
National Patient Information Reporting System: National Data Warehouse

NDW General Data Mart Technical Guide

Current Version:7.0

Created: 9/16/2011 1:39:00 PM

Published:2/7/2012

Author: NPIRS-NDW



Department of Health and
Human Services

Indian Health Service

Office of Information
Technology (OIT)

Contents

Version Control	iii
Overview	1
Design Parameters	1
System Environment	2
Architecture	3
General Data Mart Design	3
General Data Mart Schemas and Tables	4
Legacy Schema	5
End-User Access	5
Area Access	5
Partially Cleansed Data.....	5
Security Access Levels	8
Federation	9
Extract, Transform and Load (ETL)	9
Extract	9
Transform.....	9
Load	10
Baseline Loads	10
Incremental Loads	10
Backups	10
Appendix A: REG_NP Table Schema	11
Appendix B: Split Tables	12
Appendix C: Label-Based Access Control	14
LBAC Implementation	14

Version Control

Version	Date	Notes
4.0	June 2009	<p>Supersede the General Data Mart information from the <i>HOLLYWD Database Technical Guide V3.0</i>. Add:</p> <ul style="list-style-type: none"> • schemas for the Restricted Personally Identifiable Information (PII) Data Requirements • data element that allows control of Area security • schemas for each version of past userpop reports run from NDW data • new tables; pat_reg_ssn and enctrss_ssn • Architecture Section • Backup Section • ETL Section <p>Remove:</p> <ul style="list-style-type: none"> • Social Security Numbers from ENCTRSS and PAT_REG tables <p>COTR approval June 16, 2009</p>
4.1	January 2011	Annual Update; reorganized document structure, removed all explicit version number references to other documents; removed/combined several sections, changed name of Appendix A
5.0	April 2011	Final
6.0	January 2012	Uploaded to New Sharepoint; Version Number automatically incremented.
7.0	February 2012	Updated metadata with missing Target Folder and Target Path information.

Overview

The General Data Mart (GDM) contains a query optimized sub-set of the Encounter and Registration data from the National Data Warehouse (NDW). Comprised of a set of tables, views, and Materialized Query Tables (MQTs), this data mart provides access to the following data:

- All current registration data contained in the data warehouse which **includes** Personally Identifiable Information (PII)
- Non-PII current registration data, which **excludes** Personally Identifiable Information (PII)
- Historic and current user population tables
- All current encounters contained in the data warehouse
- Subset of encounter history data
- All reference tables
- Basic Meta Data, which contains information about tables and columns included in the database
- A snapshot of historic Legacy data tables. Data in tables associated with the Legacy Data Mart is not refreshed.
- Administrative data, including a timestamp for Mart refreshes and also information pertaining to file exports

Design Parameters

- The General Data Mart exists on a separate server than the NDW production database. This server is currently available and utilized for the GDM with no additional hardware required at this time.
- The HOLLYWD database houses the General Data Mart and is used as the source for establishing connections.
- All data within the General Data Mart is refreshed as stipulated in the *General Data Mart Service Level Agreement*.

- The data in the General Data Mart is a copy of NDW production data. Access to an entire table within the mart, as well as individual rows within the tables, is determined by the level of a user's security permissions. The ENCTR and REG schema tables do not use any scrambling, data validation, encryption, or other methodologies to disguise Personally Identifiable Information (PII). The REG_NP schema tables have been partially cleansed of PII data utilizing risk vs. access parameterized and approved by IHS.
- Access controls are administered that allow users to query data that is appropriate to their authorized level of access.
- Enhanced security controls which adhere to IHS standards, as outlined in this document and in separate security documents, are enforced.
- User access is maximized by allowing multiple simultaneous user query access while complying with stringent security restrictions.
- The General Data Mart is enterprise compliant to allow various environments to access the database, including ODBC, JDBC, OLE, and CLI. These are some of the environments and protocols that may be used to access this data mart, depending on user needs, access protocols, and environments.
- System resources are monitored online automatically using Query Patroller and other tools. If a query exceeds a reasonable threshold, it can be placed on hold and restarted later.

System Environment

The following sections describe the physical environment of the NDW General Data Mart:

Server:	BILBO, 64bit
Database:	HOLLYWD, DB2 Version 9.5x
AIX Version:	6.1 or above
FTP Address:	198.45.1.38
System Access:	Enterprise compliant to allow various environments to access the database, including ODBC, JDBC, OLE, CLI.
System Monitoring Tool(s):	IBM DB2 Query Patroller

Architecture

The following diagram shows the process of extracting data from the NDW using either an ETL or a Federation process, loading it into the GDM, and applying security and end-user access. All data originates from the National Data Warehouse (NDW), the source system. During the load process, tables in the NDW are converted to flat files and then loaded into the GDM tables.

