Data Model: World Cup 2018

Viji Kumar

Table of Contents

Chapter 1 - Explanatory Framework

The aim of this book is to explain, using a data model, how FIFA World Cup tournament data can be used to produce a set of searchable records to document the outcomes of the games played. This will be done by the linking of individual items of information (***data***) about an entity, e.g. the names of teams, with other items of information e.g. number of games played, goals scored etc. The data of interest are that which enable the deduction of the result of each game. Here the word ***database*** will refer to any collection of linked data in any structured format e.g. in rows and columns. To explain how databases may be constructed in accordance with the data model, I shall work through the creation and population of a database that can be used to document the FIFA World Cup 2018 tournament to be held in Russia. The intention is to use a method that is simple, rigorous and repeatable. The techniques employed and the models discussed are independent of any proprietary database management system. The models can also be extended.

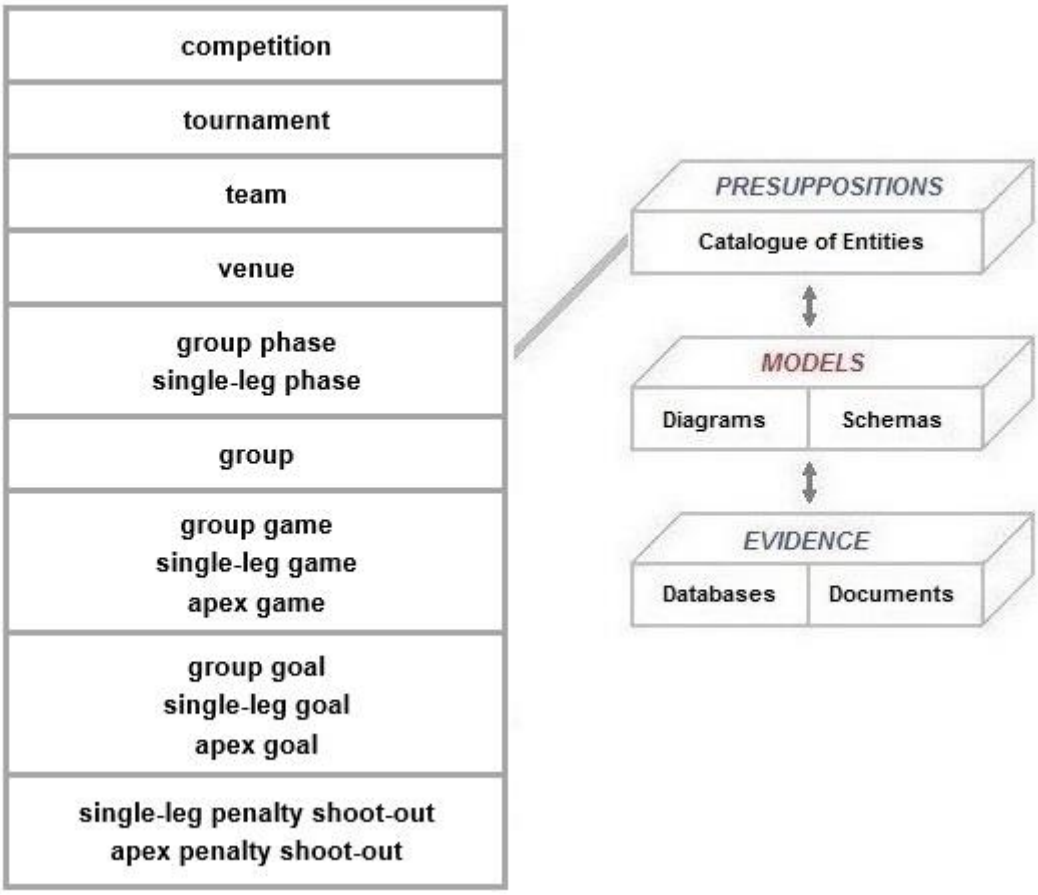
The initial analysis should produce data models and schemas or in the demotic, the ***metadata***. The outcome of that analysis is the specification of a catalogue of entities *(entities that will be linked)* about which data will have to be collected. These entity types are defined as distinct constructs with a separate existence, e.g. group games, final game, group goals etc. Entities can also refer to separate, definable and distinct constructs that have no material existence e.g. competitions, tournaments, groups etc. In this model all abstract entities have links, either directly or indirectly, to material entities e.g. all groups during a group phase of a tournament are associated with group games.

