Advanced Data Warehousing
Topic: Multi Fact

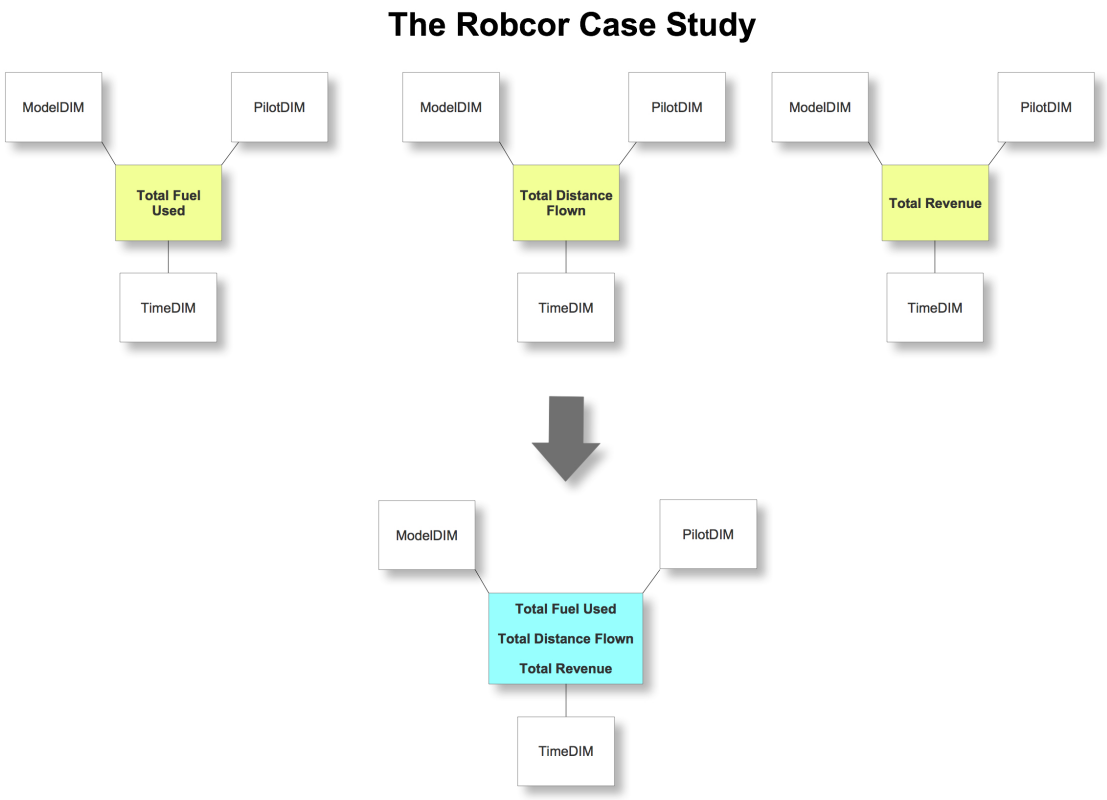
1. Introduction

Is it possible to have multi fact star schemas from one operational database?
Yes. Refer in Week-1 lecture, a data warehouse has four basic features: **(1) Integrated**, **(2) Subject Oriented**, **(3) Time Variant**, and **(4) Non-Volatile**. Pay a particular attention to the "Subject-Oriented" feature of a data warehouse. A subject-oriented data warehouse means that one star schema focuses on one subject only.

What is a "subject" (in the context of "Subject-Oriented") in a data warehouse?
A subject refers to a topic of analysis. For example, a star schema might be built for analysing sales of properties. If we want to have a data warehouse that also focuses on rental properties, it has to be a separate star schema, because one star schema focuses on one subject only. In this example, sales property is a subject, and property rental is another subject. The first star schema focuses on property sales, whereas the second star schema focuses on property rental. These two star schemas should not be combined, because they focus on a different subject. The input operational database for these star schemas might be one operational database.

Can a fact have more than one fact measures?
Yes. For example, the Robcor Charter Fact has three fact measures: Total Fuel Used, Total Distance Flown, and Total Revenue (Distance * Charge per Mile).

Does each of these fact measures focus on a different subject?
No. The reason is that these use the same dimensions, namely: TimePeriodDIM (MonthYear), PilotDIM (EmpNum), and AircraftModelDIM. If we had three separate facts: one for Total Fuel Used, another for Total Distance Flown, and another for Total Revenue, because these three facts have the same three dimensions, these facts are considered to share the same subject, and hence, the three facts should be combined as one fact with these three fact measures.



