Facts and Dimensions

# Fact and Dimension Tables

## What is a Fact Table?

A data mart exists to provide users with the ability to analyze specific business processes. Fact tables exist to measure information about those business processes. Fact tables provide **quantitative** information on the business process being described. For example, if we were tracking sales, we may want to know things like price, quantity, total sale price, discount percentage, or net margin. These values are known as **measures**.

Fact tables capture a view of the data at a specific time. This is typically either related to an event, such as a financial transaction, or is a snapshot of the data at a specific time.

A **fact table** provides quantitative data on a specific event or moment in time. It is used to measure the business process that is to be analyzed.  
  
A **measure** is a column in a fact table that contains quantitative data to be tracked.

## What is a Dimension Table?

Dimension tables help define the granularity of the fact tables and give the fact tables context. They contain **qualitative** information on the business process being described. For example, if we were tracking sales, we may want to know things like what product was sold, where was it sold, when was it sold, who sold it, or who bought it.

Unlike fact tables, dimensions are usually persistent, though they may change over time. For example, one salesperson may be associated with thousands of sales over several years.

A **dimension table** provides qualitative data, typically on a person, place, or thing. It provides context that can be used to analyze facts.

## Granularity

Granularity refers to the level of detail available in the data. More detailed data is referred to as being **fine grain**, and more aggregated data is referred to as being **coarse grain**. In data warehousing, increasing granularity provides more details, resulting in a finer grain of data.

The granularity of a fact table is affected by the number of dimensions it is associated with. For example, we could track sales by store, product, salesperson, and/or date. Each additional dimension we add increases the granularity of the fact table – sales by store and date would be coarser than sales by product, store and date.

The granularity of a fact table is also affected by the types of dimensions it is associated with. For example, we might define date as being daily or monthly. Daily data would be finer grain than monthly.

There are a few important things to keep in mind when considering granularity:

1. **The granularity of your fact table defines its maximum level of detail**

Fine grain details can be “rolled up” to a coarser grain by using aggregate functions. The opposite is not true. You can only “drill down” into a finer grain of detail if those details exist in the table.

1. **All records in a fact table should exist at the same granularity**For example, in a fact table that tracked sales by salesperson, store, and day, you would not include the profit margin for the entire store, or the total sales for the month. If you need to track those details and couldn’t match the granularity of the existing table, you should create a new table.
2. **Increasing granularity increases the number of rows**

Finer grain details require more storage and processing because they generate more rows.

1. **Granularity is constrained by your sources**Because it is not usually possible to increase the granularity of your data, your maximum granularity is constrained by your source data.

# Creating a Dimensional Model

## Basic Steps

There are four basic steps to creating a new dimensional data model:

1. **Identify the process to be tracked**

Clarify the business process you would like to analyze. This helps identify which fact tables might be required to satisfy the requirements for the data mart.

1. **Identify the granularity**A fact table’s granularity is largely defined by its relationships with dimension tables, so a clear statement of the required granularity helps identify the dimensions needed.
2. **Define Dimensions**From the required granularity, we can derive a list of required dimensions. We also define the properties of each dimension, which are the qualitative data that we would like to use to analyze our facts.
3. **Define Facts and Measures**

Define the facts using dimension keys and required measures. Measures are quantitative information that may be aggregated and analyzed using dimensions.

In an ideal scenario, the required data would be available with the proper level of detail from one or more sources. In practice, we are often constrained by our sources and our design is often heavily influenced by what is available to us.

Defining a dimensional model is an iterative process that requires consultation with subject matter experts (SMEs). It also requires consultation with business analysts (BAs) who represent the end-users of the data.

## Step by Step

Over the next few lessons we will be building a very simple data mart using WideWorldImporters as a data source. We will first determine a general structure for our data mart. In later lessons we will learn how to build and populate it.

### Requirements

We have been given the following general requirements.