The first step, listing the different types of entities necessary to describe this model of a football tournament, is uncontroversial. The entity types required are competition, tournament, team, venue, group phase, single-leg *(knock-out)* phase, group, 3 types of games and 5 types of goals including penalty shoot-outs. Essentially you count the number of goals scored in each game by each team and deduce the result. In any type of game, the two participating teams are labelled ***team one*** and ***team two***. Round robin contests, e.g. groups, are decided by the accumulation of points e.g. 3 points for a victory and a point for a draw. To effectively report on a group, a table summarising the results and ranking the teams based on the points total is required. This explanation will follow all the stages from creating the data model representing the relationships between the different entity types listed above to providing sample scoring data to test the model.

Carolus *(Father of Taxonomy)* Linnaeus's words below are an excellent starting point.

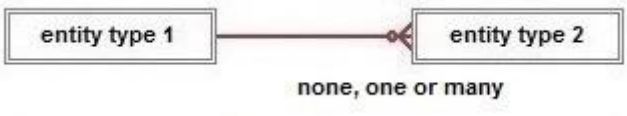
The first step in wisdom is to know the things themselves; this notion consists in having a true idea of the objects; objects are distinguished and known by classifying them methodically and giving them appropriate names. Therefore, classification and name-giving will be the foundation of our science.

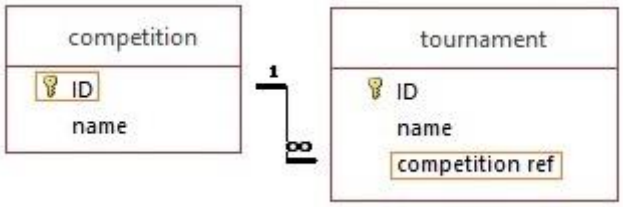
Systema Naturae (1735), trans. M. S. J. Engel-Ledeboer and H. Engel (1964)



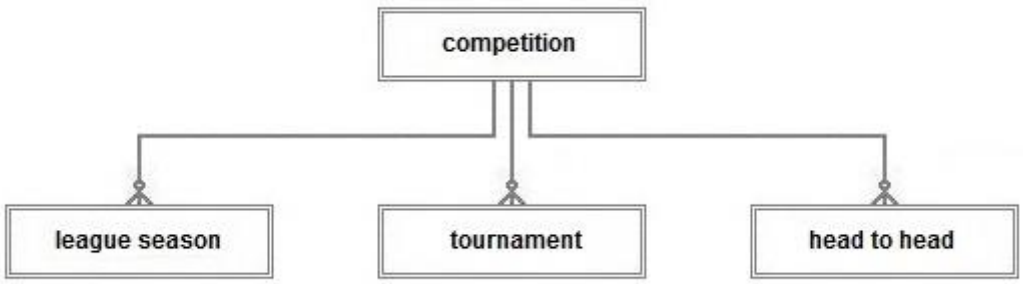
A tournament, as defined here, starts with four or more teams and then whittles the teams down with each successive phase, e.g. the 2014 FIFA World Cup in Brazil started with a group phase *(32 teams)*, followed by 3 single-leg knock-out phases *(16, 8 and 4 teams)* ending with the final and third place games. Tournament prizes are decided by the final and optionally, 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 of a tournament.)* 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.

The relationship or the link between any two entities is represented by a line starting at entity type 1 and ending with a nought and a crow's foot at entity type 2. 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 in that role. An alternate depiction of the link *(MS Access)* is also provided below showing additionally the mechanism that enforces the relationship.





The tournament model used here has 19 entity types, the 15 listed above plus 3 documenting the qualification paths to subsequent phases and 1 documenting the teams in a group. While this exercise covers a tournament, competitions may also be decided by league seasons and head to head encounters. Head to head competitions include Test series e.g. the Ashes and competitions where the two participants qualify by winning other competitions e.g. the English Community Shield and the Spanish Super Cup. This extensibility for the model is required because there are teams that play league competitions and knock out tournaments or head to head competitions during the same season. Some of the teams that play in the Football Association Challenge Cup (FA Cup) competition simultaneously play in the English Premier League competition. So, it may be necessary to collect data from different types of competitions to report comprehensively on a team's performance over any given period. Furthermore, a competition may have both, a league season followed by a tournament to decide the prizes for the season. The Football League Championship has had in recent years, a *(play-off)* tournament in addition to the league season to decide the 3 teams that are promoted to the Premier League. In England, the top tier of professional Rugby League also employs play-offs.