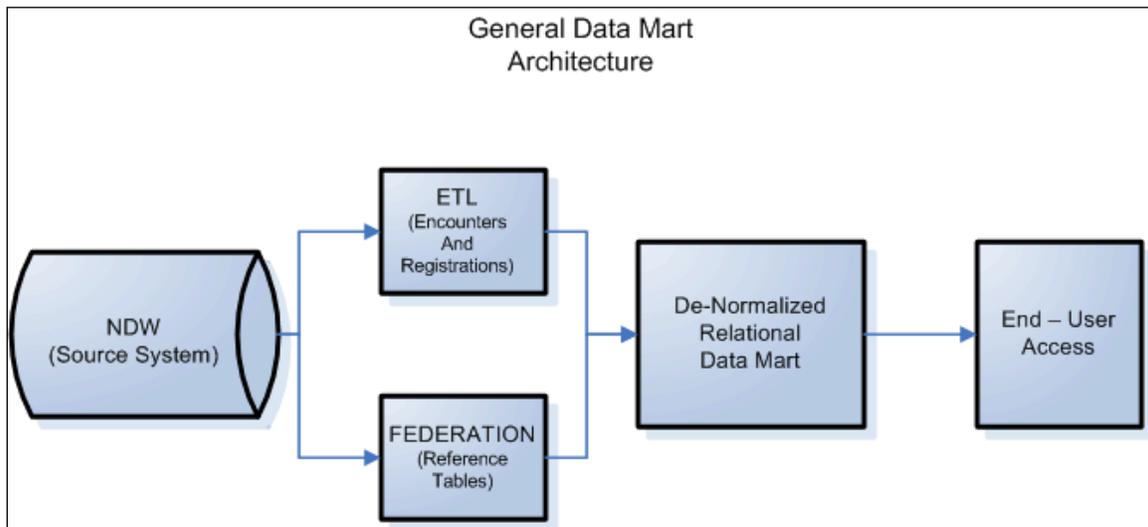


Figure 1 - *GDM Architecture*

A description of each process shown in the architectural diagram is provided in subsequent sections of this document. For general information about NDW architecture see the *NDW Production Database Technical Guide*.

General Data Mart Design

The tables in the GDM are representative of the original tables in the NDW. Transformations have been applied to achieve the goals of the GDM. A full explanation of these transformations is given in the “Extract, Transform and Load (ETL)” section of this document.

General Data Mart Schemas and Tables

The following table contains a brief description of the most commonly used schemas in the General Data Mart:

Schema	Data Type	Description
REG	Registrations	Complete set of registration tables; includes all columns and rows from the NDW REG schema tables. SSN is removed from the PAT_REG table and stored in a separate table.
REG_NP	Registrations Non-PII	Subset of the registration data excluding PII data. These tables are used as the base tables for the views/MQTs for the various security views.
ENCTR	Encounters (Current)	Complete set of encounter tables; it includes all columns and rows from the NDW ENCTR schema tables. SSN is removed from the ENCTRSS table and stored in a separate table. These tables are used as the base tables for the views/MQTs for the various security views.
ENCTR_HIST	Encounters (Historic)	Subset of NDW ENCTR_HIST schema tables; used for synchronization only. It is not to be used for reporting.
ADMIN	Admin	Subset of the generic informational tables to aid in the use of the GDM.
REF	Reference	The complete set of NDW reference tables.
META	Meta Data	Subset of the NDW META tables; contain basic information on the structure and definition of columns and tables used in the mart.
LEGACY	Legacy	Snapshot of historic data from the NPIRS Legacy tables. Since this is historical data (Pre-NDW), it is not refreshed.

For a detailed list of General Data Mart related tables, views, and MQTs, see the following documents:

- *NDW Schemas and Tables/Views/ Nicknames*
- *NDW Reference Tables*

Descriptions of data elements are available from the IHS Meta Data internet web site's Meta Data Dictionary page

at: http://www.ihs.gov/scb/metadata/index.cfm?module=lookup_dictionary&option=list&newquery=1

Legacy Schema

The Legacy schema was established to provide data from the NPIRS legacy system (Pre 2006) and is available for query only purposes. Selected legacy data is available for query by authorized users needing access to historical data. For more information, see the Legacy Data Mart Getting Started Guide. Security

End-User Access

Only authorized users are allowed access to the General Data Mart. The NPIRS Program Manager, working in collaboration with the NPIRS Investment Owner, advises the contractor on how to determine who is granted access to the General Data Mart, as well as deciding the duration of time for which that access will be granted.

