Advanced Data Warehousing

Level of Aggregation (Adding New Dimensions)

To lower down the level of aggregation of a star schema can be done by adding new dimensions to the star schema. Naturally, when a new dimension is added, in which a new attribute is added to the fact table, the level of details of the fact measure will become more details; hence lower down the level of aggregation.

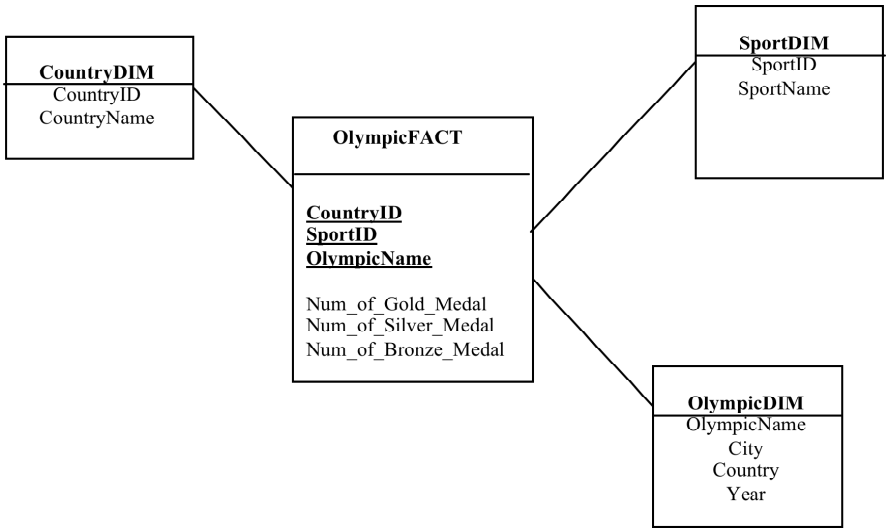
However, one must be careful when adding a star schema with a new dimension. Particularly, there are two things: First is that there are cases where adding a new dimension does not lower down the level of aggregation. Second is that when adding a new dimension, it will create double-dip values to the fact measure, and hence the fact measure will be incorrect.

Let's discuss these two things in more detail.

1. Adding New Dimensions does not Lower Down the Level of Aggregation

Consider the following Olympic star schema, consisting of three dimensions: CountryDIM, SportDIM, and OlympicDIM. The CountryDIM dimension maintains a list of countries which participated in the Olympic Games. The OlympicDIM dimension lists all the Olympic names, including the city and the year (e.g. London Olympic was in London in 2012, and Rio Olympic was in Rio, Brazil, in 2016).

Version-1 star schema with three dimensions



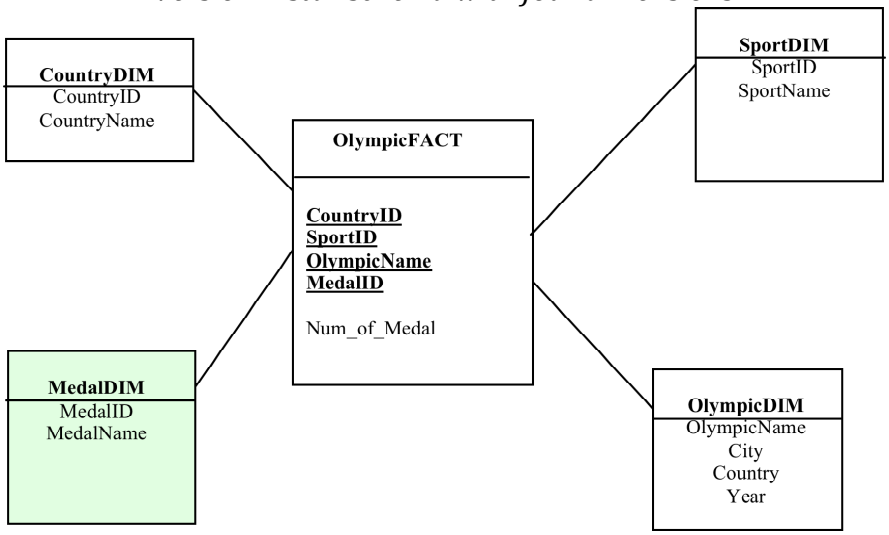
The SportDIM dimension lists all sport types in the Olympic, for example: Swimming, Athletics, Volleyball, etc. Each sport is actually a category of the sport, not the actual sport event. So for example, the sport of Swimming is a type of

