Data Warehousing

Project Technical Requirements

I anticipate that a range of topics will be covered. The following minimum technical guidelines must be met on every project for the project to be accepted. Note too these are minimum requirements. Projects meeting only the minimum level of difficulty may not be scored as high as more complex projects assuming both are technically correct.

- Time Variance – Your data warehouse must contain a time dimension. Your source data must also therefore be time-variant.

- Current Time –If selecting a data source not from the list, your data must be current. This means that your project must contain the most recent data available for a given data set. If you, for example, are doing a stock market data warehouse, then you must use stock market data that is very recent (the last several months or years) because it will be available. The most recent data for some projects may be years old, which is ok provided it is the most current data available.

- Fact Size – You need to have a minimum of 50,000 fact records in your warehouse to ensure a good sample across the dimension values. Larger fact table are encouraged, but limit your project to no larger than about 1 million records so it can be submitted easily and you don't run into performance issues. Some projects may benefit from having more than one complementary fact table as well, which is also encouraged if the business case supports it.

- Number of Measures – you need to have a minimum of two measure fields in your fact table. (One measure can be the count of rows in the fact table if it satisfies a business question.)

- Dimension tables – you need to have a minimum of four dimension tables. (It is not sufficient to simply break out fields that should be in the same dimension into multiple dimensions.) If at all possible, dimension tables should have at least one hierarchy defined. Most dimension tables in real deployments would have multiple hierarchies.

- Hierarchies – you must have at least one non-time related hierarchy. If you source does not include this data, it can be generated. Such generation must be documented.

- Type II Dimension – at least two dimensions must contain a Type II dimension field. You must correctly demonstrate at least one change in value(s) in each field over time, but it is acceptable if this change is manually constructed (i.e. faked).

- Type II Dimension Change - you must simulate a Type II dimension change in at least one field of one dimension table. It does not need to be an actual data change, and it is okay to manually change the data. You must have rows from both the new and old values for the Type II change in the fact table for appropriate time periods that make sense. The goal of this step is to make sure you are familiar with the concept of what it means to be a Type II dimension.

- Fabricating Data – it is not acceptable to create the entire data set. It is too difficult to create a good set with a large number of fact records over a range of dimension values. It is ok to make up data attributes that you cannot find, however. If you think the weather would be an interesting indicator to analyze in relation to bus schedules but it's not available, it is perfectly acceptable to create it as part of a larger project. I would rather you present a more interesting business case and make up some data, than allow the data to limit your business case.