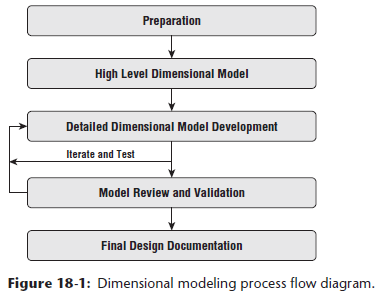
# Kimball Data Warehouse Toolkit

## Ch 18 – Dimensional Modeling Process and Tasks

* Moving past dimensional modeling *patterns*, now it’s time to turn attention to the **tasks + tactics** of the dimensional modeling process.
* This chapter, condensed from content in *The DW Lifecycle Toolkit, Second Edition (Wiley, 2008),* begins with a practical discussion of preliminary prep activities, such as identifying the participants (including business reps) + arranging logistics
* The modeling team develops an initial high-level model diagram, followed by iterative detailed model development, review, + validation
* **Throughout the process, you are reconfirming your understanding of the business’s requirements**
* **Concepts:**
* Overview of the dimensional modeling process
* Tactical recommendations for the modeling tasks
* Key modeling deliverables

### Modeling Process Overview

* Before launching into the dimensional modeling design effort, you **must involve the right players**
* Strongly encouraged: **participation of business reps during the modeling sessions**
* Their involvement + collaboration **strongly increases the likelihood that the resultant model addresses the business’s needs**
* Likewise, the organization’s **business data stewards** should participate, *especially when discussing the data they’re responsible for governing*
* Creating a dimensional model is a **highly iterative + dynamic process**
* After a few prep steps, the **design effort begins w/ an initial graphical model derived from the bus matrix**, **identifying the scope** of the design + **clarifying the grain** of the proposed fact tables + associated dimensions
* After completing the high-level model, the design team dives into the dimension tables with attribute definitions, domain values, sources, relationships, data quality concerns, + transformations
* After the dimensions are identified, the fact tables are modeled
* The last phase of the process involves reviewing + validating the model w/ interested parties, especially business reps
* **Primary goals = create a model that meets the business requirements, verify that data is available to populate the model, + provide the ETL team w/ a solid starting source-to-target mapping**
* Dimensional models unfold through a series of **design sessions** with each pass resulting in a more detailed + robust design that’s been repeatedly tested against the business needs
* The process is complete when the model clearly meets the business’s requirements
* A typical design requires 3-4 weeks for a single business process dimensional model, but time required can vary depending on team’s experience, availability of detailed business requirements, involvement of business reps or data stewards authorized to drive to organizational consensus, complexity of the source data, + the ability to leverage existing conformed dimensions
* The **key inputs** to the dimensional modeling process = the **preliminary bus matrix + detailed business requirements**
* The **key deliverables** of the modeling process = the **high-level dimensional model**, **detailed dimension + fact table designs**, and **issues log**
* The figure below shows the dimensional modeling process flow



* Although the image above portrays a linear progression, the **process is quite iterative**.
* You will make multiple passes through the dimensional model starting at a high level + drilling into each table + column, filling in the gaps, adding more detail, + changing the design based on new information
* If an outside expert is engaged to help guide the dimensional modeling effort, insist they facilitate the process *with the team* rather than disappearing for a few weeks + returning with a completed design
* This ensures the entire team understands the design + associated trade-off’s
* It also provides a learning opportunity, so the team can carry the model forward + independently tackle the next model.

### Get Organized

* Before beginning to model, you must appropriately prepare for the dimensional modeling process
* In addition to involving the right resources, there are also basic logistical considerations to ensure a productive design effort