sport, instead of the actual swimming event, such as "100m butterfly men", or "4x100m freestyle relay women". Or in the Volley Ball, it is not the actual volleyball event, such as "Volleyball Men" or "Volleyball Women". Hence, the SportDIM dimension has a high level of aggregation.

Now we are going to add a new dimension called MedalDIM, which has only three records: Gold, Silver, and Bronze. But, the fact measure is reduced from three to one: Number of Medals, instead of Number of Gold Medals, Number of Silver Medals, and Number of Bronze Medals.

The new star schema with four dimensions is as follows:

Version-2 star schema with four dimensions



The main question is whether this new star schema with four dimensions has a lower level of aggregation compared to the previous star schema with only three dimensions. To answer this question, we need to examine the data (or the records) in the fact table.

The fact table for version-1 star schema (without Medal IM) has 6 attributes: three from the dimensions, and the other three for the fact measures. The contents of the fact table are as follows:

Fact (version-1 star schema)

Country	Sport	Olympic Name	Num of Gold	Num of Silver	Num of Bronze
USA	Swimming	London 2012	16	9	6
China	Swimming	London 2012	5	1	4
Australia	Swimming	London 2012	1	6	3

The fact table for version-2 star schema (with MedalDIM) consists of 5 columns: four from the dimension, but only one fact measure.

Fact (version-2 star schema)

Country	Sport	Olympic Name	Medal Type	Num of Medals
USA	Swimming	London 2012	Gold	16
USA	Swimming	London 2012	Silver	9
USA	Swimming	London 2012	Bronze	6
China	Swimming	London 2012	Gold	5
China	Swimming	London 2012	Silver	1
China	Swimming	London 2012	Bronze	4
Australia	Swimming	London 2012	Gold	1
Australia	Swimming	London 2012	Silver	6
Australia	Swimming	London 2012	Bronze	3

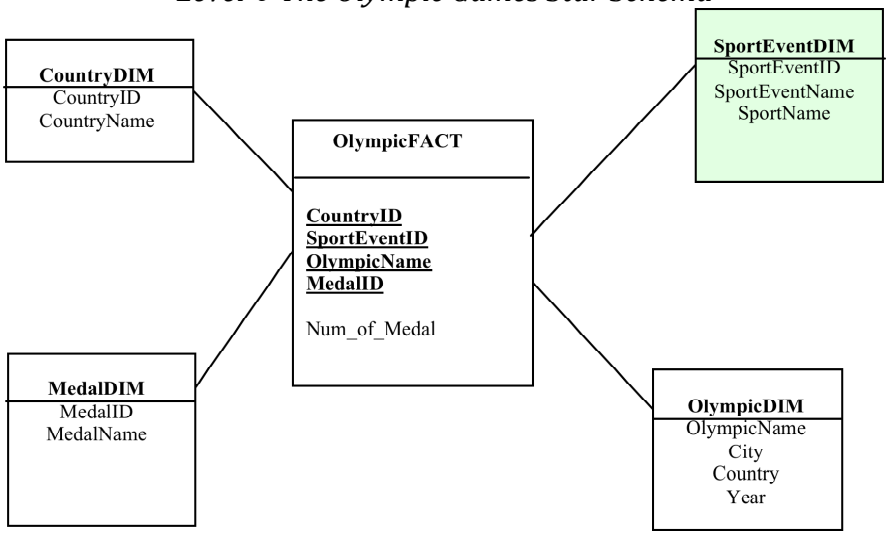
Comparing the two fact tables, it is clear that fact table from version-2 star schema is not more detail than that from version-1, or vice-versa. Hence, version-2 star schema is not of a lower level of aggregation than version-1 star schema. Both have the same level of details, and hence the same level of aggregation.