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. A portion of the data model showing some of the relationships of the apex game entity type is also specified as a 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 provides Ptolemy V's decree in 3 different scripts *(Ancient Egyptian hieroglyphs, Demotic and Ancient Greek).* That thoughtful act subsequently helped Thomas Young decipher Egyptian hieroglyphs *(thank you* **Wikipedia** *for this and many other snippets in this document)*.

There is a large amount of historical data available to test models purporting to explain how competitions work. The raw data necessary to test the model can be found in existing databases or in documents such as newspapers and books. Constructing and populating databases in accordance with

the model in this book should produce databases amenable to being queried to obtain the results of games, enable the deduction of the qualifiers from each phase and the eventual winning team. The detailed model will define each type of entity by listing attributes that are of interest e.g. the name of a competition or team, the date, time and venue of a game *et al*.

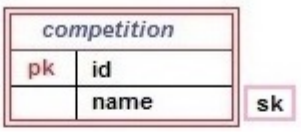
The definitions of the organisational units, e.g. league seasons, tournaments, group phases, groups, single-leg phases, will be discussed, highlighting their common ancestry. The derivation of the different variants of the other types of entities, e.g. the different types of games, goals and qualification paths will also be explained. For every goal scored *(excluding penalty shoot-out goals)*, 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 stoppage/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 discussed 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 tournament 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 construction of the models.

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 year of a team or the current coach; for the sake of simplicity I have not included such attributes. 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 e.g. what was the UEFA Cup is now the UEFA Europa League.



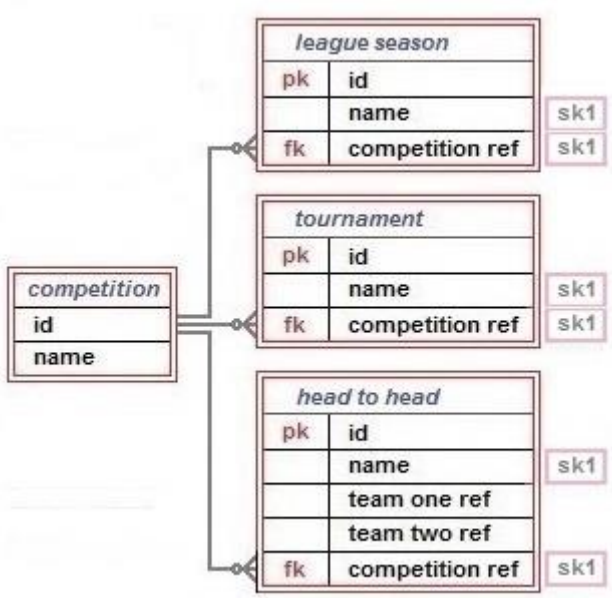
competition	
ID	name
1001	English Premier League
1002	FIFA World Cup
1003	FA Challenge Cup
1004	UEFA European Championship
1005	UEFA Champions League
1006	Football League Championship

A competition, team or venue is defined with both its attributes, its name and its id, as keys to denote that they must both 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 an entity. The name, which may change, must also be unique in this context because users of the databases may wish to search for results and other statistics about a 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. The attributes become table columns.

All league seasons, tournaments and head to heads must be linked to a competition. This is where the diagrams translate into linked tables, i.e. the model manifests itself in the real world. Therefore, in addition to an id and a name, these tables have a third attribute, the id of a competition, linking them to a competition. These 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 but the identities of the teams may not be known when a 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.

