note the addition of the apex penalty shoot-out results that may be required for the play-off final.


English Premier League
Football League Championship
competition
EPL 2019/20
FLC 2019/20
EPL 2020/21
league season
FLC 2019/20 Play-off
tournament
league to league qualifier
apex to league qualifier
league team
league game
team
apex game
FLC 2019/20 Final
league goal
apex goal
apex so penalty

Thank You for previewing this eBook

You can read the full version of this eBook in different formats:

- HTML (Free /Available to everyone)
- PDF / TXT (Available to V.I.P. members. Free Standard members can access up to 5 PDF/TXT eBooks per month each month)
- Epub & Mobipocket (Exclusive to V.I.P. members)

To download this full book, simply select the format you desire below