Why is it? It is because the fact measures in both star schemas are different. Version-1 star schema has three fact measures, whereas version-2 star schema has only one fact measure. So, in this case, adding a new dimension to version-1 star schema does not lower down the level of aggregation.

2. Adding New Dimensions may result in a double dipping in the fact measure

Before discussing the double dipping problem, let's lower down the level of aggregation of the version-2 star schema in the previous section, by changing the level of granularity of the SportDIM dimension. We now replace the SportDIM dimension (which is basically the sport category) with the SportEventDIM dimension.

Level-0 The Olympic Games Star Schema



So in the SportEventDIM, we keep the actual sport event, instead of the sport category. For example, instead of "Swimming", we now have "100m butterfly men", "4x100m freestyle relay women", etc. Naturally, this star schema has a lower level of aggregation compared to the previous star schema.

Now let's examine the records in the fact table. In comparing and contrasting the fact tables of the two stars schema, let's have a look at the content of both fact tables (level-1 and level-0 star schemas). For simplicity, we focus on the Australian records only.

Fact (level-1 star schema)

Country	Sport	Olympic Name	Medal Type	Num of Medals
Australia	Swimming	London 2012	Gold	1
Australia	Swimming	London 2012	Silver	6
Australia	Swimming	London 2012	Bronze	3

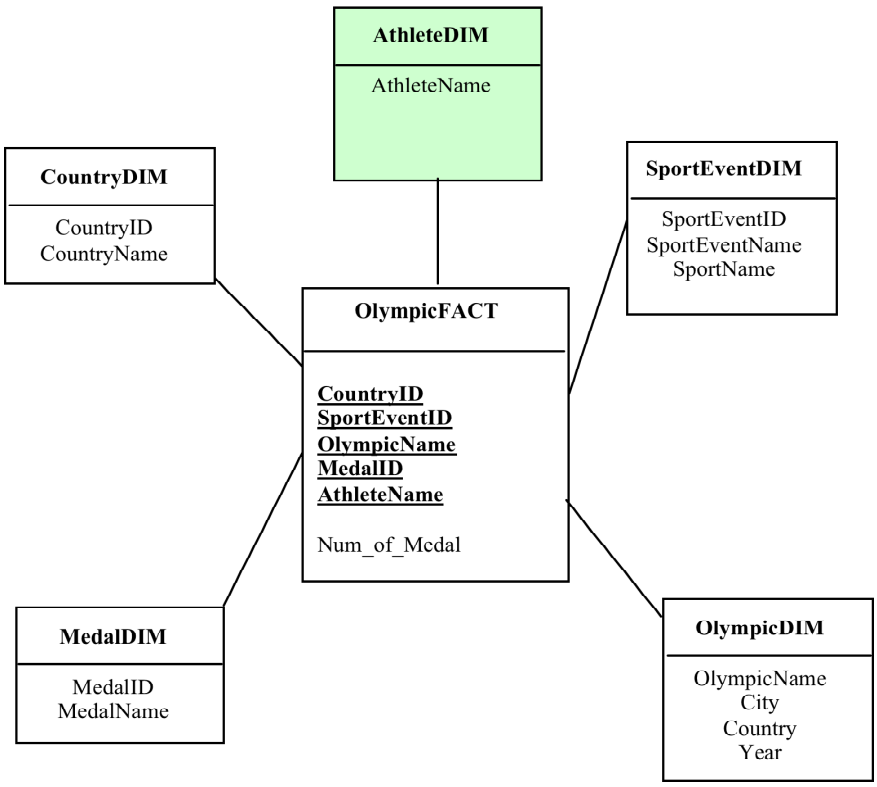
Fact (level-0 star schema)