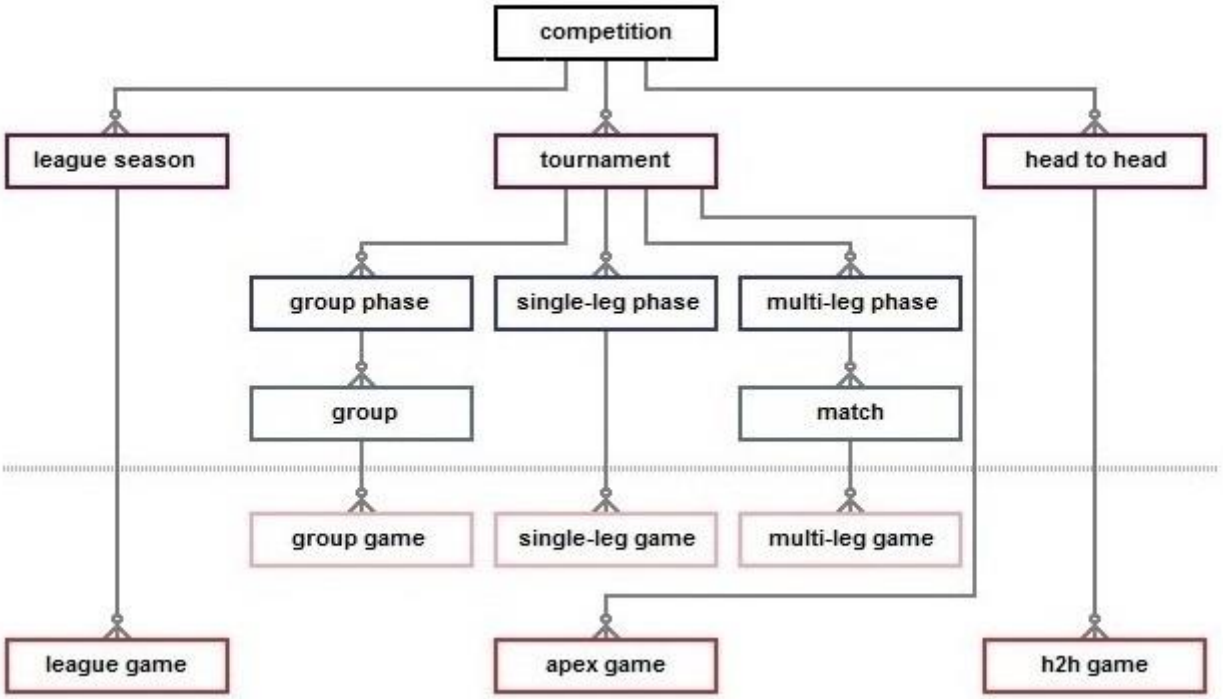


The secondary key is now a composite key that comprises both, the competition foreign key in addition to the name, to ensure that the names of the league seasons, tournaments and head to heads are unique for each competition.

tournament		
ID	name	competition ref
1001	2018	1002
1002	2017/18	1006
1003	2014	1002
1004	2017/18	1005

tournament		
ID	name	competition ref
1001	2018	FIFA World Cup
1002	2017/18	Football League Championship
1003	2014	FIFA World Cup
1004	2017/18	UEFA Champions League

The model shown below includes the organisational entities used to represent the 3 types of competitions. Links to teams and venues for the different types of games are not shown.



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 group, often has 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 deduced. 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 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.

