

InfoWorld article – The marvels of an event-based schema

The issue of schema—and what is frequently perceived as its inherent difficulties—is becoming more important every day. Organizations are increasingly encountering decentralized computing environments typified by semi-structured or unstructured external data of varying formats, often requiring integration with internal, structured data for immediate business value.

Storing each type of origin source or application data in its own database is becoming less practical because of the need to cross-reference or aggregate data horizontally. Calibrating and recalibrating relational database schema for such big data demands is far too dilatory, as are traditional master data management approaches with their modeling and schema complexities.

The rapid storing and querying of diverse data at scale necessitates a simplified, all-inclusive schema without losing any information (which often occurs in schema simplification attempts). By turning every facet of a particular line of business into an event, organizations can rapidly utilize this approach for a flexible, comprehensive schema throughout the enterprise.

[The essentials from InfoWorld: NoSQL grudge match: MongoDB vs. Couchbase Server • Review: MongoDB learns cool new tricks • The essential guide to MongoDB security. • How to work with MongoDB in .Net. | Go deep into analytics and big data with the InfoWorld Big Data and Analytics Report newsletter.]

The results are a swiftly queryable homogenous schema that reduces back-end complexity for the timely incorporation and

integration of the new, diverse data sources typifying the big data era.

Everything is an event

The ease of the event-based schema approach lies in the fact that any data-related occurrence for a line of business can be transformed into an event. The list is innumerable: customer complaints, product feedback, transactional data, email, or mobile communication, changes in health and more. Users simply have to specify their data sources and categorize all incoming data in terms of events. No matter what the data are, they are easily categorized according to event type, actors, start and end time, location, and any other factors relative to the data.

Thus, if an event type is a phone call, the actors are the participants on the phone, the location is where the call was accepted or originated (or both, the choice is entirely up to the organization), and all the other appropriate metadata needed to adequately describe the event. The same approach is used for transactional data, analytics outputs, inputs, recommendations, or anything else. The simplicity of the overall schema is that whatever the events are, they are all *described the same way* according to type, actors, time, and all the other factors organizations want. The uniformity of this approach encompasses all data, and is only possible today at scale because of contemporary big data advances in distributed databases and parallel processing.

The value of taxonomies

One of the primary enablers of this newfound approach to schema is the taxonomies which underpin this methodology. Since events can quite literally be anything, they need taxonomies to describe them in a standardized manner. The best

taxonomies do so across the enterprise and have a hierarchy of terminology for individual business units too. Event types (which provide the greatest amount of flexibility in this schema) must be specified by taxonomies, which in turn enable the all-inclusive nature of this schema.

Taxonomies are also able to be queried. It is therefore necessary to have every important business concept input in the underlying repository as a taxonomy to provide the requisite detail for events and the schema they facilitate. The tandem of taxonomies and the event-based schema approach drastically decreases the multiple ways of referring to the same concepts across different databases, which plagues the majority of organizations that have different definitions and meanings for common terms such a customer.

Without data loss

It's important to note the capacity for individuality and specifics facilitated by this homogenous schema approach. Each event may have different attributes which are not applicable to others. In order to denote that degree of specificity, yet still preserve the simplicity of this simplified schema, organizations can define different key-value pairs correlating to those attributes for each type of event. This aspect of the approach ensures that all data attributes are included in the schema.

Organizations can use those key-value pairs to denote certain aspects of metadata that are not expressly identified by the events themselves. Or, they can simply use them for the particulars of certain circumstances within an event. Either way, they still get a simplified schema that is highly descriptive for any data type, format, or structure.

Flexible ease

The flexibility and ease of implementation of the event-based approach are its most cogent boons. Unlike conventional schema in relational databases, users do not have to think of all query possibilities in advance to render the schema according to them. They simply have to outline the various information their data—as events—can include, which produces the additional benefit of making their data uniform while decreasing their schema's complexity. When working with various data types at the speed of big data, this method is ideal for integrating and aggregating assorted information for the holistic value it provides.