EXCERPT

Warehouse Design for the Future

The star-schema dimensional architecture is typically adopted for the flexibility it offers to support future data context, structure, and organizational changes with a minimum of impact or burden. Toward that end, any clinical data warehouse has to be designed with an eye toward future requirements that might impact it; and emphasis must be been placed on enabling such future changes without introducing extra risk into the development of any initial version. Although not actually implemented, and not even fully designed at the early stages, the following future scenarios should be considered in the design of a clinical data warehouse in an effort to improve the initial design while enabling future expansion.

Telemetry Data

With increased automation of medical procedures and treatments, a consequent increase in data volumes can be expected, and this is already being seen in the capture, storage, and analysis of medical telemetry data. There are a number of ways in which a warehouse design can handle the increased volumes associated with telemetry data.

One option would be to simply store all telemetry data received at the warehouse ingest process, basically necessitating an increased grain on the Clock dimension to handle the grain required by inbound telemetry data (typically down to seconds or fractions of seconds). This increased grain is why the Clock dimension design typically includes a sub-dimension of seconds, but that the rows required for Seconds (or sub-seconds) would be added to the dimension opportunistically; or only as data points arrive. In this manner, detailed second or sub-second rows in the dimension would only exist where there was data recorded, preventing an unnecessary explosion of row counts in the Clock dimension to cover times of day with no likely data. In the extreme, though, adding all rows to the dimension to account for telemetry data measured in tenths of seconds would multiply the row count (initially 1,440) in the dimension by 600. That volume is manageable. Such expected future volume is why the Calendar and Clock dimensions are not designed as a single dimension: because the 864,000 rows would eventually be needed for every day in the Calendar if Clock were merged in.

Another option for telemetry data would be to not store all of the raw data, but to aggregate it during the ingest process. This option is the most likely simply because few users typically express interest in historical raw telemetry data. For this option, the warehouse ETL process would pre-process the telemetry data according to pre-defined algorithms that involve defining time periods and desired data points. For example, Mount Sinai Hospital in New York includes surgical anesthesia telemetry data in its clinical data warehouse by capturing median values every 10 minutes. Generalized routines to capture descriptive data (e.g. min, max, mean, median, std. dev.) over parameterized time intervals (e.g. 1, 5, 10, 15, 30, 60 minutes) would not be difficult to introduce, and would support the requirements of telemetry being collected under a variety of acuities (e.g., surgery, ICU, telemetry bed, and telecare).

Specimen Data

The trend toward specimen instantiation was introduced above during the discussion of the design of the Patient dimension, and is considered a short-term need within medical informatics. Storing such specimen is consistent with the current design of the CDM without changes to the database design. The need to keep track of individual pathologies, clinical specimens, transplanting organs, and fetus *in utero* will add detail to the CDM at the specimen level in future years.

More strategically, any clinical warehouse is likely to be challenged to store even more specimen-level data in the future, sometimes not even associated with specific patients. With the growth of bio-banking and gene-banking requirements, a clinical data warehouse is increasing likely in the future to need to capture and store clinical results against, and physical placement of, significant numbers of specimen. Indeed, the number of specimen is likely to eclipse the number of patients in the typical warehouse in coming years.

Cohort Data

While a Cohort sub-dimension has been designed into the Patient dimension, no functionality has been designed around it. As the use of the warehouse matures in the future to move beyond tactical applications (i.e. queries by Patient or Provider) to broader studies involving disease management or population ecology, the data desired from the warehouse will increasingly be associated with cohorts of patients rather than with patients themselves.

For example, Mount Sinai Hospital in New York is using their clinical data warehouse to track a cohort of over 10,000 patients who lived near the World Trade Center on September 11, 2001 (and who have largely scattered around the United States over the intervening years). Another cohort includes the 5,800 patients who have consented to participate in their nascent bio-banking initiative. By defining a cohort as a distinct dimensional entry in the warehouse, facts can be stored and queried about the cohort using all of the tools typically used to query data at the patient level.

Cohort data, when implemented, will take advantage of many of the data aggregation features described above for telemetry data; except that where telemetry data will be aggregating time, cohort data will be aggregating patients. For example, facts stored for a cohort might include the median value for a lab result across the cohort by month and region.

Storing and using cohorts in the warehouse will not require any design changes to the database; only design and implementation of the needed data aggregation tools, or additional sources of cohort-level data. In addition, the Research Study sub-dimension of Encounter typically needs to be fully implemented in order to provide further context for cohort-level data; but this sub-dimension has already been considered for the Encounter dimension as well.