Country	SportEvent	Olympic Name	Medal Type	Num of Medals
Australia	4x100m Freestyle Relay Women	London 2012	Gold	1
Australia	4x100m Medley Relay Women	London 2012	Silver	1
Australia	4x200m Freestyle Relay Women	London 2012	Silver	1
Australia	100m Breaststroke Men	London 2012	Silver	1
Australia	100m Freestyle Men	London 2012	Silver	1
Australia	200m Individual Medley Women	London 2012	Silver	1
Australia	100m Backstroke Women	London 2012	Silver	1
Australia	4x100m Medley Relay Men	London 2012	Bronze	1
Australia	100m Butterfly Women	London 2012	Bronze	1
Australia	200m Freestyle Women	London 2012	Bronze	1

Note that by changing the level of granularity from Sport to Sport Event, the level of granularity of the fact changes. So, instead of having just one record for the swimming with the bronze medal, now we have the break down, which is the three bronze medal records. The fact table with Sport Event naturally has a higher level of granularity (or a lower level of aggregation) compared to the fact table with Sport. Because the fact table with Sport Event has 1s in the number of medals (i.e. the fact measure), this star schema is level-0 star schema; no aggregation.

What will happen if we decide to add a new dimension called "AthleteDIM". The rationale behind this is very simple, that is to drill down to each winning athlete. The new star schema is then as follows:

A new star schema with an Athlete Dimension



Incorrect Fact Table

Country	SportEvent	Athlete	Olympic Name	Medal Type	Num of Medals
Australia	4x100m Freestyle Relay Women	Alicia Coutts	London 2012	Gold	1
Australia	4x100m Freestyle Relay Women	Cate Campbell	London 2012	Gold	1
Australia	4x100m Freestyle Relay Women	Brittany Elmslie	London 2012	Gold	1
Australia	4x100m Freestyle Relay Women	Melanie Schlanger	London 2012	Gold	1
Australia	4x100m Medley Relay Women	Emily Seebohm	London 2012	Silver	1
Australia	4x100m Medley Relay Women	Leisel Jones	London 2012	Silver	1
Australia	4x100m Medley Relay Women	Alicia Coutts	London 2012	Silver	1
Australia	4x100m Medley Relay Women	Melanie Schlanger	London 2012	Silver	1
Australia	4x200m Freestyle Relay Women	Bronte Barratt	London 2012	Silver	1
Australia	4x200m Freestyle Relay Women	Melanie Schlanger	London 2012	Silver	1
Australia	4x200m Freestyle Relay Women	Kylie Palmer	London 2012	Silver	1
Australia	4x200m Freestyle Relay Women	Alicia Coutts	London 2012	Silver	1
Australia	100m Breaststroke Men	Christian Sprenger	London 2012	Silver	1
Australia	100m Freestyle Men	James Magnussen	London 2012	Silver	1
Australia	200m Individual Medley Women	Alicia Coutts	London 2012	Silver	1
Australia	100m Backstroke Women	Emily Seebohm	London 2012	Silver	1
Australia	4x100m Freestyle Relay Men	Hayden Stoeckel	London 2012	Bronze	1
Australia	4x100m Freestyle Relay Men	Christian Sprenger	London 2012	Bronze	1
Australia	4x100m Freestyle Relay Men	Matt Targett	London 2012	Bronze	1
Australia	4x100m Freestyle Relay Men	James Magnussen	London 2012	Bronze	1
Australia	100m Butterfly Women	Alicia Coutts	London 2012	Bronze	1
Australia	200m Freestyle Women	Bronte Barratt	London 2012	Bronze	1

