Determinant vs. Non-Determinant Dimensions

1. The Olympic Games Case Study

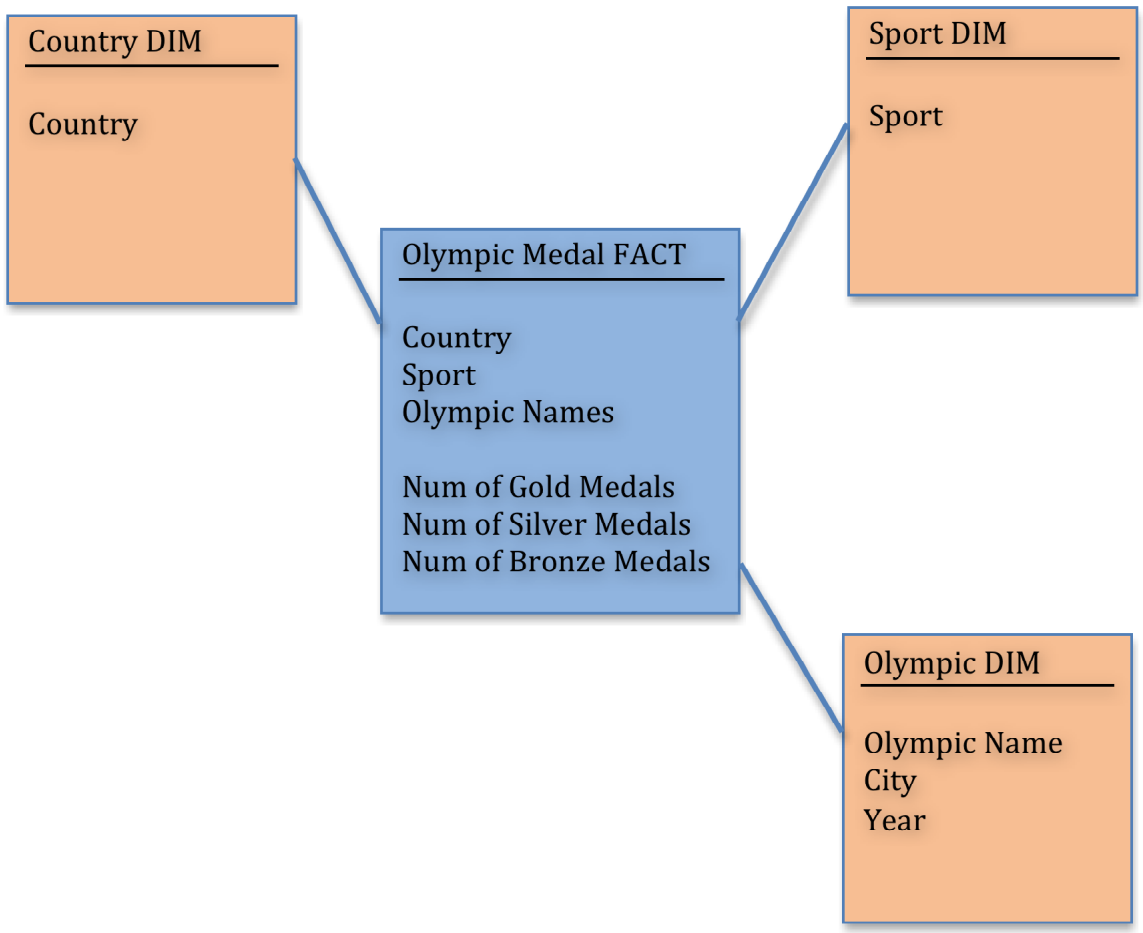
The Olympic Games committee maintains an operational database that stores all matches, games, as well as the medal winners of the Olympic Games over the years (https://www.olympic.org/).

We would like to build a data warehouse to analyze the medal counts, by each country, sport, and at which Olympic Games.

There are two possible star schemas. Version-1 star schema contains 3 dimensions, whereas Version-2 star schema contains 4 dimensions. For simplicity, a very minimal number of attributes are included in each dimension.

