

Better Storytelling

Dimensional design techniques that bind events into stories

Storytelling is, and always has been, a powerful means of sharing knowledge, because it simultaneously and elegantly satisfies several innate human needs – including the need for entertainment, sharing the accumulated wisdom of experience, and providing avenues of self-expression. In this edition of the Data Warehouse (DW) Architect, we will describe some design techniques that amplify the dimensional data warehouse's already robust abilities to "tell the stories" that are encased within the millions or billions of events sitting within its' subject areas – in a way that taps into this deep-rooted human need.

The DW Architect's challenge is to select dimensional approaches that allow smart, non-technical users to look across millions of overlapping customer stories that occur asynchronously, and to uncover their commonalities and patterns - while still maintaining the integrity of each individual event. We aspire to provide this capability using only simple, dimensionally friendly query tools, rather than needing PHD-level SQL or data mining techniques. In practice, this means adding new dimensional support to the detailed data in the dimensional data warehouse, allowing users to creatively nudge these out-of-phase events into novel alignments.

The example chosen to showcase these design techniques is the clickstream "Page Visit" subject area for a web retailer – where each fact represents a web site visitor's viewing of a page of web content - organized into a classic Kimball star schema. The techniques presented, however, should be applicable to any subject area in which one event is a part of a larger and interesting sequence of events (or "story") – such as multi-leg voyages and shipments, airline flights, telephone calls or network traffic routing, etc. More profoundly though, these techniques can also be applied to second-tier subject areas that consolidate multiple customer, prospect, or partner touch points – allowing analysts to scrutinize the patterns of exchanges that lead to changes in customer behaviors over time.

Trumping Time

Calendar "Time" is a dimension that is used in almost every report query – but viewing events *solely* in chronological order can veil predictable sequences of events at lower levels. Here are some dimensional techniques for dragging these micro-level sequences into the light of day – effectively doing on-the-fly realignment of events based on their common story milestones.

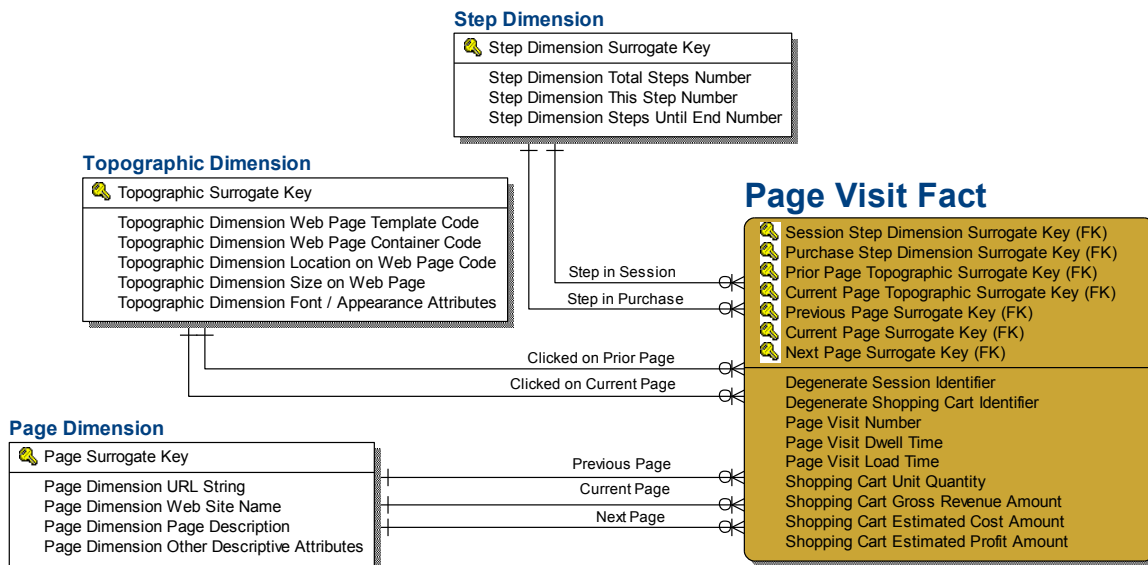


Figure 1 – Storytelling Dimensions

The “Step Dimension”, shown in Figure 1, is a spin on the Kimball “Hierarchy Helper Table” concept (see Ralph’s [“Help for Hierarchies”, 9/1998](#)). An example of the first few rows for a Step Dimension is shown in figure 2.