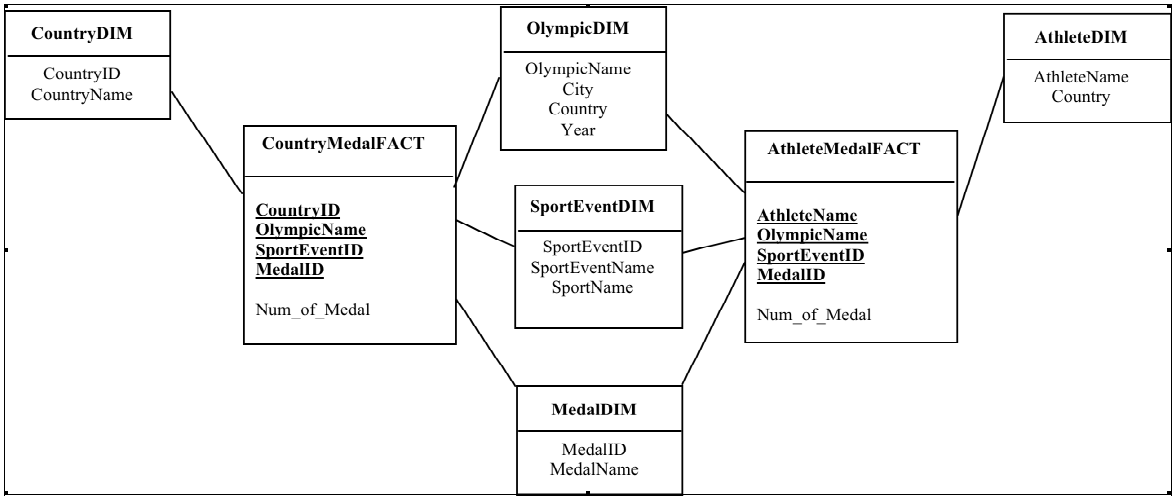
By looking at the fact table, the last record in the above example shows that Bronte Barrat was the athlete who got the broanze medal in 200m freestyle women. So, the drilling down to the winning athlete seems to be correct. From the fact table, it is also correct to see that Bronte Barrat received 2 (two) medals, one bronze (200m freestyle women), and one silver (4x200m freestyle relay women).

However, the above fact table is incorrect because if we query number of gold medals that Australia received in Swimming in London Olympic 2012, it will return 4, instead of 1. Therefore, adding a new dimension AthleteDIM will result in an incorrect fact measure (e.g. number of medals). The reason is because it is double dipping the count of medals.

Double dipping of number of medal count is due to different focus or different subject of the star schema. One focus is from the Country point of view, and another focus is from the Athlete point of view. The two cannot be mixed and combined into one star schema.

The solution is to have two star schemas: one star schema focuses on country, and the other focuses on athlete. The multi-fact star schema is shown as follows:

A multi-fact star schema for the Olympic Games (still incorrect!!!)



The fact tables look like the following:

CountryMedalFact

Country	SportEvent	Olympic Name	Medal Type	Num of Medals
Australia	4x100m Freestyle Relay Women	London 2012	Gold	1
Australia	4x100m Medley Relay Women	London 2012	Silver	1
Australia	4x200m Freestyle Relay Women	London 2012	Silver	1
Australia	100m Breaststroke Men	London 2012	Silver	1
Australia	100m Freestyle Men	London 2012	Silver	1
Australia	200m Individual Medley Women	London 2012	Silver	1
Australia	100m Backstroke Women	London 2012	Silver	1

Australia	4x100m Medley Relay Men	London 2012	Bronze	1
Australia	100m Butterfly Women	London 2012	Bronze	1
Australia	200m Freestyle Women	London 2012	Bronze	1