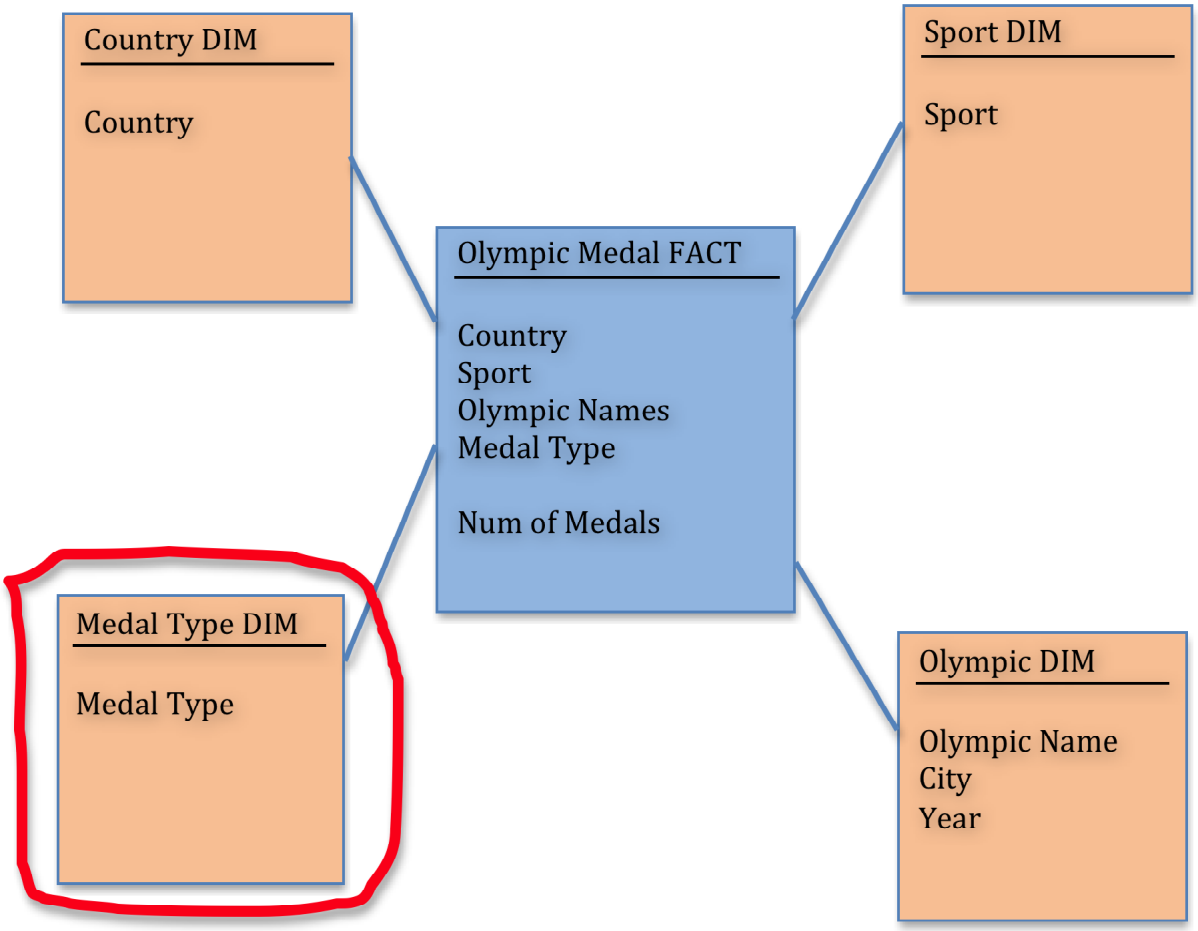
Version-1 star schema has CountryDIM, SportDIM, and OlympicGamesDIM as the dimensions. Three fact measures are included in the fact table, namely number of gold medals, silver medals, and bronze medals.

Version-1 Star Schema



Version-2 star schema has four dimensions – with an additional of Medal Type DIM (which is either Gold, Silver, or Bronze), but there is only one fact measure in the Fact table, which is Num of Medals.

Version-2 Star Schema with the Medal Type DIM



What is the difference between these two versions of star schema?
Is Medal Type DIM a Determinant Dimension?
In order to answer these questions, we need to visualize two-column tables for each dimension category.

2. Two-Column Tables for Version-1 Star Schema (without the Medal Type DIM)

The first two-column table is from the Country point of view, which is as follows:

Country	Num of Gold	Num of Silver	Num of Bronze
USA	733	602	488
China	199	143	133
Australia	167	170	189

These are the number of Gold, Silver and Bronze medals that these country got on all Olympic Games (several Olympic Games) recorded in the operational database. Note that this two-column table methodology is to help the data warehouse designer to visualize the view of the fact measures from each dimension.

The second two-column table is from the Sport point of view. Assuming that in the operational database, it records 20 past Olympic Games, and at each Olympic, there is only one gold for 100m Butterfly Men, for instance.

Sport	**Num of Gold**	**Num of Silver**	**Num of Bronze**
Swimming 100m Butterfly Men	20	20	20
Swimming 400m Freestyle Women	20	20	20
Swimming 4x100m Medley Relay Men	20	20	20

The third two-column table is from the Olympic Name point of view.