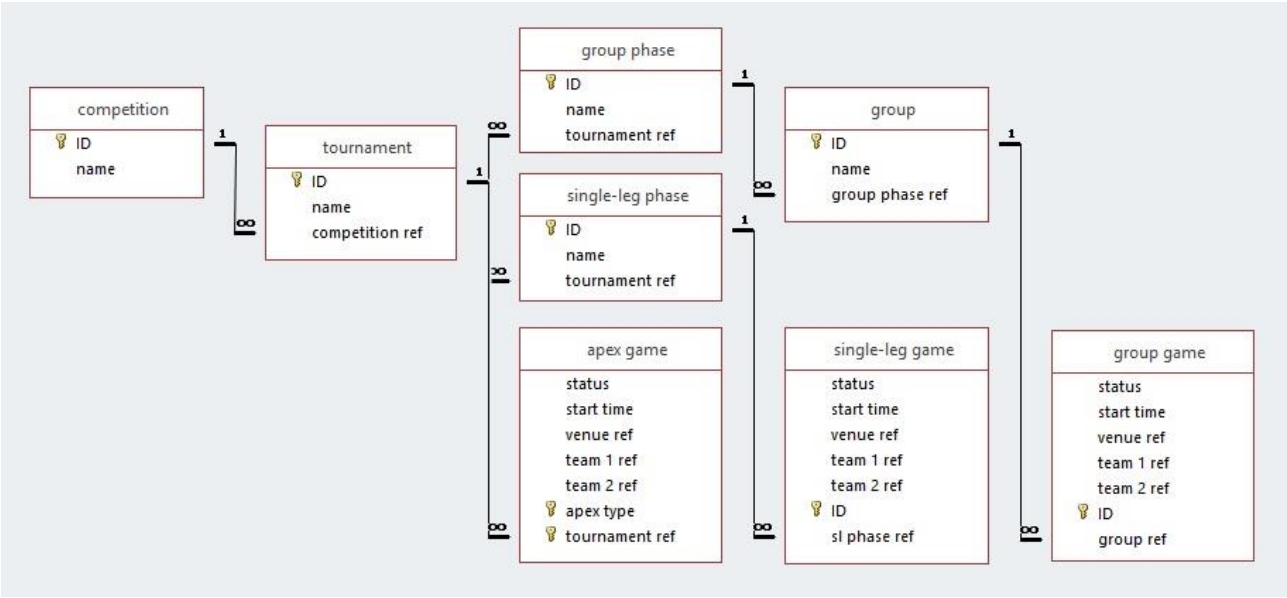
The model is used below to describe the last two tournaments of the UEFA European Championships. There were 53 and 54 teams eligible to play in Euro 2012 and Euro 2016 respectively, including the non-qualifying hosts.

The teams in both qualifying group stages were divided into 9 groups; some with 6 participants and some with 5. To decide the best runner-up, group performances are viewed for the top five teams in each group. In both cases there were 8 teams from the qualifying groups that took part in a play-off with the four winners also qualifying. The four play-off matches were home and away fixtures.

Returning to the detail of the 2018 FIFA tournament model; there is an initial group phase of 8 groups. The 4 teams in each group play each other once, a total of 48 *(8 x 6)* group games. The first two teams

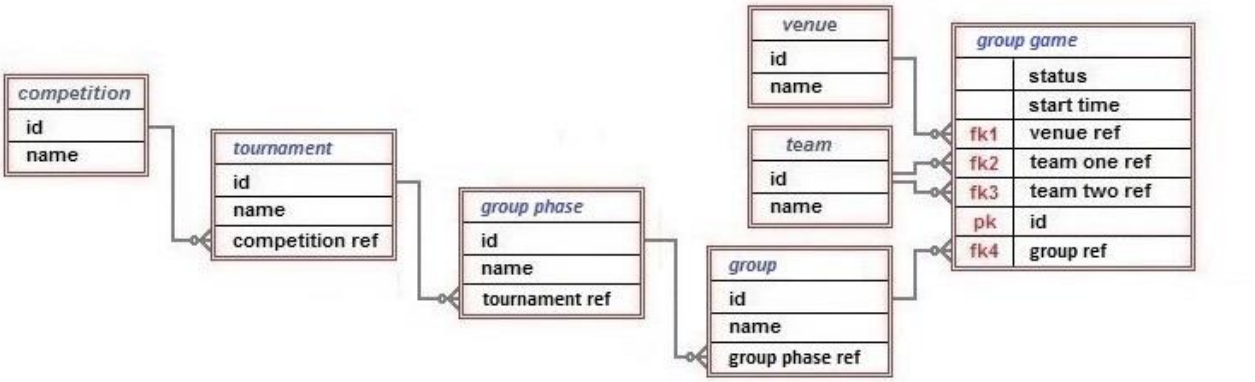
in each group qualify for the next phase. The next three rounds are single-leg knockout rounds producing the 4 teams who will contest the two apex games; the final and the third-place game.

The diagram below documents, using MS Access, the attributes of interest.



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; no data (*to be played or the score is not known*), incomplete data (*game in progress with all known scores*) and complete data (*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 *(local time at the venue)* and the venue reference are also common attributes for all types of games in this model. The two participating team references *(the team one and team two roles)* are attributes of group games but are not common to all types of games in the extended model e.g. for head to head games during a series. The group reference is the only attribute specific to a group game. Again, a unique identifier, id, is the prime key. The two team references, the venue reference and the group reference are all foreign keys.