*We would like to track sales information for WideWorldImporters. We would like to analyze which products are being ordered to determine how things like brand, colour, and price may impact gross sales. In particular, we would like to identify preferences for specific customers so that we may more effectively predict which products they may be interested in. Similarly, we would like to be able to determine if certain cities have brand, colour, and price preferences. We would also like to determine whether certain salespeople have more success selling certain products, selling to certain customers, or selling in certain cities.*

### Business Processes and Granularity

Properly defining both our business process and granularity would may require consultation with our BAs and/or SMEs. By the end of it, we should have a strong understanding of the process to be modelled and have a simple statement that serves to define our granularity.

At its simplest, we are tracking customer orders. We might define our maximum granularity as “An order for one or more of a stock item by a customer in a city via a sales person.”

### Define Dimensions

Similar to entities in an ER model, dimensions are often the nouns in our requirements. Specifically, we should be looking for nouns that help us define our business process and its granularity, so properly identifying the granularity will help a great deal in identifying dimensions.

Let’s review again and highlight anything that may be a good fit. We can reduce the list later.

*We would like to track sales information for WideWorldImporters. We would like to analyze which products are being ordered to determine how things like brand, colour, and price may impact gross sales. In particular, we would like to identify preferences for specific customers so that we may more effectively predict which products they may be interested in. Similarly, we would like to be able to determine if certain cities have brand, colour, and price preferences. We would also like to determine whether certain salespeople have more success selling certain products, selling to certain customers, or selling in certain cities. We need to track the date of each order to be able to track the impact of our change over time.*

This leaves us with the following distinct concepts: **sale; product; brand; colour; price; customer; city; salesperson; date; order; time.**

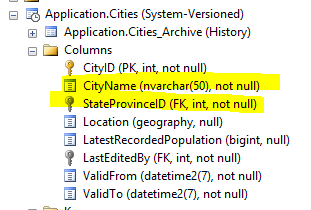
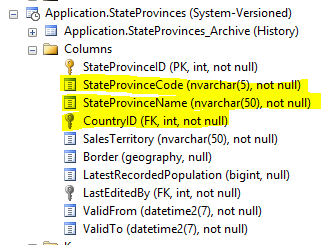
* Based on the description, **sales** are what we want to track, not something we want to use to describe the thing we are tracking. For our purposes, we have determined that **order** is synonymous with sale.
* **Brand** and **colour** sound more like properties of a product that we might use to analyze our sales.
* **Price** also sounds like a property of our product. As a clear quantity that could be associated with a sale, it might also make a good measure. In either case, it does not likely make a good dimension.
* Based on the wording of our requirements, **time** is certainly synonymous with **date**.
* Based on the requirements, it is debatable whether **city** is a property of a customer or a dimension on its own. Along with time, location is a very common dimension so we will have a bias towards defining it as a dimension instead of a property.

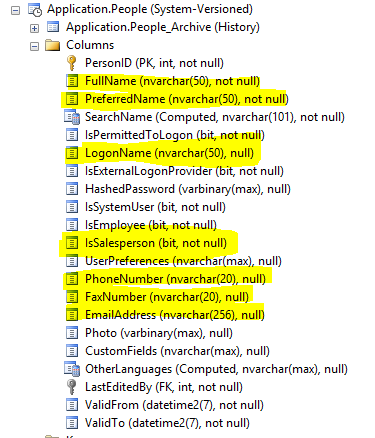
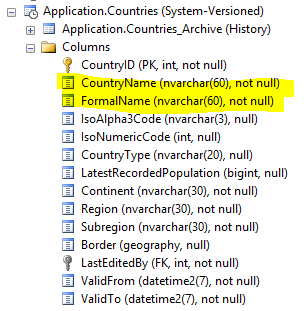
The two most popular dimensions is the date and the location of