End-users can access the data via several methods (i.e., ODBC, JDBC, OLE, CLI, etc.). NPIRS works with the users to provide tables for queries but does not support the applications used for accessing the GDM.

The end user can access data only after passing through 3 types of security:

- Multi-layer authentication (AIX server security)
- Database security
- Label-based access control (LBAC)

Label-based access control (LBAC) is used to allow users to query data that is appropriate to their authorized level of access. An explanation of LBAC implementation is given in Appendix C.

Security controls are appropriate for a query based database and adhere to IHS standards including NIST 800-53.

Area Access

Area users may access only Area based data from the ENCTR, REG or REG_NP schemas and data relevant to their security access. Area Access is controlled by the Security Administrator and based on the user's ID.

Partially Cleansed Data

Partially cleansed data refers to tables with selected Personally Identifiable Information (PII) removed. Classification of PII data columns for exclusion (e.g. partially cleansed) was determined by mitigating the risk of unauthorized use of IHS data. For more

information about the partially cleansed data columns see the document entitled *Analysis & Recommended Restricted PII Data – General Data Mart*.

Tables are identified by their schemas. The REG_NP schema relates to data that is not generally PII sensitive (partially cleansed). An explanation of how the REG_NP schema and table-level security is used in implementing data security is given in Appendix A. The ENCTR schema tables generally do not contain PII data.

The following columns are considered PII data and have been excluded from the partially cleansed tables:

- FATHER_FIRST_NM
- FATHER_LAST_NM
- FATHER_MID_NM
- FIRST_NM
- FULL_NM
- LAST_NM
- MID_NM
- MAIL_ADDR_1
- MAIL_ADDR_2
- MOM_MAIDEN_FIRST
- MOM_MAIDEN_LAST_NM
- MOM_MAIDEN_MID
- NM_SUFIX
- NM_TITLE
- PLCY_NBR
- PLCYHLDR_FIRST_NM
- PLCYHLDR_FULL_NM
- PLCYHLDR_LAST_NM
- PLCYHLDR_MID_NM
- SSN
- SSA_VERIF_L_NM
- SUSP_SSN_FG

Security Access Levels

Only *authorized* users are allowed access to the General Data Mart, and are assigned one of the following current security levels:

- **National Level 1** access allows a user to view all data for all Areas.
- **National Level 2** access allows a user to view partially cleansed data for all Areas.
- **Area Level 1** access allows a user to view all data within his/her specified Area.
- **Area Level 2** access allows a user to view partially cleansed data within his/her specified Area.

Additional security levels may be assigned or created in the future.

Below is a graphical depiction of the different levels of access.

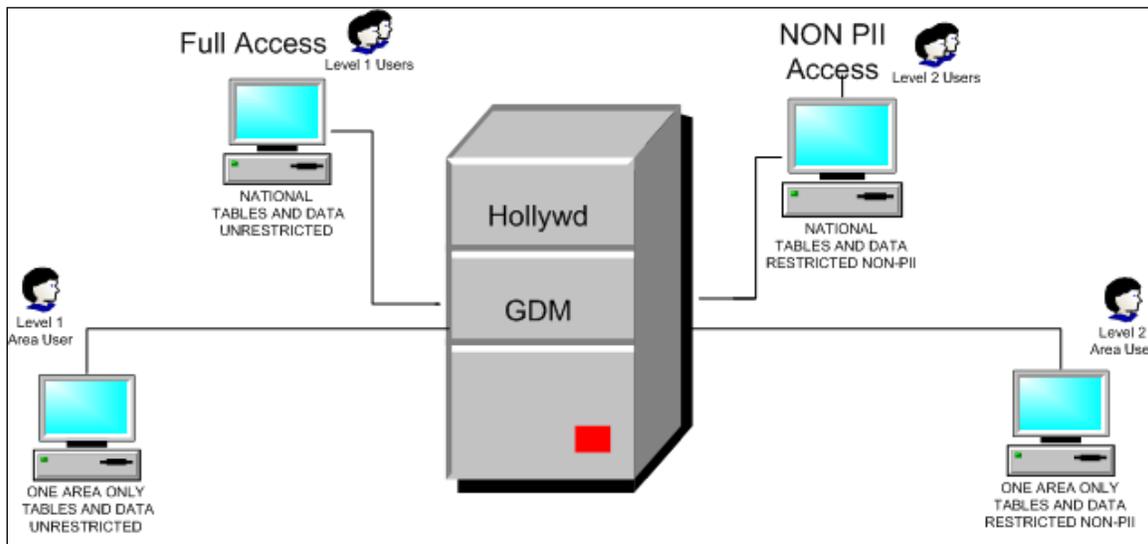


Figure 2 - Levels of User Access

In all cases, authorized users have **Read-Only** privileges.

