

Kimball Design Tip #1: Guidelines For An Expressive Clickstream Data Mart

By Ralph Kimball

The clickstream is the record of page events collected by a web server. In the raw data, there is one record for every click made by a visitor that the web server can detect. The clickstream contains unprecedented detail about every "gesture" made by a visitor to the web site.

The clickstream data source is huge. Even moderately busy commercial web servers may generate 100 million page event records each day. We must reduce the volume of data to manageable proportions for our most important analyses. In this design tip we will seek a way to NOT crawl through 100 million records, while still keeping a useful level of detail to analyze web visitor behavior.

In the raw clickstream data there are hints of a number of interesting dimensions, including Calendar Day, Time of Day, Visitor, Page Object Requested, Referring Context (what prior page contained the link clicked), and Action (basically either Get Object from the web server, or Post Object to the web server).

The recommended grain of the clickstream behavior fact table is

One Fact Record = One Visitor Session.

If an average session consists of 20 page events, then the number of fact records in our example is reduced to 5 million per day, which is comparable to the experience of medium sized retailer data warehouses.

The recommended dimensionality of this fact table is

- * Web Server Day (calendar date as recorded by the web server)
- * Web Server Time (seconds since midnight recorded by the web server marking the start of the session)
- * Visitor Day (calendar date as experienced by the visitor)
- * Visitor Time (seconds since midnight recorded by the visitor marking the start of the session)
- * Visitor (generic name "Visitor" for anonymous visitors, unique system generated name for unregistered visitors who have accepted a cookie, and true name for registered visitors)
- * Starting Page (identity of first page in session: the page that attracted the visitor from elsewhere on the web)
- * Ending Page (identity of the last page in session: maybe this is a session killer)
- * Referring Context (the URL of the page the visitor came from, if available)
- * Session Diagnosis (a simple descriptive tag indicating what kind of session this was)

The recommended numeric facts in this design are:

- * Number pages visited
- * Total dwell time (something of an estimate, because we can't account for the visitor's real activities)

This design can be a very powerful base from which to evaluate visitor behavior on a web site. The most important dimension is the Session Diagnosis dimension. You must have a sophisticated back

room ETL (Extract-Transform-Load) process to create good session diagnoses out of the detailed page event sequences.

For further reading on these subjects, download the following free article from the Intelligent Enterprise magazine archive: www.intelligententerprise.com/990501/warehouse.shtml

Another article will appear in January 2000 in Intelligent Enterprise discussing the Page and Session Diagnosis dimensions in more detail.

Follow-up to Design Tip #1: Guidelines For An Expressive Clickstream Data Mart

January 7, 2000

I had a number of interesting comments from the first design tip, which recommended a dimensional design for clickstream data. Several people asked me why the design tip recommended a fact table grain of a record = a complete session, when the January 5, 1999 Intelligent Enterprise article found at www.intelligententerprise.com/990501/warehouse.shtml recommended a grain of the individual page event. These people asked me if I had changed my mind.

No I have not changed my mind, but I understand the problem better. There are at least three useful grains at which to represent clickstream data:

- 1) Fact record = individual page event. This level, described in the IE article, can give detailed maps and trajectories of every web visit, if you keep every record. But for very busy sites, there is too much data. You will spend all your time and money collecting and storing the data rather than analyzing it. Several people told me that statistical sampling techniques with as little as 1% of the total volume of data could very usefully depict site usage patterns that would lead to important decisions about the use of the web site, even if all their individual visitors were not present in the data. I like this suggestion very much. You will probably need a professional statistician to help you choose a robust small sample of your data.
- 2) Fact record = each complete visitor session. This is the level I discussed in the first design tip. In this case, you can realistically strive for complete coverage of all visitors, although you are not seeing their complete maps and trajectories taken through your web site. But you can do extensive demographic and web site effectiveness analyses. Remember that you have the Entry Page, the Exit Page, and the Session Diagnosis dimensions.
- 3) Fact record = web page by calendar day. This grain is one of several similar rollup levels that can be useful for seeing the total pattern of hits in various parts of your web site. Clearly the advantage of this grain is the sharply reduced size of the data, but like any aggregate fact table, you have suppressed several of the behavioral dimensions like Visitor and Session Diagnosis.

I guess the answer to the grain question is that eventually you want them all. Just like most of the other data warehouses we build.