#### Identify Participants, Especially Business Representatives

* The best dimensional models result from a **collaborative team effort**
* *No single individual is likely to have the detailed knowledge of the business requirements + the idiosyncrasies of the source systems to effectively create the model themselves.*
* Although the data modeler facilitates the process + has primary responsibility for the deliverables, we believe it’s critically important to get SMEs from the business involved to actively collaborate
* Their insights are invaluable, especially because they are often the individuals who have historically figured out how to get data out of the source systems + turned it into valuable analytic information
* Although involving more people in the design activities increases the risk of slowing down the process, the **improved richness + completeness of the design justifies the additional overhead**
* It’s always helpful to have someone w/ keen knowledge of the source system realities involved
* Might also include some physical DBA and ETL team reps so they can learn from the insights uncovered during the modeling effort + resist the temptations to apply 3NF concepts or defer complexities to the BI applications in an effort to streamline the ETL processing
* **Remember: the goal is to trade off ETL processing complexity for simplicity + predictability at the BI presentation layer**
* Before jumping into the modeling process, take time to consider the *ongoing* stewardship of the DW/BI environment
* If the organization has an active data governance + stewardship initiative, it is time to tap into that function
* If there is no preexisting stewardship program, it’s time to initiate it
* An enterprise DW/BI effort committed to dimensional modeling must also be committed to a **conformed dimension strategy to ensure consistency across business processes**
* An active data stewardship program helps the organization achieve its conformed dimension strategy
* Agreeing on conformed dimensions in a large enterprise can be a challenge, + the difficulty is usually less a technical issue + more an organizational communication + consensus building one
* Different groups across the enterprise are often committed to their own proprietary business rules and definitions
* Data stewards must work closely w/ the interested groups to develop common business rules + definitions, + then cajole the organization into embracing the common rules and definitions to develop enterprise consensus
* Over the years, some have criticized the concept of conformed dimensions as being “too hard.”
* Yes, it’s difficult to get people in different corners of the business to agree on common attribute names, definitions, + values, but that’s the crux of unified, integrated data
* **If everyone demands their own labels + business rules, then there’s no chance of delivering the single version of the truth promised by DW/BI systems**
* And finally, one of the reasons the Kimball approach is sometimes criticized as being hard from people who are looking for quick solutions is because they have spelled out the detailed steps for actually getting the job done

#### Review the Business Requirements

* **Before the modeling begins, the team must familiarize itself with the business requirements**
* 1st step = carefully review the **requirements documentation** (Chapter 17)
* It’s the **modeling team’s responsibility to translate the business requirements into a flexible dimensional model that can support a broad range of analysis**, *not just specific reports.*
* Some designers are tempted to skip the requirements review + move directly into the design, but the resulting models are typically driven exclusively by the source data without considering the added value required by the business community
* Having appropriate business representation on the modeling team helps further avoid this data-driven approach

#### Leverage a Modeling Tool

* Before jumping into the modeling activities, it’s helpful to have a few tools in place.
* Using a spreadsheet as the initial documentation tool is effective because it enables you to quickly + easily make changes as you iterate through the modeling process
* After the model begins to firm up in the later stages of the process, you can convert to whatever modeling tool is used in your organization
* Most modeling tools are dimensionally aware w/ functions to support the creation of a dimensional model
* When the detailed design is complete, the modeling tools can help the DBA forward-engineer the model into the database, including creating the tables, indexes, partitions, views, + other physical elements of the database

#### Leverage a Data Profiling Tool

* **Throughout the modeling process, the teams needs to develop an ever-increasing understanding of the source data’s structure, content, relationships, + derivation rules**
* **Need to verify the data exists in a usable state, or at least its flaws can be managed, + understand what it takes to convert it into the dimensional model.**
* **Data profilinguses query capabilities to explore the actual content + relationships in the source system rather than relying on perhaps incomplete or outdated documentation**
* Can be as simple as writing some SQL statements or as sophisticated as a special purpose tool.
* The major ETL vendors include data profiling capabilities in their products

#### Leverage or Establish Naming Conventions

* The issue of **naming conventions** inevitably arises during the creation of the dimensional model
* The **data model’s labels must be descriptive + consistent *from a business perspective***
* **Table + column names become key elements of the BI application’s interface**
* A column name such as “Description” may be perfectly clear in the context of a data model but communicates nothing in the context of a report
* Part of the process of designing a dimensional model is agreeing on common definitions + labels
* **Naming is complex because different business groups have different meanings for the same name and different names with the same meaning**
* People are reluctant to give up the familiar and adopt a new vocabulary.
* Spending time on naming conventions is one of those tiresome tasks that seem to have little payback but is worth it in the long run
* Large organizations often have an IT function that owns responsibility for naming conventions
* A common approach is to use a naming standard with 3 parts: **prime word**, **qualifiers** (if appropriate), + **class word**
* Leverage the work of this IT function, understanding that sometimes existing naming conventions need to be extended to support more business-friendly table + column names
* If the organization doesn’t already have a set of naming conventions, establish them during the dimensional modeling