Shared Servicing

Certain markets within information technology are seeing shifts toward shared services, and it is foreseeable that the warehouse could eventually be used in such a manner to provide shared informatics services beyond the current inhouse-based scope. To the extent that the extended users of the warehouse would require similar clinical and medical informatics, virtually no changes would be required in the design of the warehouse; although each new user of the environment would likely need to provide an entirely new and different ETL subsystem. These partners would require a load of large numbers of new organizations, facilities, caregivers, and accounts; but would not require structural changes to the database itself. The design of the Organization dimension to include a comprehensive natural hierarchy is intended to facilitate the extra layer of logical security that such an expansion of users would necessitate.

Network Care

An expansion of a clinical warehouse to include care provided outside of existing environment of in-house facilities is foreseeable in the future. Such network care data would take advantage of all of the dimensionality of the warehouse, but like with shared servicing, it would require the addition of numerous organizations, caregivers, facilities, and accounts to the dimensions.

The expected biggest shift in the addition of network care data would be the need to store and manage any contractual and financial data that might be received along with the new clinical data. If such data were deemed beyond the scope, then almost no changes would be required in the typical clinical warehouse. But if contractual and financial data associated with network care were required, it could be problematic.

Contractual data might include agreements with network providers to provide certain services at certain facilities at certain rates to certain cohorts of patients. Financial data could include billings, payments, debit/credit management, and patient reimbursements. Most of this data could be handled through expansion of the organization and accounting dimensions, and an increase in inter-dimensional control facts among the key dimensions. But it is also foreseeable that such an arrangement could stretch the warehouse dimensional model too far, and eventually necessitate the addition of a new dimension to handle the contractual/payer perspective.

Ontological Support

The healthcare information technology field is increasingly adopting and adapting standardized vocabularies, taxonomies, and ontologies for the identification and cross-referencing of data from disparate applications that might not themselves use shared vocabularies. Such ontologies can be made available to the warehouse through the loading of a variety of public domain clinical ontologies available over the internet from the biomedical community.

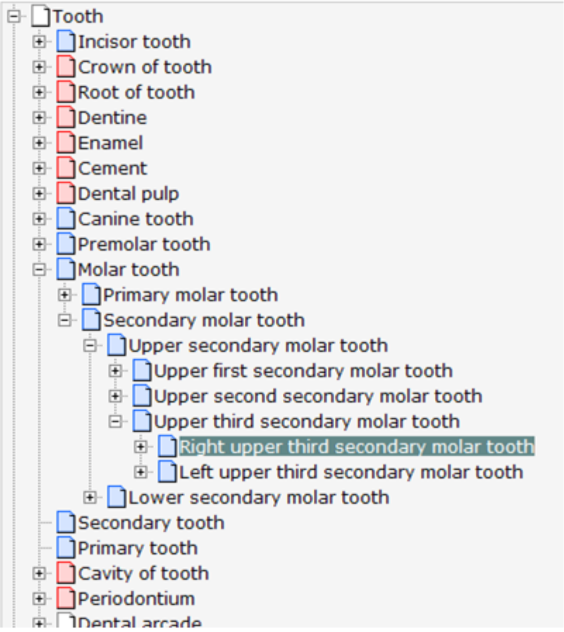
Loading more extensive ontologies into the warehouse would provide for a richer more extensive ability to analyze the data within and across vocabulary domains, and to correlate data using relationships in the ontologies that are unknown or unavailable within the warehouse's clinical data sources. The challenge in using ontologies within a clinical warehouse isn't loading them; the design of each dimension can easily accommodate the addition of any number of ontological sub-dimensions. The challenge is in developing the necessary tools to take advantage of those ontologies, primarily in cross-referencing existing warehouse dimensional concepts into the ontological terms. Such analysis would need to be conducted largely on a case by case basis, but doesn't present any major challenges to the actual warehouse design.

Anatomy Example

While there are many ontologies available within the healthcare sphere, an anatomy ontology presents an excellent example precisely because a typical clinical warehouse design based on in-house clinical systems typically lacks the kinds of data cross-referencing and integration that bringing in an ontology would address, and so serves as a good early candidate for ontology adoption.