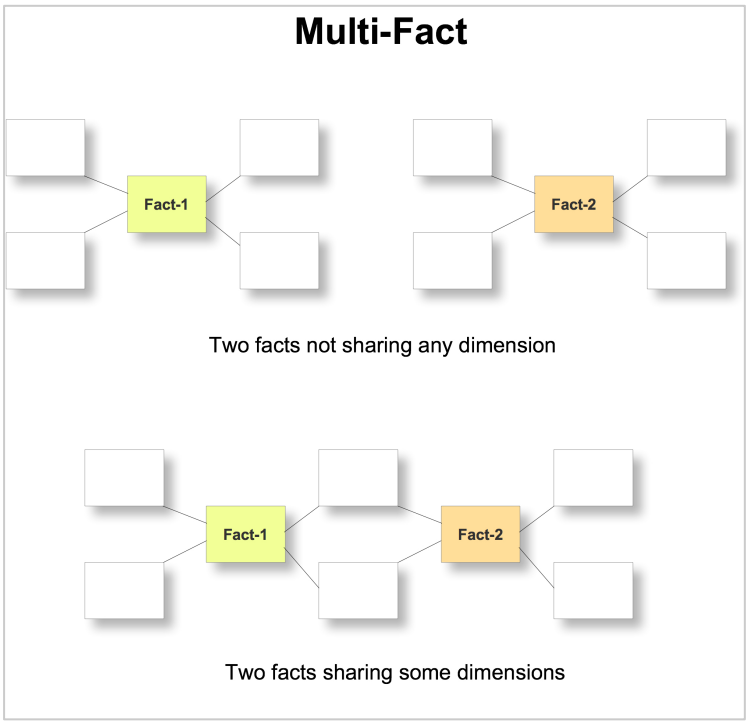
Can two star schemas having identical dimensions be merged into one star schema automatically?
No. If two star schemas having exactly the same dimensions, they may not be merged into one star schema automatically. In the robcor case study (refer to the data cleaning of robcor lecture notes), there are two star schemas: one for the pilot, and the other for the co-pilot. The dimensions for these two star schemas are identical, namely: ModelDIM, PilotDIM, and TimeDIM. The fact measures are also identical, namely: Total Fuel Use, Total Distance Flown, and Total Revenue.

If we merge the two star schemas by union-in the two fact tables, and re-calculating the three fact measures to incorporate the figures from the pilot and co-pilot, the results of the three fact measures will be incorrect. Therefore, the two star schemas cannot be merged into one star schema, although the two star schemas have the exactly identical dimensions and fact measures.

When can then two star schemas be merged into one star schema?
Two star schemas can be merged into one star schema if they follow the following two criteria: (1) the dimensions of the two star schemas are identical, but (2) the fact measures of the two star schemas must be different, and hence, there is not re-calculation of the fact measures after the two facts are merged into one table.

In the example above, if we had three star schemas for the robcor case study, and each star schema has one fact measure, which is different to each other (e.g. one fact has total fuel used, the second fact has total hours flown, and the third fact has total revenue), then these three facts can be combined into one fact.

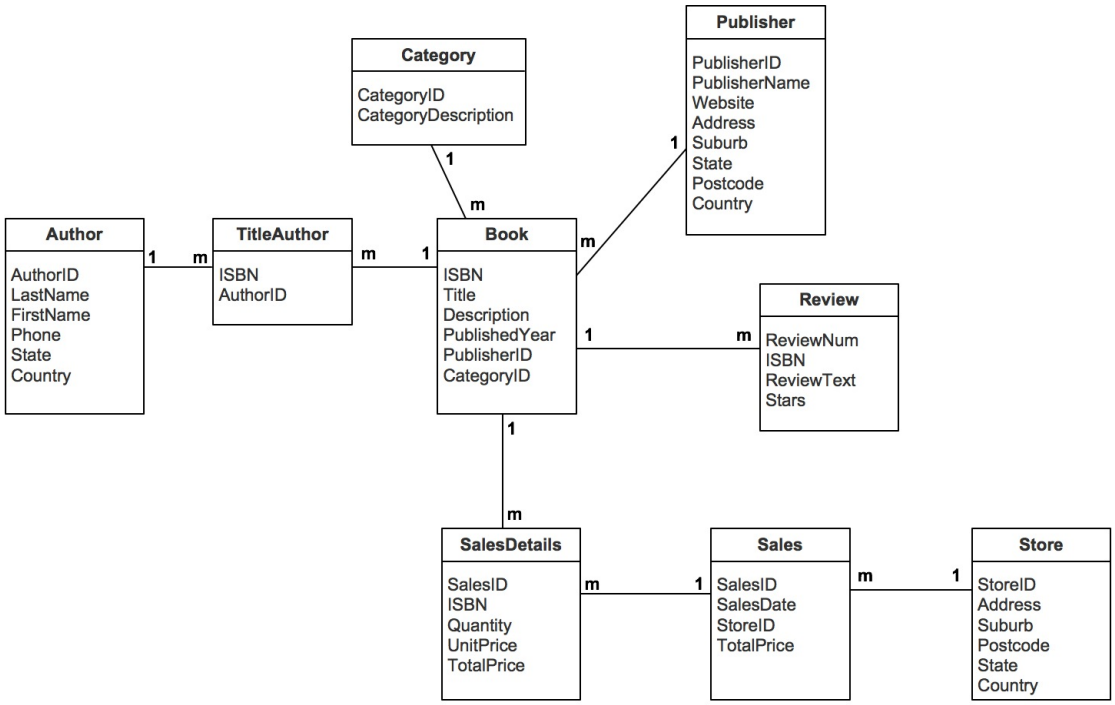
In case of two different star schemas, can dimensions be shared by two facts?
Yes. If different facts have the same dimensions, these dimensions can be shared by different facts. Because the analysis is based on the numerical measures stored in the fact, having shared dimensions does not change the nature of the data warehouse, in which a star schema focuses on one subject only – a data warehouse is still subject-oriented.