Olympic Name	**Num of Gold**	**Num of Silver**	**Num of Bronze**
London 2012	302	304	356
Beijing 2008	302	303	353
Athens 2004	301	301	327

All the three two-column tables above seem to be reasonably correct. The first columns are the categories, while the other columns are the fact measures which are numeric and aggregate values. Because these three two-column tables make sense, we are confident that version-1 star schema is correct.

3. Two-Column Tables for Version-2 Star Schema (with the Medal Type DIM)

The two-column tables for the first three dimensions, namely Country, Sport, and Olympic Names are as follows:

Country	**Num of Medals**
USA	1823
China	475
Australia	526

Sport	**Num of Medals**
Swimming 100m Butterfly Men	60
Swimming 400m Freestyle Women	60
Swimming 4x100m Medley Relay Men	60

Olympic Name	Num of Medals
London 2012	962
Beijing 2008	958
Athens 2004	929

The question is whether these two-column tables make sense. If we look at the country, it makes sense to see how many medals Australia has received in all Olympic Games; the same with Sport, and Olympic Names. Finding how many medals (regardless the medal types) for each country, for each sport, and for each Olympic seems to be reasonable.

The fourth two-column table for version-2 star schema is the Medal Type, which is as follows:

Medal Type	Num of Medals
Gold	4115
Silver	4095
Bronze	4474

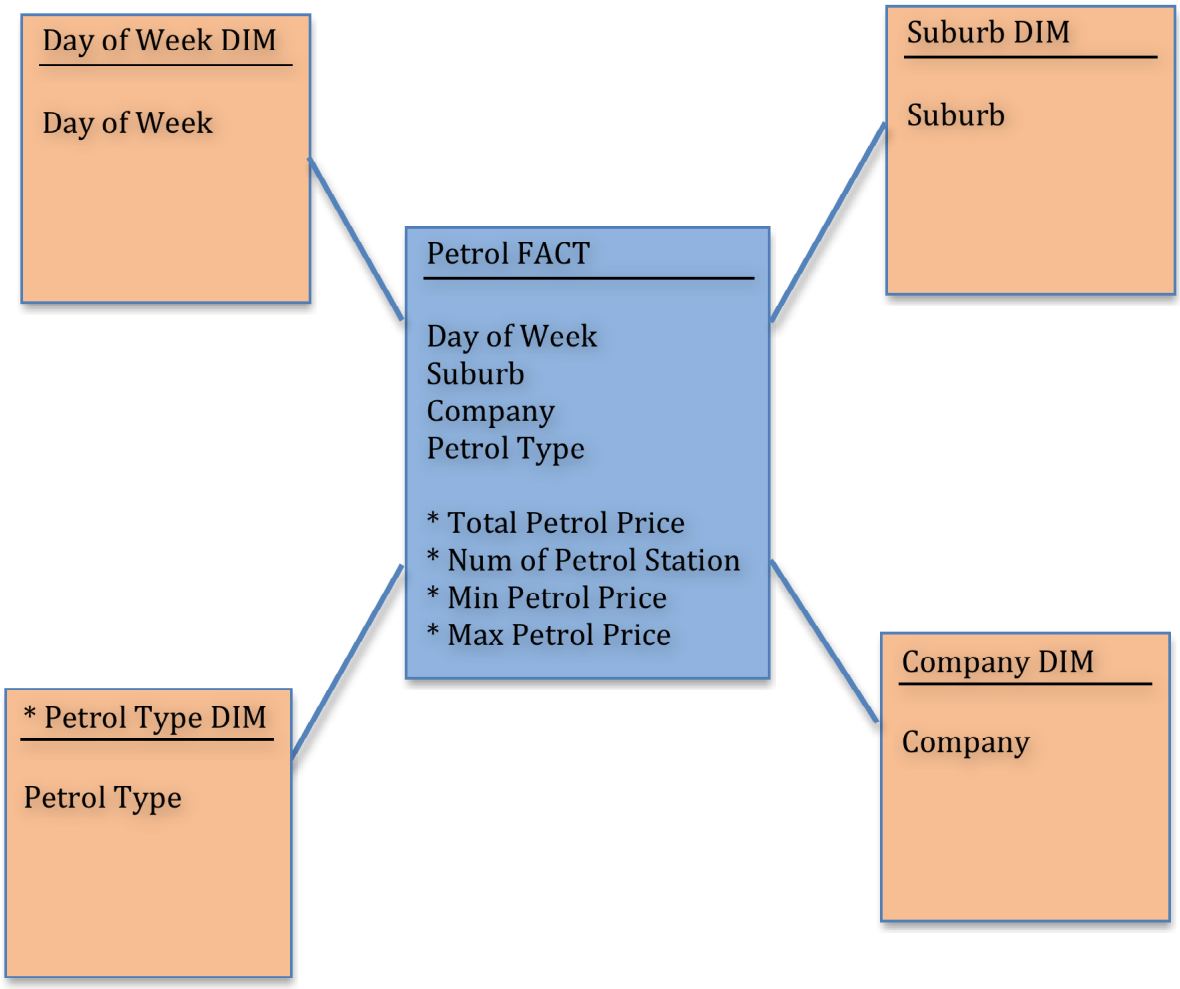
This two-column table on Medal Type seems to be reasonable too. Hence, version-2 star schema (with Medal Type DIM) is correct.