AthleteMedalFact

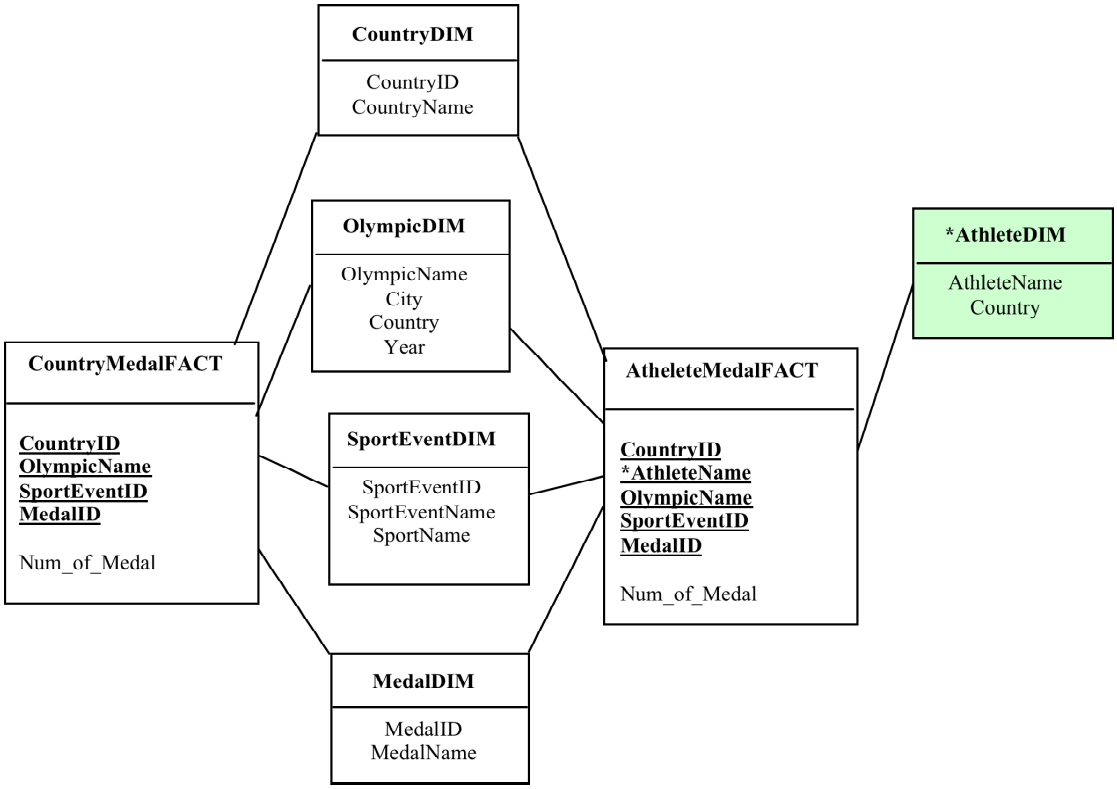
Athlete	SportEvent	Olympic Name	Medal Type	Num of Medals
Alicia Coutts	4x100m Freestyle Relay Women	London 2012	Gold	1
Cate Campbell	4x100m Freestyle Relay Women	London 2012	Gold	1
Brittany Elmslie	4x100m Freestyle Relay Women	London 2012	Gold	1
Melanie Schlanger	4x100m Freestyle Relay Women	London 2012	Gold	1
Seebohm	4x100m Medley Relay Women	London 2012	Silver	1
Leisel Jones	4x100m Medley Relay Women	London 2012	Silver	1
Alicia Coutts	4x100m Medley Relay Women	London 2012	Silver	1
Melanie Schlanger	4x100m Medley Relay Women	London 2012	Silver	1
Bronte Barratt	4x200m Freestyle Relay Women	London 2012	Silver	1
Melanie Schlanger	4x200m Freestyle Relay Women	London 2012	Silver	1
Kylie Palmer	4x200m Freestyle Relay Women	London 2012	Silver	1
Alicia Coutts	4x200m Freestyle Relay Women	London 2012	Silver	1
Christian Sprenger	100m Breaststroke Men	London 2012	Silver	1
James Magnussen	100m Freestyle Men	London 2012	Silver	1
Alicia Coutts	200m Individual Medley Women	London 2012	Silver	1
Emily Seebohm	100m Backstroke Women	London 2012	Silver	1
Hayden Stoeckel	4x100m Freestyle Relay Men	London 2012	Bronze	1
Christian Sprenger	4x100m Freestyle Relay Men	London 2012	Bronze	1
Matt Targett	4x100m Freestyle Relay Men	London 2012	Bronze	1
James Magnussen	4x100m Freestyle Relay Men	London 2012	Bronze	1
Alicia Coutts	100m Butterfly Women	London 2012	Bronze	1
Bronte Barratt	200m Freestyle Women	London 2012	Bronze	1

If we examine carefully, table AthleteMedalFact is also incorrect, because when we ask a question "how many gold medals for 4x100, freestyle relay women in London 2012 Olympic Games", the answer would be 4, which is incorrect.