2. A Case Study – Book Sales

The following is an E/R diagram of an operational database. The system stores information about books, including the authors, publishers, book categories, as well as the reviews that each book has received. The 'stars' attribute in the Review entity records the star rating for each review (e.g. 5 stars for excellent to 1 star for poor, etc). One book may receive many reviews. For simplicity, it is assumed, as also shown in the E/R diagram, that a book will only have one category.

The E/R diagram also includes entities related to sales of books, and the stores which sale the books. Each store has many sales transactions (i.e. the Sales entity), and each sales transaction may include several books (i.e. the SalesDetails entity). The TotalPrice attribute in the SalesDetails entity is basically quantity multiplied by unitprice, whereas the TotalPrice attribute in the Sales entity is the total price for each sales transaction.



You are required to design a small data warehouse for analysis purposes. The analysis is needed for identifying at least the following questions:

- What are the *total sales* for each bookstore in a month?
- What is the *number of books* sold for each category?
- What is the book category that has the highest *total sales*?
- What is the *number of reviews* for each category?
- *How many* 5-star *reviews* for each category?

Based on the above requirements, it is clear that there are three fact measures:

1. Total sales,
2. Number of books sold, and

3. Number of reviews

Total sales and *number of books sold* would be useful for the management to understand how the book sales perform. *Number of reviews* would be useful for the marketing department to understand people's perception on certain books and would be able to use this information to launch any marketing campaign.

Potentially, there can be many dimensions. However, based on the above requirements, we limit the dimensions to the following four dimensions:

1. Store dimension,
2. Time dimension,
3. Book category dimension, and
4. Star rating dimension.

Let's examine the three fact measures (e.g. (*i*) total sales, (*ii*) number of books sold, and (*iii*) number of reviews) against the above four dimensions.

For the **store** dimension, we would like to find out total sales for each store, and number of books sold for each store. However, it doesn't make any sense if we ask how many reviews (or 5-star reviews) from customers to a certain store, because review's rating is not applicable to the store – it is only applicable to books. Therefore, the store dimension should not be connected to the fact measure: number of reviews.

For the **time** dimension (assuming we measure the time by month), we would like to find out total sales and number of books sold for each month, for example. This totally makes sense. To find out number of reviews given by customer in each month seems make sense too. However, if we inspect the E/R diagram closely, the month (or the date) is associated with the purchase of a book, not when a review is given by a customer. Consequently, time dimension should not be connected to the fact measure: number of reviews.

It seems clear now that the first two fact measures: total sales and number of books sold, are related to the book sales, whereas the third fact measure: number of reviews, is not about book sales, but is associated with reviews of the books. Therefore, these three fact measures are not within one common subject; in fact, they are two different subjects: **book sales**, and **book reviews**.

Let's continue with the third dimension: **book category** (or category). Finding out total sales and number of books sold for each book category seems to be sensible. For example, how many fiction novels were sold, and what is the total amount of sales for fiction novels. How about number of reviews? Finding out how many 5-star reviews for fiction novels is also reasonable, because each book (or each novel) will receive reviews, and therefore calculating how many reviews for each particular book category is also reasonable. In other words, the book category dimension is applicable to all of the three fact measures: total sales, number of books sold, and number of reviews.

Now the fourth dimension: **star rating** dimension, that describes the meaning of each star rating: from one star to five star, as an example. Finding out the total sales for books with 5-star ratings seems to be fine. Because one book may receive many different kinds of star ratings, we need to aggregate this, and come up with one star rating, such as 3.6 stars. And then this will be classified as 3.0 to 3.9 stars in the star rating dimension. Hence, the review star dimension is also applicable to all of the fact measures.

In summary, the first subject: "book sales" will be the first fact, with all the four dimensions (e.g. Store, Time, Category, and Star Rating dimensions). The fact measures are total sales, and number of books sold.

The second subject is "reviews", with only two dimensions: Category, and Star Rating dimensions. There is only one fact measure, which is number of reviews.

The star schema will then have two fact entities: Book sales fact, and Review fact. Note that the category and star rating dimensions are shared by the two facts.