The ontology of interest is the Foundational Model of Anatomy (FMA) ontology. FMA is an evolving computer-based knowledge source for biomedical informatics concerned with the representation of classes or types and relationships necessary for the symbolic representation of the phenotypic structure of the human body in a form that is understandable to humans and is also navigable and interpretable by machine-based systems like a data warehouse. Specifically, the FMA is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy. Its ontological framework can be applied and extended to all other species.

In this example scenario, the elements of the FMA would be loaded directly into a warehouse's Anatomy dimension (or into the Annotation dimension if no Anatomy dimension has yet been implemented in the design). Existing warehouse anatomy references would be converted, shifting from being specific rows in the dimension to being new context references to their associated FMA entries. For example, instead of tooth "31" being a row in the dimension (D_ANATOMY), its existing context reference in R_ANATOMY would be shifted to point to the appropriate tooth entry in the FMA (refer to Figure 1 below).



- Tooth
  - Incisor tooth
  - Crown of tooth
  - Root of tooth
  - Dentine
  - Enamel
  - Cement
  - Dental pulp
  - Canine tooth
  - Premolar tooth
  - Molar tooth
    - Primary molar tooth
    - Secondary molar tooth
      - Upper secondary molar tooth
        - Upper first secondary molar tooth
        - Upper second secondary molar tooth
        - Upper third secondary molar tooth
          - Right upper third secondary molar tooth
          - Left upper third secondary molar tooth
      - Lower secondary molar tooth
  - Secondary tooth
  - Primary tooth
  - Cavity of tooth
  - Periodontium
  - Dental arcade

Figure 1. Example FMA Entry for Individual Tooth

Each tooth would be mapped against an FMA anatomical entry for that tooth, such as Right Upper Third Secondary Molar (FMA:55696). The source data includes definitions for all 52 human teeth (20 primary, 32 permanent). Because such descriptions are not available within the many source data feeds, such annotation would support more extensive analysis by users potentially familiar with the more extensive anatomical terminology.

Additional anatomical mappings can be obtained by parsing the descriptive fields in the diagnosis, procedure, and material dimensions for their anatomical components. In this manner, an x-ray of the left knee in the Procedure dimension becomes correctly mapped to the left knee in the Anatomy dimension. A myocardial infarction in the Diagnosis dimension gets correctly mapped to the affected heart anatomy.

The FMA ontology also contains fiat hierarchy descriptions unavailable in those other dimensions (e.g., it knows that a hand [FMA:9712] is part of an arm [FMA:24890], and a foot [FMA:9664] is part of a leg [FMA:24979], also that a hand [FMA:9712] could be a right hand [FMA:9713] or a left hand [FMA:9714]), empowering simple anatomical queries that would be extremely difficult to accomplish using the other dimensions and fields. Simple aggregates like Set of Teeth [FMA:75150] can be used to obtain all teeth in a single simple query. The ontology hierarchies also include the relationship types that are needed to populate the Perspective in the H_ANATOMY table, including is_a, part_of, and union_of (e.g., Left Hand is_a Hand, Hand is_part_of Arm).

Generally, all of the current anatomy information in a non-ontology warehouse design becomes context references into the richer ontology-based dimensional data. The dimension will end up with zero, one, or two context reference entries pointing at each anatomical entity in the dimension. Each tooth will end up with multiple context references pointing at it; one being from within the ontology itself, and others coming from available non-ontology source data. Some of these will be hierarchic parents of used anatomical entities that can be used in queries (e.g., Mandible [FMA:52748] to find all lower teeth).

Natural Language Annotation

A clinical warehouse design allows for the capturing of large amounts of data that will be dimensionalized according to the nature of the data received with each data point during the ingest process. The dimensionalization of the data provides multiple search and access paths for that data to warehouse users. However, a huge amount of additional classifying data will remain hidden within the text portions of the facts stored in the warehouse.

The dimensional warehouse design anticipates a desire to be able to continuously search and annotate such data using a variety of natural language algorithms that will be developed and mature over the coming years. The presence of the Annotation dimension in the general warehouse design is specifically aimed at supporting this future requirement since not all annotation opportunities will involve data associated with an existing dimension in the warehouse. In conjunction with the loading of various biomedical ontologies into the warehouse, an annotation function would be very powerful.

