# Assignment 7 (Group Project, part 2) - Dimensional Modeling

**Objectives**: Gain experience designing, implementing a dimensional database model in a relational database and based on transactional data. Analyse your transactional data from Assignment 5 and define and create one or more fact tables and one or more dimensions that will work together in a star schema. Implement SQL, Python, or R code that will extract the data from the transactional environment, transform it as appropriate, and load it into these new tables. Explore the data. Record the whole process in a notebook so it may be reproducible and easily communicated to others. Gain experience working collaboratively to achieve these goals.

**Grading criteria**: The tasks should all be completed. As everyone will work with data of their choosing, not everyone will have the same results. Describe your work in prose along with your code to inform the reader about the process, from sources of data through any decisions made about key concepts, techniques, and tools. The notebook itself should be reproducible; from start to finish, another person should be able to use the same code to obtain the same data and same results that you obtain, or at least similar data and similar results, should the data sources chosen vary over time.

Give your notebook a clear name like `assignment-07-our-last-names`. Document the contributions of each group member. Acknowledge any assistance you received. Zip your .ipynb along with any scripts or other files together, and upload your zipfile to Blackboard by the deadline. Only one group member should submit on behalf of your group.

**Deadline**: Tuesday, December 1, 7pm (before the start of class)

## Part 1 - Design a data warehouse from your transactional data

Starting from the same data and database your used in Assignment 5, analyse the records to determine which fact(s) and dimension(s) you wish to model and implement. Discuss your decisions, and explain the grain of your intended fact(s). Define a relational database model (two or more tables) that implements a star schema like those we studied in class. Include an explicit schema and create tables appropriately, and create a database model diagram if it might be helpful. Apps like Visio (MS) and Gliffy (free, on the web) have templates for ER diagrams.

## Part 2 - Extract, transform, and load data

Use SQL, Python, or R to pull data from the source tables into your new dimensional model tables. The week 10 lecture notebook provides examples of populating a dimensional model using only SQL. Note that you might need to look up documentataion on handling date/time functions and formats, for example, as these vary from one RDBMS to another (e.g. SQLite or MySQL) and from one programming language to another (e.g. Python or R). Whatever tools you use, document your entire process with code and describe each key step in prose.

## Part 3 - Explore data from data warehouse

With your star schema populated, perform some exploratory queries using SQL. Use each of the dimensions you created to bring out different facets of the processes measured in your fact tables. If any of your queries prove to be slow, use `EXPLAIN` to determine whether and where to add indexes, and create them, remembering to show again how the updated query plans should improve with the new indexes available. For this part, you might find it helpful to produce one or more plots.

Consider the benefits and potential drawbacks or missing pieces of your new dimensional model. How might it make the work of an analyst easier? Are any fields or dimensions missing? If this were a professional project, which additional tasks would be required to perform this work well, and what other data might you want to combine with the data you already have to support more powerful analyses?