Step Dimension			
Step Dimension Surrogate Key	Total Steps Number	This Step Number	Steps Until End Number
1	1	1	0
2	2	1	1
3	2	2	0
4	3	1	2
5	3	2	1
6	3	3	0
7	4	1	3
8	4	2	2
9	4	3	1
10	4	4	0




Figure 2 – Populating the Step Dimension

It describes each event’s position within a story sequence by explicitly proclaiming: “This Page Visit is associated with Session Step Surrogate Key 10, which means that it is the fourth Page Visit of a session consisting of four total Page Visits – and is therefore zero Page Visits from the end”.

Users can constrain on and group by the Step dimension, and quickly identify the Page Visits belonging to sessions of any length, pages that are session killers (zero steps from the end), the pages that precede these session killers, etc. It allows analysts to quickly and easily realign Page Visits by their position within their respective sessions – to

answer business questions about the beginnings and endings of sessions for any subset of interesting customers, calendar time and promotions – a subject of endless inquiry for web retailers.

Including an additional "Purchase Path Step" role for the dimension allows analysts to easily scrutinize sequence of Page Visits that lead to "purchases" (the placement of something in a shipping cart). Because more than one purchase can occur in a session, the Purchase Path Step dimension should "rollover" (start back at step one) for page visits in the session that follow a shopping cart add. Page Visits that don't lead to a shopping cart add should point to a "Not Applicable" Purchase Path Step dimension instance.

Session Steps			Purchase Steps		
1 of 7	6 until end		1 of 3	2 until end	
2 of 7	5 until end		2 of 3	1 until end	
3 of 7	4 until end		3 of 3	0 until end	
4 of 7	3 until end		1 of 2	1 until end	
5 of 7	2 until end		2 of 2	0 until end	
6 of 7	1 until end		NA	NA	
7 of 7	0 until end		NA	NA	

Example: A session that consisted of seven total Page Visits, with purchases in the third and fifth Page Visits is represented in Figure 3:

Notice that by constraining on Purchase Steps that are one step from the end, only Page Visits that immediately precede Shopping Cart Adds will be returned – regardless

Figure 3 – Example of Session and Purchase Steps

of their position in the Session sequence or in absolute time. This is powerful medicine - because it allows simple query tools to identify and scrutinize the web page "pathways" that consistently lead to purchases – helping retailers to re-architect their sites and internal promotions to better drive web traffic into these "purchase vortexes".

The Step dimension could be used in any number of alternative roles to identify page visit steps that lead to abandoned shopping carts, etc. Techniques for using a common dimension in multiple roles are described more fully in Ralph's ["Data Warehouse Role Models" article from 8/1997](#).

The Step dimension is easily built in a spreadsheet, and pre-populating it with up to 300 total steps requires only about 45,000 rows – a reasonable dimension size that covers enough potential steps for most typical web site, travel, or network traffic analytic applications. A special "Over Max Steps" dimensional instance can be created to handle any exceptionally high step number situations that emerge.

The "Page Dimension" contains a row for each web page on all web sites that publish clickstream data to the data warehouse, and is typically a slowly changing dimension. The data model shown has three roles for this dimension: Prior Page, Current Page, and Next Page – referring to the pages associated with the previous, current and next Page Visits of the session respectively. Having all three of these relationships available allows standard query tools to easily produce reports that show "where did they come from" and "where did they go next" for any web page. A special "Not Applicable" instance is used for the Previous Page and Next Page roles of the first and last page visits of a session respectively. Reporting page-to-page traffic this way is a whole lot easier than using correlated sub-queries.

The “Topographic Dimension” is used for web sites whose web content is structured – perhaps using page style “templates” consisting of several “containers” for dropping content into. It stores the template and container names and codes, along with any other descriptive information about the placement (screen geography) of the container on the page, and any special appearance characteristics of the container. Two distinct roles are represented in the model – one for the container that was clicked on the prior Page Visit in the session, and one for the container on the current page that will be clicked next. This data modeling technique simplifies finding “hot spots” of high value web page real estate, and in identifying the template and container presentation styles that successfully attract different types of visitors.