#### Coordinate Calendars and Facilities

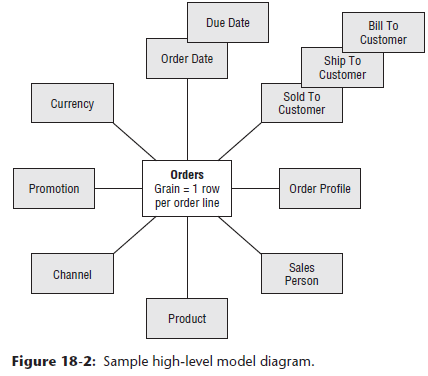
* Last, but not least, you need to schedule the design sessions on participants’ calendars.
* Rather than trying to reserve full days, it’s more realistic to schedule morning + afternoon sessions that are 2-3 hours in duration for 3 or 4 days each week
* This approach recognizes that the team members have other responsibilities + allows them to try to keep up in the hours before, after, + between design sessions
* The design team can leverage the unscheduled time to research the source data + confirm requirements, as well as allow time for the data modeler to update the design documentation prior to each session
* As mentioned earlier, the modeling process typically takes 3-4 weeks for a single business process, such as sales orders, or a couple of tightly related business processes such as healthcare facility + professional claim transactions in a set of distinct but closely aligned fact tables
* There are a multitude of factors impacting the magnitude of the effort
* Ultimately, the **availability of previously existing core dimensions allows the modeling effort to focus almost exclusively on the fact table’s performance metrics, which significantly reduces the time required**
* Finally, you must reserve appropriate facilities
* It is best to set aside a dedicated conference room for the duration of the design effort (no easy task in most organizations where meeting room facilities are always in short supply)
* Whiteboards would be nice, too, + some basic supplies

### Design the Dimensional Model

* Chapter 3 🡪 **4 key decisions made during the design of a dimensional model**:
* **Identify** the **business process**.
* **Declare** the **grain** of the business process.
* **Identify** the **dimensions**.
* **Identify** the **facts**.
* 1ST step of identifying the business process = typically determined at the conclusion of the requirements gathering
* The prioritization activity described in Chapter 17 establishes which bus matrix row (+ hence business process) will be modeled
* With that grounding, the team can proceed with the design tasks.
* The modeling effort typically works through the following sequence of tasks + deliverables, as illustrated flow shown earlier
* **High-level model** defining the model’s **scope + granularity**
* **Detailed design** with table-by-table attributes + metrics
* **Review and validation** with IT + business representatives
* **Finalization** of the design documentation
* As with any data modeling eff ort, dimensional modeling is **an iterative process**.
* You will work back + forth between business requirements + source details to further refine the model, changing the model as you learn more
* Depending on the design team’s experience + exposure to dimensional modeling concepts, you might begin w/ basic dimensional modeling education before kicking off the effort to ensure everyone is on the same page regarding standard dimensional vocabulary + best practices

#### Reach Consensus on High-Level Bubble Chart

* **Initial task in the design session = create a high-level dimensional model diagram for the target business process**
* Creating the **1st draft is relatively straightforward b/c you start with the bus matrix**
* Although an experienced designer could develop the initial high-level dimensional model + present it to the team for review, this does not allow the entire team to participate in the process
* The high-level diagram graphically represents the business process’s dimension + fact tables
* Shown below, can refer to this diagram as the **bubble chart**for obvious reasons



* This entity-level graphical model **clearly identifies the grain of the fact table + its associated dimensions to a non-technical audience**
* **Declaring the grain requires the modeling team to consider what is needed to meet the business requirements + what is possible based on the data collected by the source system**
* The **bubble chart *must* be rooted in the *realities of available physical data sources***
* A single row of the bus matrix may result in multiple bubble charts, each corresponding to a unique fact table with unique granularity
* **Most of the major dimensions will fall out naturally after you determine the grain**
* One of the **powerful effects of a clear fact table grain declaration is you can precisely visualize the associated dimensionality**
* **Choosing the dimensions may also cause you to rethink the grain declaration**
* If a proposed dimension doesn’t match the grain of the fact table, either the dimension must be left out, the grain of the fact table changed, or a multivalued design solution needs to be considered
* The above graphical representation serves several purposes
* Facilitates discussion within the design team before the team dives into the detailed design, ensuring everyone is on the same page before becoming inundated with minutiae
* Also is a helpful introduction when the team communicates with interested stakeholders about the project, its scope, + data contents.
* To aid in understanding, it is helpful to **retain consistency across the high-level model diagrams for a given business process**
* Although each fact table is documented on a separate page, arranging the associated dimensions in a similar sequence across the bubble charts is useful