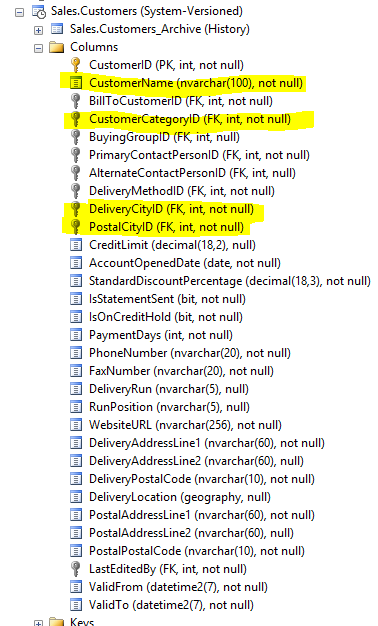
This leaves us with the following dimensions: **product; customer; city; salesperson; date**

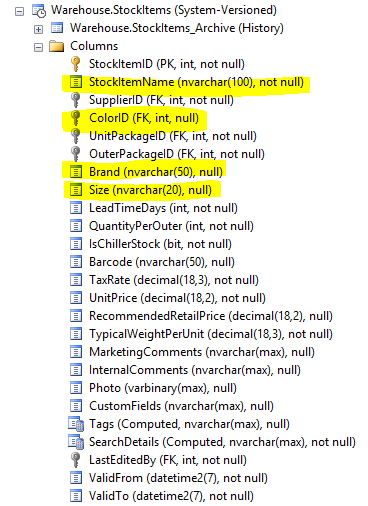
A few of our dimension properties have been identified in our requirements; products should include brand, colour, and maybe price. Aside from that, we have not been provided with much.

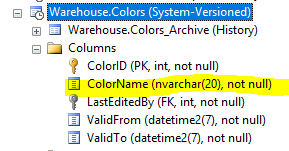
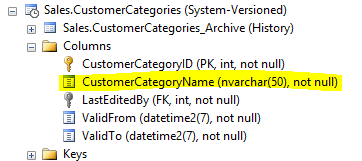
Getting a full list of appropriate properties will require consultation with your BAs and SMEs, but in practice our properties are typically defined by the sources we have available. Let’s look at some of the sources we have available to look for appropriate **qualitative** datato use as properties for our dimensions.









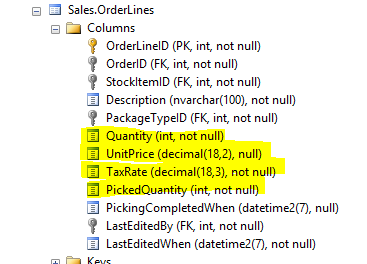
### Define Facts and Measures

Using our definition above, it is pretty clear what fact we want to track. We need to track **orders**, which in the requirements is also described as **sales**. The requirements use the terms interchangeably, so we will consider them one and the same.

Measures are quantities that we want to track about our fact. Our requirements don’t mention much aside from maybe **price** that could make a good measure. So where do we get our measures from?

In practice it requires collaboration with our BAs and SMEs, but it is often heavily influenced by the data that is available to us. Let’s look at our likely source tables for anything that might make a good measure. Since measures are quantitative data, we’re usually looking for numeric data.

Based on the data being requested, Sales.OrderLines seems like our best bet. Excluding IDs, we have the following numeric columns:



**Quantity, UnitPrice,** and **TaxRate** all seem like a good fit for the kind of analysis described. **PickedQuantity** is the quantity picked from stock, so that doesn’t seem like a good fit for our fact table.

In the requirements, we also have the following statement:  
  
“…*determine how things like brand, colour, and price may impact* ***gross sales****.”*

In this context, gross sales seems to be a qualitative measure; specifically, the total price of an order. It is unclear from this context whether that is intended to include taxes or not, but it is reasonable to expect that both may be required at some point. In a normalized ER data model, these would violate the third normal form. In a dimensional model, it is acceptable to include them for performance and ease of use.

## Result

Let’s look at the model we have produced. We will be reviewing keys in a data mart in another class, so we will exclude them from now. We will also exclude any dimension properties for the sake of brevity.