The teams that are in a group can be deduced from reviewing the group game participants, but that relationship is also explicitly recorded in another table *(not shown above)* and is used when generating the group points table. The table below shows the details for the group A games.

group game						
status	start time	venue ref	team 1 ref	team 2 ref	ID	group ref
no data	14/06/2018 18:00:00	1001	1001	1032	1001	1001
no data	15/06/2018 17:00:00	1012	1022	1015	1002	1001
no data	19/06/2018 21:00:00	1003	1001	1022	1003	1001
no data	20/06/2018 18:00:00	1010	1015	1032	1004	1001
no data	25/06/2018 18:00:00	1007	1015	1001	1005	1001
no data	25/06/2018 17:00:00	1008	1032	1022	1006	1001

The venue, team and group names will be required to translate the foreign keys in the table of group games to make it meaningful for humans. The names can be found in the tables shown below. The four group A teams, Russia, Uruguay, Egypt and Saudi Arabia, are shown in red.

team	
ID	name
1001	Russia
1002	Germany
1003	Brazil
1015	Uruguay
1016	Croatia
1017	Denmark
1020	Sweden
1021	Tunisia
1022	Egypt
1030	Panama
1031	South Korea
1032	Saudi Arabia

venue	
ID	name
1001	Luzhniki Stadium (Moscow)
1002	Otkrytiye Arena (Moscow)
1003	Krestovsky Stadium (St Petersburg)
1004	Kaliningrad Stadium
1005	Kazan Arena
1006	Nizhny Novgorod Stadium
1007	Cosmos Arena (Samara)
1008	Volgograd Arena
1009	Mordovia Arena (Saransk)
1010	Rostov Arena (Rostov-on-Don)
1011	Fisht Olympic Stadium (Sochi)
1012	Central Stadium (Yekaterinburg)

group		
ID	name	group phase ref
1001	A	1001
1002	B	1001
1003	C	1001
1004	D	1001
1005	E	1001
1006	F	1001
1007	G	1001
1008	H	1001

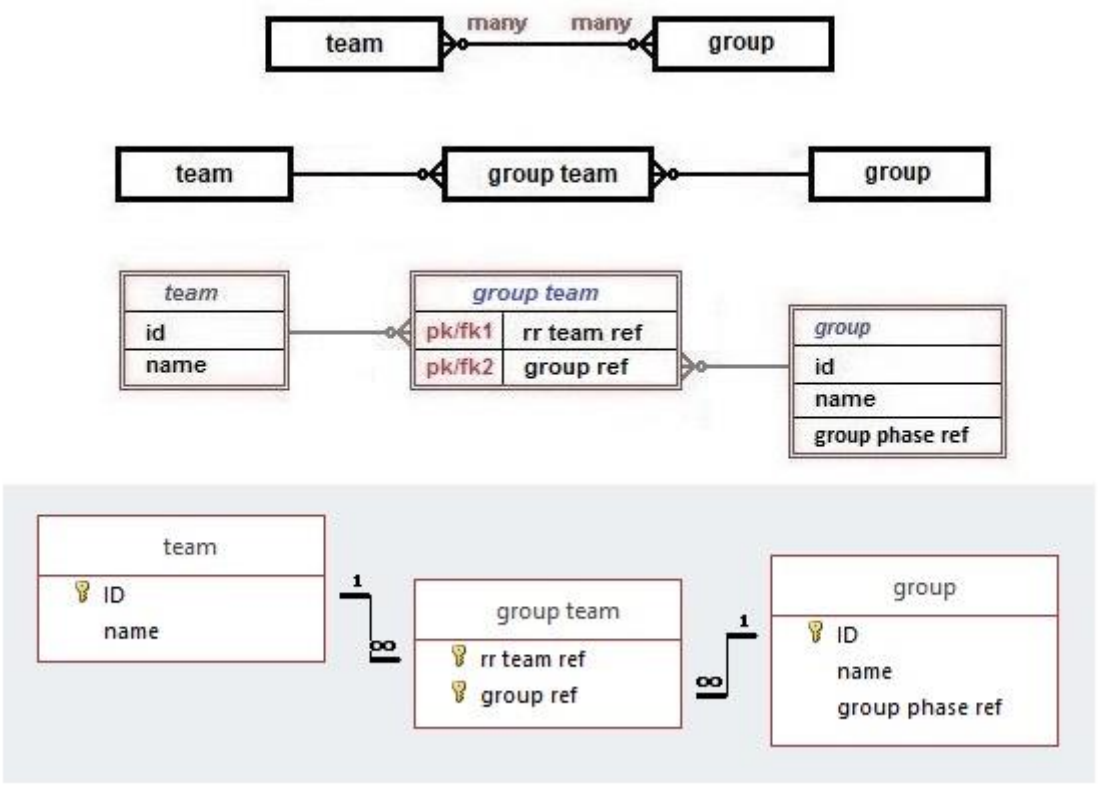
group phase		
ID	name	tournament ref
1001	Round 1	1001