#### Develop the Detailed Dimensional Mode

* After completing the high-level bubble chart designs, it’s time to focus on the details.
* The **team should meet on a very regular basis to define the detailed dimensional model, table by table, column by column**
* The **business reps should remain engaged during these interactive sessions**, as you **need their feedback on attributes, filters, groupings, labels, + metrics**
* **Most effective to start with the dimension tables and then work on the fact tables**
* **Launch the detailed design process with a couple of straightforward dimensions**, + the date dimension is always a favorite starting point
* **Enables the modeling team to achieve early success, develop an understanding of the modeling process, + learn to work together as a team.**
* The **detailed modeling identifies the interesting + useful attributes w/in each dimension + appropriate metrics for each fact table**
* Also want to **capture the sources, definitions, + preliminary business rules that specify how these attributes + metrics are populated**
* **Ongoing analyses of the source system + systematic data profiling during the design sessions helps the team better understand the realities of the underlying source data**

##### Identify Dimensions and their Attributes

* **During the detailed design sessions, key conformed dimensions are defined**
* Because the DW/BI system is an *enterprise* resource, **these definitions must be acceptable *across the enterprise***
* The data stewards + business analysts = key resources to achieve organizational consensus on table and attribute naming, descriptions, + definitions
* The design team can take the lead in driving the process + leveraging naming conventions, if available, but **it is ultimately a *business* task to agree on standard business definitions + names, as the column names must make sense to the business users**
* Can take some time, but it is **an investment that will deliver huge returns for the users’ understanding + willingness to accept the dimensional model**
* Don’t be surprised if the governance steering committee must get involved to resolve conformed dimension definition + naming issues.
* At this point, the modeling team often also wrestles w/ the *potential* inclusion of **junk dimensions** or **mini-dimensions** in a dimensional model
* It may not be apparent that these more performance-centric patterns are warranted until the team is deeply immersed in the design

##### Identify the Facts

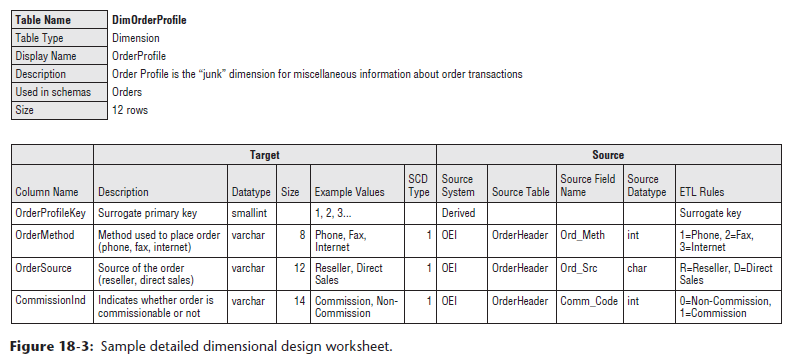
* **Declaring the grain crystallizes the discussion about the fact table’s metrics because the facts must all be true to the grain**
* The **data profiling effort identifies the counts + amounts generated by the measurement event’s source system**
* However, **fact tables are not limited to just these base facts**
* There **may be additional metrics the business wants to analyze that’re derived from the base facts**

##### Identify Slowly Changing Dimension Techniques

* After the dimension + fact tables from the high-level model diagram have been initially drafted, you then **circle back to the dimension tables**
* **For each dimension table attribute, define how source system data changes will be reflected in the dimension table**
* Again, **input from the business data stewards is critical to establishing appropriate rules**
* Also helpful to ask the source system experts if they can determine whether a data element change is due to a source data correction

##### Document the Detailed Table Designs

* The **key deliverables of the detailed modeling phase are the design worksheets**, as shown below



* The **worksheets capture details for communication to interested stakeholders including other analytical business users, BI application developers, + most important, the ETL developers who will be tasked w/ populating the design.**
* **Each dimension + fact table should be documented in a *separate* worksheet**
* At a minimum, the supporting information required includes attribute/fact name, description, sample values, + a SCD type indicator for every dimension attribute
* In addition, the detailed fact table design should identify each FK relationship, appropriate degenerate dimensions, + rules for each fact to indicate additive, semi, or non-additive
* **Dimensional design worksheet = 1st step toward creating the source-to-target mapping document**
* The *physical design* team will further flesh out the mapping w/ physical table + column names, data types, + key declarations