Complete Data Model

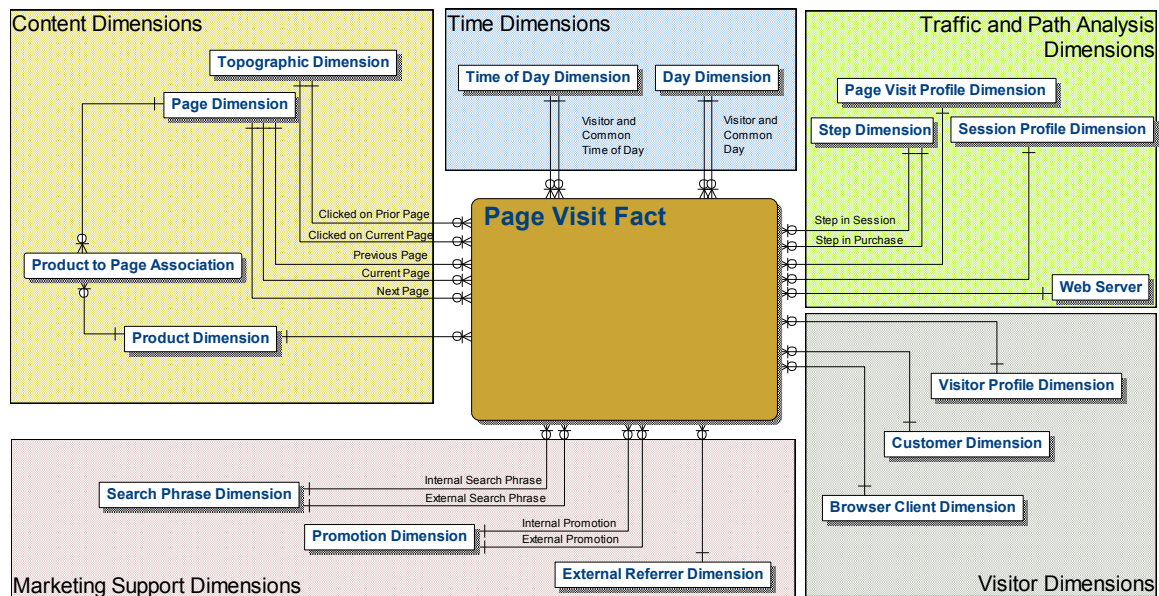


Figure 4 – Complete Clickstream Dimensional Data Model

The complete clickstream dimensional data model (figure 4) probably includes more dimensions than are needed in any one implementation, but they are shown nonetheless to convey the set of design possibilities available. Let’s take a quick tour around the model:

The “Day” (one row per day) and “Time of Day” (one row per second in a day) dimensions are standard Kimball modeling designs for capturing the date and time of an event down to the second, shown here with two roles –for capturing time from the Visitor’s and standard time perspectives. If needed, the web server’s time perspective can be captured too.

The “Session Profile” dimension, classifies the entire session into areas of interest to analysts, like: “Successful Sale”, “Just Browsing”, etc. It can also be used to describe the absence or presence of specific actions in the session – such as “Shopping Cart(s) Created”, “No Shopping Carts Created”, or “Shopping Cart(s) Abandoned”. Notice that by simultaneously constraining on both the Session Profile and Session Step Dimensions, users can walk backward and forward through many similar stories concurrently –

looking for interesting patterns – like “What are the top ten Pages visited in the waning Page Visits of Sessions that had Shopping Carts abandoned?”

Similarly, the “Page Visit Profile” dimension classifies each Page Visit into interesting buckets, such as: “Just Browsing”, “Shopping Cart Add”, “Shopping Cart Removal”, or “Page Failed to Load”, etc. The “Web Server” dimension captures specifics about the web server platform that served up the page, including its manufacturer and model, web server software type, release and version, operating system release and version, processor quantity and clock speeds, RAM and DASD capacities, etc. This information helps server farm technicians scrutinize the real-world performance of their web platforms – through maintenance cycles, patches, hardware upgrades, and other changes.

