

Summary of the PDW to DW Transition

The Pilot Data Warehouse

The Data Quality Action Team and the Division of Information Resources (DIR) initiated the Pilot Data Warehouse (PDW) Project to test assumptions and gain knowledge about issues that needed to be better understood and addressed in order for IHS to implement an enterprise-wide, national data warehouse environment for IHS, Tribal, and Urban healthcare systems.

As part of the PDW Project, the team designed, tested, and implemented processes to load data from export files into the PDW database. The PDW was specifically designed to gather and store data using minimal cleansing or transformation techniques. Some of the conclusions of the testing included:

- Updating of RPMS registration records is an important issue to be dealt with during the design phase of DW1.
- An algorithm must be developed to identify duplicate occurrences of the same encounter transmitted from multiple sources.

The Final Report was published in October 2002 and is available on the web at http://www.ihs.gov/CIO/DataQuality/warehouse/PDW_Final.pdf.

Major Interval Enhancements

In preparation for the first iteration of the production, enterprise-wide implementation of the data warehouse, a number of changes were implemented (or are currently being worked on) either because they were required, to better comply with DIR's overall IT architecture, or because they were felt to be advantageous in the long run and appropriate to do at this point in the project. These changes include:

New Export Process

The PDW tested exports were processed using current procedures, (i.e., an ASCII file is transmitted to the Area office, run through BXP, and transmitted to NPIRS/ORYX). For DW-1 we have programmed a completely new, and as yet untested, export process for RPMS. As dictated by DIR's architecture they will be:

- Unified registration and encounter record export
- HL7 formatted messages from the sites
- Transmitted to the Integration Engine where they will be translated into ASCII files and transmitted to the DW.

Normalization of the DW Model

Taking advantage of the conversion to HL7 messages, we chose to normalize the DW database structures so that we will now be able to accept all data entered at local sites for many types of data (diagnosis, ADA codes, immunizations, patient education, etc.) rather than being limited to an arbitrary, set number defined by the fixed length of the export record itself. We have similarly modified the post Interface engine ASCII record layouts and ETL and table structures in the DW database to accommodate this.

New ETL Processes

In addition to the changes made to accommodate the increased normalization, additional and extensive changes were made in the staging tables and their structures to allow the ETL process to be more efficient and logical. These structural changes, in turn, required programming new ETL processes for these tables. In

addition we specifically developed new logic for updating registration records and added it to the database model.

New Look Up Tables

In moving from PDW to DW, we also researched, designed, setup, and then populated over 100 look up tables (tables that define the standard domain values for a specific field such as patient education codes, provider discipline and affiliation, clinic type, service type, etc.). We used this as an opportunity to work closely with the RPMS DBA and appropriate users to improve the coordination of the "Standard Code Book" process and assure that there is better synchronization of these specific code sets between all users.

IHS HL7 Implementation Guide

Concurrent with its development of this entirely new export process, the DW team produced a document that summarizes the HL7 message system and details its specific implementation for the data warehouse (e.g., which message segments comprise the IHS-specific data messages to DW-1; the order, content, and format of individual data components; etc.). By doing so, we hope to provide a basis for standardizing both the syntax and the semantics of the data provided by all the various sites, non-RPMS as well as RPMS, thus improving the quality, understandability, and completeness of all derived information.

Unduplication of Encounter records

The PDW team did not address the unduplication of encounter records from disparate sources, but deferred this issue to DW-1. This includes the troublesome issue of CHS data. We are addressing this issue with users as we prepare for the DW-1 load.

Technical Review of Original Design

A team of 20 technical resources, including developers, analysts, system administrators and database administrators, carefully reviewed IBM's proposed DW-1 database model and ETL process design. Not all members participated in every review, but each section was evaluated by a minimum of 4 people.

In this review we identified a few key issues, summarized below, that the group has or is planning to address prior to loading data into the DW-1.

Critical Issues

These issues would have prevented our successfully loading the data.

Patient Registration Load

As originally designed, the logic for updating RPMS Patient Registration information relied upon the source system providing record level information that, on further review, we realized it would be very difficult to provide. In addition, the original logic relied upon the source system to provide a record identifier that the warehouse would use to locate the exact record(s) to update. However, the identifiers vary by data type, are not consistently maintained across modules, and therefore cannot be utilized by the warehouse.

We therefore designed new logic that will allow us to accurately update registration data in the warehouse. However, this logic still needs to be coded and tested.

Staging Table Load

As originally designed, export files were to be loaded into the warehouse staging tables using an IBM utility called Autoloader. Autoloader would have allowed us to easily load from ASCII to DB2 without having to write any programs or scripts. But we have not been able to get this utility to run successfully.

Therefore our DBAs have implemented an automated way for us to load data into the staging tables without using Autoloader. This has already been unit tested.

Outstanding Issues

These are issues that we have not yet solved. Although they will not prevent us from loading data into the warehouse they could cause serious problems with data quality if not rectified prior to the load.

Nationally Unique Identifiers

The RPMS developers have implemented a method to assign a nationally unique identifier to each registration and visit record. The warehouse will rely on these identifiers to update information and to accurately link visit records to patient records (critical for User Population, GPRA, ORYX, and other reporting).

We implemented the unique registration identifier in NPIRS during the Comprehensive Patient Registration Reload over a year ago. We recently discovered infrequent and sporadic, but widespread instances of duplication – i.e., this nationally unique identifier is not always unique even at the facility level.

The Software Development Team is currently investigating this.

General Issues

The review team identified several improvements that are easier to implement now than they would have been in future versions of DW. We have already implemented these changes or will do so prior to the DW-1 load.

- Pare down the number of staging tables, which also reduces the number of ETL processes that need to be developed and maintained.
- Implement an orientation engine that will identify and process non-standard formats that are transmitted to the warehouse without having to create new staging tables or ETL processes, which the current environment would require.
- Separate history from current data instead of attempting to manage “state” tables that rely on flags to identify current records. This will be easier to manage and should significantly improve load program performance.
- Consider using Embarcadero to replace stored procedures. This should also improve performance, although we won't know until we do some benchmarking.

Issues That Still Put The Load At Risk

Critical Issues

- Resolution of the Unique Registration Record ID problem
- Adequate staffing and hardware for the Interface Engine
- Installation of additional disc space
- Successful testing of the RPMS export process, from the program creating the RPMS export through the interface engine processing, ETL processes into staging tables, and finally the ETL processes into the target tables.
- Successful testing of the new and modified staging and target tables and associated ETL processes.

Other Important Issues

- CHS data handling issues
- Unduplication of encounter records from disparate sources
- Processing of Non-RPMS site exports