The group A games table below has the names substituted for the ids of the teams, venues and group.

group game						
status	start time	venue ref	team 1 ref	team 2 ref	ID	group ref
no data	14/06/2018 18:00:00	Luzhniki Stadium (Moscow)	Russia	Saudi Arabia	1001	A
no data	15/06/2018 17:00:00	Central Stadium (Yekaterinburg)	Egypt	Uruguay	1002	A
no data	19/06/2018 21:00:00	Krestovsky Stadium (St Petersburg)	Russia	Egypt	1003	A
no data	20/06/2018 18:00:00	Rostov Arena (Rostov-on-Don)	Uruguay	Saudi Arabia	1004	A
no data	25/06/2018 18:00:00	Cosmos Arena (Samara)	Uruguay	Russia	1005	A
no data	25/06/2018 17:00:00	Volgograd Arena	Saudi Arabia	Egypt	1006	A

Chapter 3 – The Round Robin Model

This chapter aims to explain how to construct a round robin competition model using the group A games of the FIFA World Cup 2018 for illustration. The participation of teams in groups is an example of a many-to-many linkage. A group may have none, one or many participating teams and a team may have participated in, be participating in or be eligible to participate in none, one or many groups.



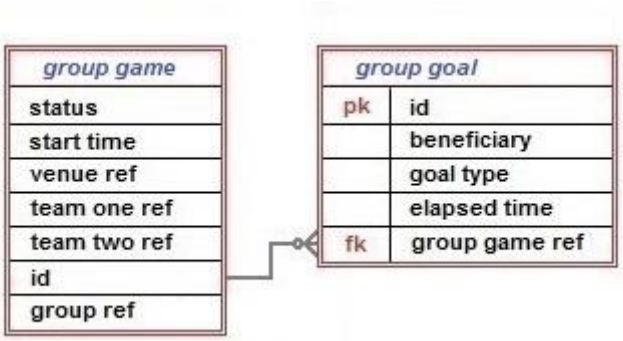
The decomposition of that single many-to-many link into 2 one-to-many links is achieved by the creation of an intermediate entity, the ***group team***. It has just two attributes and they are also foreign keys. The primary key of the group team type of entity is composed of the two foreign keys referencing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

group team	
π team ref	group ref
1001	1001
1015	1001
1022	1001
1032	1001

group team	
π team ref	group ref
Russia	A
Uruguay	A
Egypt	A
Saudi Arabia	A

For group games, the score will only involve goals during 90 minutes of play, without the option of extra

time or a penalty shoot-out. A reference to the group game is the only attribute specific to a group goal entity type. 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*). Group goals scored during stoppage time are recorded as being scored in the 45th or 90th minute.