So in conclusion, both star schemas (with or without Medal Type DIM) are correct. Now going back to the original question: *Is Medal Type DIM a Determinant Dimension*?

4. Determinant or Non-Determinant Dimensions

A "*Determinant Dimension*" is a dimension that the fact measure relies on, and consequently, all data retrieval from the data warehouse must include this dimension. If the data retrieval from the data warehouse does not include this determinant dimension, the retrieval result will not make sense at all.

In the previous case study on Petrol Price, Petrol Type DIM is a Determinant Dimension (refer to the star schema below). Note that a Determinant Dimension is denoted by a star. The fact measures affected by the determinant dimension are also starred.



The fact measures: Total Petrol Price, Num of Petrol Station, Min Petrol Price, and Max Petrol Price, depend on the Petrol Type, which is indicated by the Petrol Type DIM. That means analyzing the min petrol price from the day of week point of view, must include the petrol type. Otherwise, it doesn't make sense to retrieve data to show that on Monday the lowest (min) petrol price is, for example, 109.90cents. As it does not indicate which petrol type it is, this lowest petrol price is meaningless. Therefore, a better data retrieval is to retrieve the record to show that for example, on Monday, the min "Unleaded" petrol price is 109.90cents (e.g. Unleaded is a petrol type obtained from the Petrol Type DIM).

Now going back to the Olympic Games case study (refer to Version-2 star schema with Medal Type DIM). *Is Medal Type DIM a Determinant Dimension*?

The answer to this question can be answered by another question. To retrieve the data from version-2 star schema, must we have the information from Medal Type DIM? The answer is clearly no, because we can simply retrieve a record from the fact to show that Australia in London 2012 Olympic Games received 10 medals in Swimming. This covers three dimensions, namely Country (Australia), Olympic Name (London 2012), and Sport (Swimming). In this example, Medal Type DIM is not involved, and the information retrieved still makes sense.

So the answer to the question whether Medal Type DIM is a Determinant Dimension or not, the answer is clearly No!!

The next question is: what is the difference between the Olympic Games case study and the Petrol Price case study. Both are very similar, but the Olympic Games Medal Type DIM is not a determinant dimension, whereas the Petrol Type DIM is a determinant dimension.

The answer is the aggregate function used in the fact measure. In the Olympic Games case study, the fact measure function is COUNT, which is count of medals. The breakdown of the medals is gold, silver, and bronze; but the main aggregate function of the fact measure is number of medals, which is a count. If the fact measure is a count, then the dimension (e.g. Medal Type DIM) is not a determinant dimension, because we can still analyze the fact measure which is the total medals from other dimensions, without the medal type dimension.

On the other hand, the Petrol Price case study uses AVG, MIN, and MAX as the aggregation functions. Note that we do not store average as a fact measure, but total price and number of stations. These two fact measures will be used to calculate the average. Although the average is not explicitly stored in the fact, implicitly, the total price and number of stations represent the average. If the aggregate function to calculate the fact measures is not COUNT, then a determinant dimension is needed. In this case, Petrol Type DIM is hence a determinant dimension, because Min Petrol Price, for example, does not have any meaning without petrol type.

5. Version-1 (without Medal Type DIM) vs. version-2 (with Medal Type DIM)

As both versions in the Olympic Games case study are correct, let's compare and contract these two versions. In order to do this, let's have a look at the records in the respective fact tables.

The fact table for version-1 star schema (without Medal Type DIM) 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 Medal Type DIM) consists of 5 columns: 4 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

From the storage point of view, it is clear that version-1 is the winner. It has only 3 records, whereas in version-2, the same information is represented in 9 records.

From the modeling point of view, some may prefer version-2, because the model is concise and more compact.

When the number of fact measure is reasonably large (like in the Petrol Price case study), the star schema with a determinant dimension looks very slim and compact – hence, it is easy to understand. But consequently, the storage requirement increases as well.

In contrast, with many different petrol types, if the star schema does not use a determinant dimension, the number of attributes in the fact will dramatically increase, and the schema looks more complex and crowded; but the storage cost is lower.