The query on the AthleteMedalFact will be correct if the AthleteDIM dimension is ALWAYS used in any data retrieval on the AthleteMedalFact. Hence, AthleteDIM MUST BE a determinant dimension.

The correct multi-fact star schema is then shown as follows: (Note that we can still keep the CountryDIM in the CountryMedalFACT schema, because AthleteDIM is a determinant dimension).

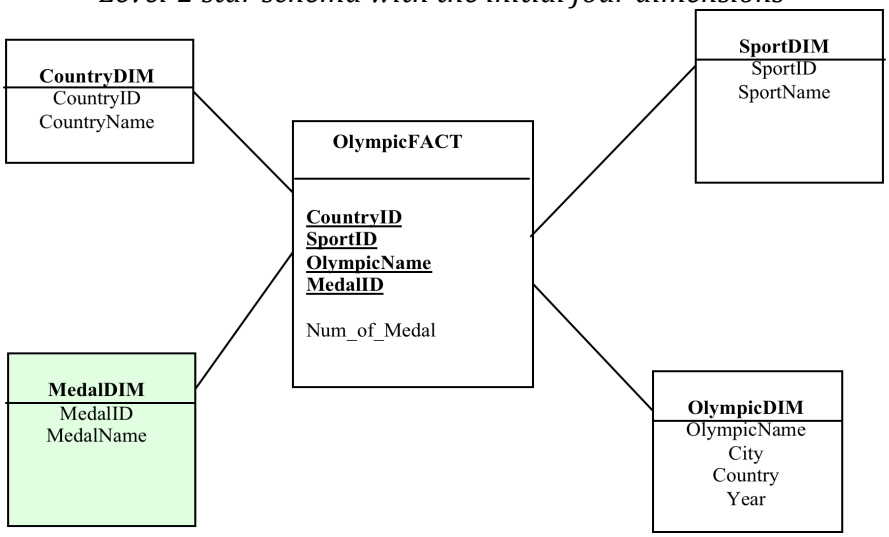
AthleteDIM is a Determinant Dimension (the correct multi-fact star schema)



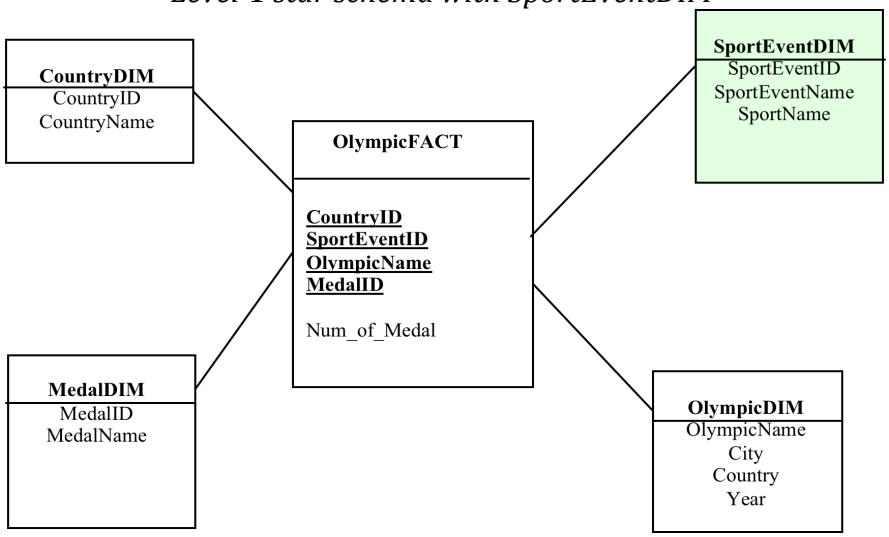
3. The Final Star Schemas

We can avoid the multi-fact schema, by having different level of granularity for the fact table with AthleteDIM. The levels of aggregation for the Olympic Games case study are then as follows:

