Conceptual Model for Research-Driven Data Marts

Shelley A. Rusincovitch, Howard C. Shang
Duke Health Technology Solutions, Duke University Health System

Introduction
The enterprise data warehouse (EDW) is an important asset in healthcare and supports many secondary use activities, including clinical research. Data warehousing is a proven and valued tool that leverages established methodologies, experience with large volumes of data, and existing tools and platforms. However, EDW development, maintenance, and integration of new data sources can be time-consuming and costly endeavors. Researchers frequently utilize project-specific data sources and definitions. Incorporating these domains directly within an EDW would result in greater operational efficiency, but would traditionally require the investment of significant resources in order to reach full integration of all new data sources. In addition, many research data sources cannot be integrated at an enterprise level due to data use agreements, confidentiality considerations, and project-specific authorizations.

Methods
We present a conceptual model with specific application for research. Our model is centered on the development of project-specific data marts using a consistent set of practices. Information is connected to consumers via a suite of stackable informatics services and platform functionalities applied to the data marts. Integrated security is a foundation for all systems and processes.

Table 1. Processes and components of the project-specific data mart.

<table>
<thead>
<tr>
<th>Process/Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Sources</td>
<td>Both internal and external data sources can be utilized independently for any given project. Every data source may have distinct requirements and can be constrained in usage.</td>
</tr>
<tr>
<td>Project Authorizations</td>
<td>Contracting, authorizations, and data use agreements are dependent on the context of each data source.</td>
</tr>
<tr>
<td>Source System Analysis</td>
<td>A methodical analysis of data and systems architecture is applied for each individual data source with best practices of profiling and metadata development.</td>
</tr>
<tr>
<td>ETL Processes</td>
<td>ETL (Extract-Transform-Load) processes programmatically mediate data transfers, data transformations, identity matching, address standardization and geocoding, and unstructured data. A future state is intended to incorporate standardized terminologies/ontologies.</td>
</tr>
<tr>
<td>Project-Specific Data Mart</td>
<td>The data mart emphasizes best practices in clinical domain structuring. It is deliberately intended to be system agnostic to take advantage of rapidly-evolving platforms, and allows more agile adoption on an individual project basis than a more centralized system would require. A metadata layer is incorporated into the data mart to enable source data to be consistently translated into project-specific data element definitions.</td>
</tr>
<tr>
<td>Informatics Services and Functionality</td>
<td>A catalog of services is incorporated to meet each project’s objectives and scope. Existing platforms such as DEDUCE and DISCERN provide a robust framework of functionality that can be deployed as appropriate.</td>
</tr>
<tr>
<td>Consumers</td>
<td>The model is intended to meet the needs of project consumers with different roles and responsibilities. The project retains control of decisions about access and disclosure.</td>
</tr>
<tr>
<td>Integrated Security</td>
<td>Security and patient confidentiality are an integral part of all systems and maintained through complimentary mechanisms and policies.</td>
</tr>
</tbody>
</table>

Figure 1. An instance of the project-specific data mart model.

Discussion
The existing Duke Medicine EDW is a unique and widely utilized resource. Our team has previously developed and implemented innovative tools, platforms, and services to facilitate use of the EDW by our clinical, operational, and research communities.

In this expansion of existing infrastructure, each data mart remains distinct from the EDW. Linkages and associations within the EDW, and to other data marts, can be enabled as appropriate to project needs. Data marts are separate from the EDW presentation layer because integration of research data must be sensitive to data use agreements and patient consent. This organization also allows more flexibility and consideration to project requirements. We expect that this model will be both cost- and resource-efficient.

References

Acknowledgements
Robert M. Califf, MD†
Bimal R. Shah, MD, MBA†
Monica M. Horvath, PhD‡
and the DHTS Information Management Team

Contact Information
Shelley Rusincovitch, Senior Analyst§ shelley.rusincovitch@duke.edu 919-668-7954
Howard Shang, Senior Director¶ howard.shang@duke.edu 919-861-4481