Simple annotation involves continuously scanning the facts in the warehouse (particularly focusing on large text facts like pathology and radiology reports, physician summaries, and nursing notes) looking for keywords against which the facts can be further annotated. The simple search looks for phrases that have been defined in any one of the warehouse dimensions. For example, if the Disease Ontology were loaded into the Diagnosis dimension (or into the Annotation dimension if not otherwise wanted in the Diagnosis dimension) a simple annotation tool would look through the fact tables looking for entries in the ontology. For facts with terms identified, the Disease Ontology entry would be added to the dimensionalization of that fact in the Diagnosis (or Annotation) dimension. Such annotation is particularly powerful and useful for keywords that are difficult to discern otherwise. For example, Niemann-Pick Disease has no specific ICD-9 diagnosis code, and so the diagnosis is often hidden in data warehouses because the disease is typically only found as text within a pathology report. An annotation against this term using the Disease Ontology would aid researchers looking for this disease within the general clinical warehouse.

As the number of available ontologies loaded into a warehouse increases, the richer and richer the terrain for annotation of facts becomes. It becomes reasonable to expect to more effectively find and use data embedded in the large text blocks ingested from the clinical setting, so that if physicians, surgeons, anesthesiologists, pathologists, radiologists, cytologists, or nurses make mention of the disease, pathology, and anatomy terms found in the ontologies; those facts will be effectively cataloged by the warehouse annotation tools.

Bioinformatics Data

The trend toward integrating bioinformatics data into clinical data repositories is being driven largely by the push toward translational medicine; the tying of data developed benchside to the care of patients at bedside. This creates a need for translation bioinformatics with strong integration of clinical data. If and when such requirements begin to filter through to the clinical warehouse, it will place stress on several of the warehouse dimensions and fact tables in predictable ways; some of which are likely to require the addition of added dimensionality to the clinical warehouse.

There are many existing features of the clinical warehouse design discussed above that would support the introduction of bioinformatics data: a) Cohort data to define study groups, b) Specimen data to track individual experiments, c) Research Study encounters to group experiments, d) Telemetry data to handle the storage or aggregation of large data volumes, e) Ontologies to be able to pull in the various bioinformatics perspectives, and f) Annotation to support cross-study citations and researcher notes.

Most importantly, the general warehouse design supports the desired level of research and clinical data integration. For example, since the warehouse typically carries the clinical record of patients, then tissue specimen drawn from some of those patients that were used in a microarray study to draw conclusions about gene expression under various experimental conditions would automatically be able to correlate microarray observations with the clinical history of the tissue donors. It is precisely this lack of integration that is currently holding back much bioinformatics research today.

The introduction of bioinformatics data into the warehouse could begin using existing dimensions in the design, but would quickly overwhelm those dimensions; necessitating the addition of new dimensionality to support the grain of bioinformatics research. In particular, the warehouse design lacks the System and Biology dimensions in which data would be dimensionalized into the *science* of medicine. The System dimension would likely include subdimensions for Ecologies, Populations, Organisms, Organs, Tissues, Cells, Proteins, and Genes. The Biology dimension would likely include sub-dimensions for Genome, Regulome, Transcriptome, Proteome, Epigenome, and Metabalome.

Bioinformatics data is likely to need other new dimensions, among them several that could probably be mapped as subdimensions into typically existing clinical dimensions for the near-term. These would likely include Phenotype to track diseases, appearances, and traits; Process to track physiology and reactions to events and treatments; Development to track life and growth stages; Pathology to track resistances and perturbations; Environment to track nature and settings; and Chemistry to track pharmacology, elements, molecules, and macromolecules.

Translational bioinformatics is too young a discipline to be able to be sure of the exact dimensional impacts such data would have on a typical clinical warehouse, but it clearly represents the more extreme view of what could drive change in a clinical warehouse in the future. It is encouraging to note that beyond the addition of more dimensionality to the design, few other changes to the underlying clinical warehouse architecture and design would be foreseen.

Not surprisingly, the new dimensionality required to support translational bioinformatics would directly and immediately support the clinical impact of the warehouse. The new dimensionality required for research data tends to represent the very things that clinical warehouse researchers are looking for as well. Without the bioinformatics dimensionality in the warehouse, much of that research perspective remains in the mind of the user, not in the data content of the clinical warehouse.

The anticipated annotation features for the general clinical warehouse would find a rich target environment within the masses of clinical data for annotating pathology observations and disease observations against clinical patient results and diagnoses. Correlations of clinical data among anatomy, process, and development could aid in the analysis of clinical cases and outcomes. That's the whole point of the translational bioinformatics movement! It seems that the general clinical data warehouse design described above is well position to provide such support in the future.