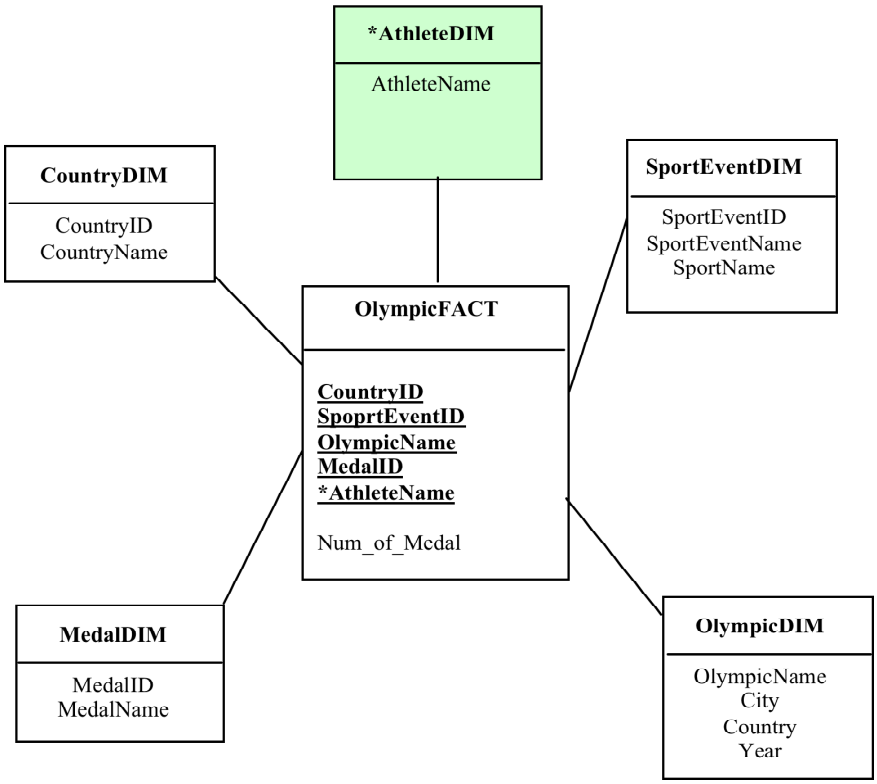
Level-2 star schema with the initial four dimensions



Level-1 star schema with SportEventDIM



Level-0 star schema with AthleteDIM as a Determinant Dimension



4. Summary

In general, adding a new dimension will result in a lower level of aggregation. However, one must take into account the following issues:

1. Adding a new dimension *will not lower down the level of aggregation*, especially when the fact measure has changed.
   For example, the fact measure changes from three fact measures (number of gold medals, number of silver medals, and number of bronze medals) to one fact measure (number of medals). By adding a new MedalDIM to the star schema does not change the level of granularity of the star schema, when the fact measures have also changed.

2. Adding a new dimension *will create an incorrect fact table*, if the new dimension is double dipping the fact measure count.
   For example, adding a new AthleteDIM will result in double dipping in the calculation of the number of medals.

3. Adding a new dimension *will create a multi-fact table*.
   For example, one fact table focuses on Country, whereas the other fact table focuses on Athlete.

4. The new fact table(s) might have a *determinant dimension*.
   For example, the Athlete dimension in the second fact (e.g. the AthleteMedalFact) is a determinant dimension.

5. The star schema that has a *determinant dimension* will become the lowest level of aggregation in a data warehouse architecture
   For example, the star schema with AthleteMedalFACT will become level-0 star schema, but the AthleteDIM must be a determinant dimension.

THE END