##### Track Model Issues

* Any **issues, definitions, transformation rules, + data quality challenges discovered** during the design process should be captured in an **issues tracking log**
* Someone should be assigned the task of capturing + tracking issues during the sessions, + the PM (if participating in the design sessions) often handles this responsibility because they’re typically adept at keeping the list updated + encouraging progress on resolving open issues
* The facilitator should reserve adequate time at the end of every session to review and validate new issue entries + their assignments
* Between design sessions, the design team is typically busy profiling data, seeking clarification + agreement on common definitions, + meeting w/ source system experts to resolve outstanding issues

##### Maintain Updated Bus Matrix

* **During the detailed modeling process, there are often new discoveries about the business process being modeled**
* Frequently, these findings **result in the introduction of new fact tables to support the business process, new dimensions, or the splitting or combining of dimensions**
* **Must keep the bus matrix updated throughout the design process because it is a key communication + planning tool**
* Chapter 16 🡪 the ***detailed* bus matrix often captures *additional* information about each fact table’s granularity + metrics**

#### Review + Validate the Model

* Once the design team is confident about the model, the process moves into the **review + validation phase to get feedback from other interested parties**, including:
* IT resources, such as DW/BI team members not involved in the modeling effort, source system experts, + DBAs
* Analytical or power business users not involved in the modeling effort
* Broader business user community

##### IT Reviews

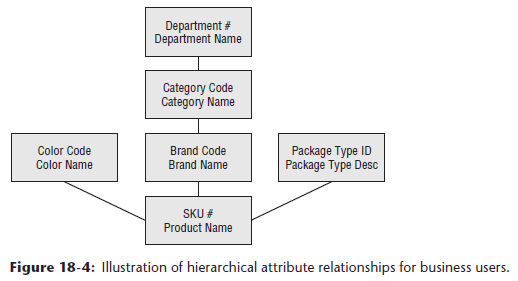
* Typically, the **1st review** of the detailed dimensional model is with **peers in the IT organization**
* This audience is **often composed of reviewers who are intimately familiar with the target business process because they wrote or manage the system that runs** **it**
* Are **also at least partly familiar with the target data model because you’ve already been pestering them with source data questions**
* IT reviews **can be challenging because the participants often lack an understanding of dimensional modeling**
* In fact, most of them probably fancy themselves as proficient 3NF modelers + their tendency will be to apply transaction processing-oriented modeling rules to the dimensional model
* Rather than spending the bulk of your time debating the merits of different modeling disciplines, it is **best to proactively provide some dimensional modeling education as part of the review process**
* When everyone has the basic concepts down, **begin with a review of the bus matrix, which gives everyone a sense of the project scope + overall data architecture, demonstrates the role of conformed dimensions, + shows the relative business process priorities**
* Next, illustrate how the selected row on the matrix translates directly into the high-level dimensional model diagram
* This gives everyone the entity-level map of the model + serves as the guide for the rest of the discussion
* **Most of the review session should be spent going through the dimension + fact table worksheet details**
* It is also a good idea to **review any remaining open issues for each table as you work through the model**
* **Changes to the model will likely result from this meeting**
* Remember to assign someone to capture the issues + recommendations

##### Core User Review

* In many projects, this review is NOT required because the core business users are members of the modeling team + are already intimately knowledgeable about the dimensional model
* Otherwise, this review meeting is similar in scope + structure to the IT review meeting
* Core business users = more technical than typical business users + can handle details about the model
* In smaller organizations, can often combine the IT review + core user review into one session

##### Broader Business User Review

* This session is as much education as it is design review
* Want to **educate people without overwhelming them, while at the same time illustrating how the dimensional model supports their business requirements**
* **Start with the bus matrix as the enterprise DW/BI data roadmap, review the high-level model bubble charts, + finally, review the critical dimensions, such as customer and product**
* Sometimes the bubble charts are supplemented with diagrams similar to below to illustrate the hierarchical drill paths within a dimension:



* Be sure to **allocate time during this education/review to illustrate how the model can be used to answer a broad range of questions about the business process**
* Can often **pull some examples from the requirements document + walk through how they would be answered**

#### Finalize the Design Documentation

* After the model is in its final form, the **design documentation should be compiled from the design team’s working papers**
* This document typically includes:
* Brief description of the project
* High-level data model diagram
* Detailed dimensional design worksheet for each fact and dimension table
* Open issues