The 11 ***sample*** group A goals shown below are examples of goals that may be scored during the group phase of Russia 2018. 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 be deduced using the team sheets, the scorer and the type of goal. The beneficiary attribute is then unnecessary and should be removed *(Occam's razor)*. That's all the data required in this model to produce the group game results and group points tables.

group goal				
ID	beneficiary	goal type	elapsed time	group game ref
1001	team two	open play	14	1001
1002	team one	penalty	73	1001
1003	team two	open play	23	1002
1004	team one	own goal	56	1002
1005	team two	penalty	68	1002
1006	team two	open play	87	1003
1007	team one	penalty	3	1004
1008	team one	open play	42	1004
1009	team one	open play	27	1005
1010	team one	own goal	56	1005
1011	team one	penalty	89	1005

An example of a data entry form to record the above data is shown below. If required, the id is can be generated automatically and does not necessarily have to be an integer. The role of the beneficiary can be ascertained from the drop-down list of the game details.

group goal

ID: 1001

beneficiary: team two

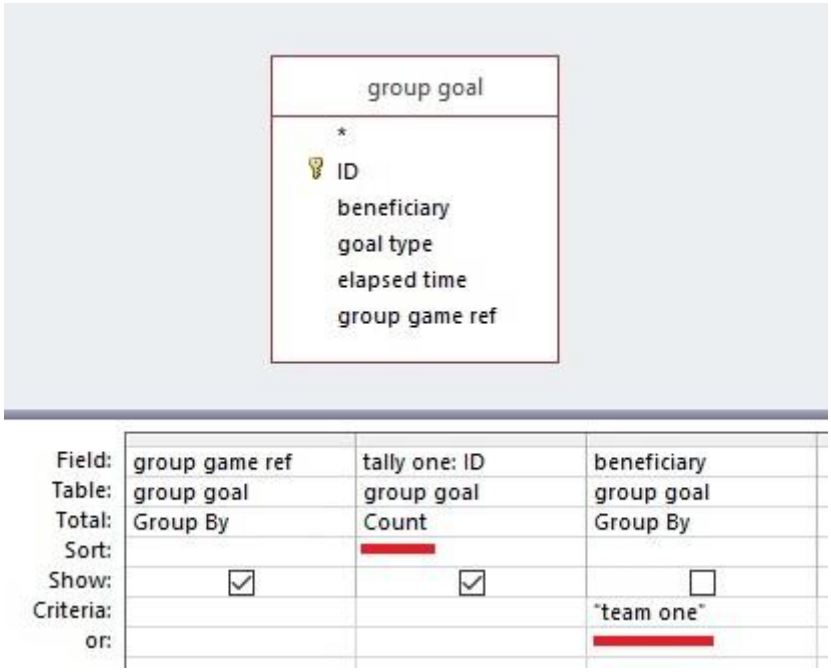
goal type: open play

elapsed time: 14

group game ref: 1001

ID	status	start time	venue	team one	team two	group
1001	complete data	14/06/2018 18:00:00	Luzhniki Stadium (Moscow)	Russia	Saudi Arabia	A
1002	complete data	15/06/2018 17:00:00	Central Stadium (Yekaterinburg)	Egypt	Uruguay	A
1003	complete data	19/06/2018 21:00:00	Krestovsky Stadium (St Petersburg)	Russia	Egypt	A
1004	complete data	20/06/2018 18:00:00	Rostov Arena (Rostov-on-Don)	Uruguay	Saudi Arabia	A
1005	complete data	25/06/2018 18:00:00	Cosmos Arena (Samara)	Uruguay	Russia	A
1006	complete data	25/06/2018 17:00:00	Volgograd Arena	Saudi Arabia	Egypt	A

The goal data is used to obtain the tallies of goals scored by each team during the game. There will be no tallies for teams who do not score. The tallies for team one and team two *(group tally one and group tally two respectively)* are generated separately.



group tally one	
group game ref	tally one
1001	1
1002	1
1004	2
1005	3

group tally two	
group game ref	tally two
1001	1
1002	2
1003	1

To produce group game results, I have used another query to bring together the two tallies and to record a zero for the teams that did not score during a game. In MS Access this is done by using the null zero function; Nz(), which returns a zero if there are no goals scored by a team. The results are only generated when the game status has been set to “complete data”. I have also generated a ***outcome*** attribute for each game to denote if it is a group game without a known outcome, (0), a win for team one, (1), a win for team two, (2) or a draw, (3).

group outcome			
group game ref	tally one	tally two	outcome
1001	1	1	3
1002	1	2	2
1003	0	1	2
1004	2	0	1
1005	3	0	1
1006	0	0	3

An IF, THEN, ELSE programming construct is also required to deduce the scores and the game outcomes for completed games.