The model includes separate dimensions for capturing information about “Browser Clients” (web browser devices – typically PCs) and “Customers” (persons). A many-to-many relationship exists across these entities: A Customer uses one or more Browser Clients (PCs at home and work) and a Browser Client can be used by one or more Customers. The “Browser Client” dimension captures whatever information is available about the specific platform, such as its IP Address, Cookie GUID, and browser type and release/version. The “Customer” dimension carries all of the organization’s available information about the individuals that visit their web site(s). Both of these dimensions are typically “slowly changing”.

Because there are often millions of dynamic Browser Client and Customer profiles for a busy retailer, the “Visitor Profile” dimension is a mini-dimension that breaks out attributes that might otherwise make them “rapidly changing dimensions”. It classifies the Browser Client and the Customer at the point in time of the session, in terms of Customer Segmentation (i.e.: recency, frequency, and intensity), “familiarity” of the Browser Client and the Customer, and the level of Customer Recognition achieved in the session. “Familiarity” means simply: have we seen this Customer or Browser Client before? They are independently classified as either “New” or “Returning”. “Level of Recognition” is a classification of the confidence that we have in the true identity of the customer - either “Authenticated” (a positive identification occurred somewhere during the session), “Assumed” (no positive ID, the customer identity is assumed based on a UID found in cookie on the Browser Client), or “Unknown”.

The “Search Phrase Dimension” is shown with two roles: one for capturing the search phrase keywords typed into external search engines that drive traffic to the site, and one for capturing the search phrase keywords typed into internal web site searches. These dimensions capture the search keyword strings entered by users – verbatim. Although primitive, this technique should nonetheless be adequate for supporting simple keyword pattern matching and substring queries. Each unique set of keywords (search phrase) typed into an external search engine would result in a new row in the “Search Phrase” dimension, and would be associated with every Page Visit of the session. The search phrases typed into an internal search web page are associated with only that particular Page Visit. If your organization needs more robust keyword analytic capabilities, have a look at Ralph’s [“The Keyword Dimension”, 10/2000](#).

The “Promotion” dimension contains a row for each external and internal web promotion that drives traffic to, and within, the site. Every Page Visit in the session is associated with the External Promotion and the External Referrer that initially drove the visitor to

the site, but only the Page Visits that immediately follow clicks on internal promotional containers carry a defined "Internal Promotion".

The Product Dimension contains a row for each product available for purchase – and is typically a slowly changing dimension. Page Visits to single product web pages, such as product detail pages, carry the foreign key to the specific product that was showcased – otherwise they point to "Not Applicable" or "Multiple Products Displayed" product dimension records. In situations where products undergo predictable marketing cycles, an optional attribute in the product dimension called "Product Lifecycle Description" can be used to describe the stage of the product at the time of the Page Visit. Possible values are: "Controlled Introduction", or "Arrival of First Competitor", or "Market Saturation" – whatever is appropriate for your industry. It can either be folded into the Product dimension, or broken out as a separate mini-dimension for companies that have many products. This is another example of a "story-telling" technique, because it allows the analyst to re-align events by the "product story", and greatly simplifies comparisons of measurements taken across different products at analogous stages in their lifecycles.

The optional "Product to Page Association" table captures the many-to-many relationship between web pages and the products that they display. Including this outrigger table in queries explodes the granularity of the facts into "product impressions" – counting the number of times that products were displayed in any way on all web pages. The DW Architect should consider concealing this table from the end user community, including it only in pre-built views that rename all of the fact table metrics to "Impression" measures – in order to avoid possible misunderstandings.

Buy the ETL Team Lunch

Of course, all of this increased storytelling power comes with a price, and as usual, the data warehouse ETL team will be asked to step up and accept the additional work and complexity. In particular, these techniques demand that the ETL software hold in memory and profile a complete image of the individual "story" before its constituent event facts can be written out. In our clickstream example, this means that web events need to be consolidated, standardized in granularity, sessionized, and sorted by session prior to being handed off to a downstream layer that buffers each session, assigns surrogate keys to each Page Visit, and then writes the dimensional Page Visit facts out. Thankfully, there are some terrific clickstream data preparation tools on the market today that can handle some of the consolidation, standardization and sessionization tasks. Type "Clickstream Collection Technology" into Google to get a list of some of the players in this arena.