Federation

The reference tables are sent to the GDM using a federation process. Federation is a DB2 process of connecting external databases or data sources through a Distributed Relational Database Architecture (DRDA) on an enterprise system, enabling a user to access distributed data regardless of where it physically resides. This allows applications to access multiple remote tables at various locations and to have them appear to the end user as if they were a logical whole. For more information about federation see the *NDW Production Database Technical Guide*.

Extract, Transform and Load (ETL)

All tables except the reference tables are sent to the GDM using an ETL process. Many of the tables are split into smaller tables in order to expedite queries. The specific schedule of the ETL process may be found in the *Service Level Agreement – General Data Mart*. ETL processes typically run from the instance owners. For more information about overall ETL processes, see the *NDW Production Database Technical Guide*.

Extract

The ETL process first extracts the data from the NDW source system as a flat file which includes appropriate transformations. The data is subsequently loaded into the GDM tables.

Transform

The transform stage is a series of rules or functions that are applied to the extracted data from the source before it is loaded into the end target. The transformations for the GDM are created using Structured Query Language (SQL).

Three types of transformations are utilized:

- Subsets of tables.
 - Subsets are created to remove Personally Identifiable Information so that users with lower level access can query the data.
 - Split large tables into smaller subsets to expedite queries (See Appendix B).
- Formatting data.

- ICD9 codes are transformed into the industry standard format. ICD9 is a code set used to classify diseases, injuries, and causes of death by etiology and anatomic localization enabling clinicians, statisticians, politicians, health planners and others worldwide to speak a common language.
- Primary keys and indexes.
- Primary keys are created to ensure uniqueness within a table and to facilitate the joining of tables.
- Indexes are created to expedite queries.

Load

There are two types of load processes, incremental and baseline. The tables are refreshed per the *Service Level Agreement – General Data Mart*.

Baseline Loads

Baseline loads are loads that replace all the records in the tables. The Admin, Registration, Encounter History, and Meta Data tables are refreshed using this technology in every ETL.

The encounter tables are loaded using the baseline process 4 times per year for quality control.

Incremental Loads

Incremental loads use the Insert, Update and Delete (IUD) method of loading data into the tables. Incremental loads are utilized for the encounter tables routine refresh loads. During the incremental load new records are inserted and existing records are updated or deleted. Deletions are based on the encounter history tables.

Backups

Backups are performed on a weekly and monthly basis using the Tivoli Storage Manager. If the mart becomes corrupted or is lost, it can be reconstructed using the backup or baseline extract/import process defined in the “Extract Transformation and Load (ETL)” section.

Backups are also critical for use in the event the Emergency Management Plan is activated. More information about the Emergency Management Plan can be found in the abridged version of the *Emergency Management Plan (EMP) for the National Patient Information Reporting System (NPIRS)*.

Appendix A: REG_NP Table Schema

The following tables are directly related to the implementation of security on the GDM:

- **REG_NP.PAT_REG**
- **REG_NP.DEMOGR**
- **REG_NP.INSUR_ELIG**
- **REG_NP.CHART**
- **REG_NP.USERPOP**
- **REG.PAT_REG_SSN***
- **ENCTR.ENCTRSS_SSN***
- **REG.USERPOP_SSN***
- **REG_VERx.USERPOP**

** The REG.PAT_REG_SSN, REG.USERPOP_SSN and the ENCTR.ENCTRSS_SSN tables allow only limited user access, based on the individual user's security access level, to enforce the highest level of patient confidentiality.*

A schema for Registration tables with Non-PII (REG_NP) data houses the tables that have been partially cleansed of Personally Identifiable Information (PII). The REG_NP schema is comprised of registration tables for use by both National and Area Level 2 users.

Views are used to pull the data from the REG schema source tables. A binary security tag column (SECURITY_TAG) is created based on the region abbreviation codes. The data is put into a flat file, and then the weekly driver picks up the file and prepares it for loading into the GDM. LBAC security utilizes the security tag, enabling the administrator to implement row security and restrict access to the data according to the user's group authorization.

The columns for each table are listed in the document titled *NDW Schemas Tables Views and Nicknames*.

Appendix B: Split Tables

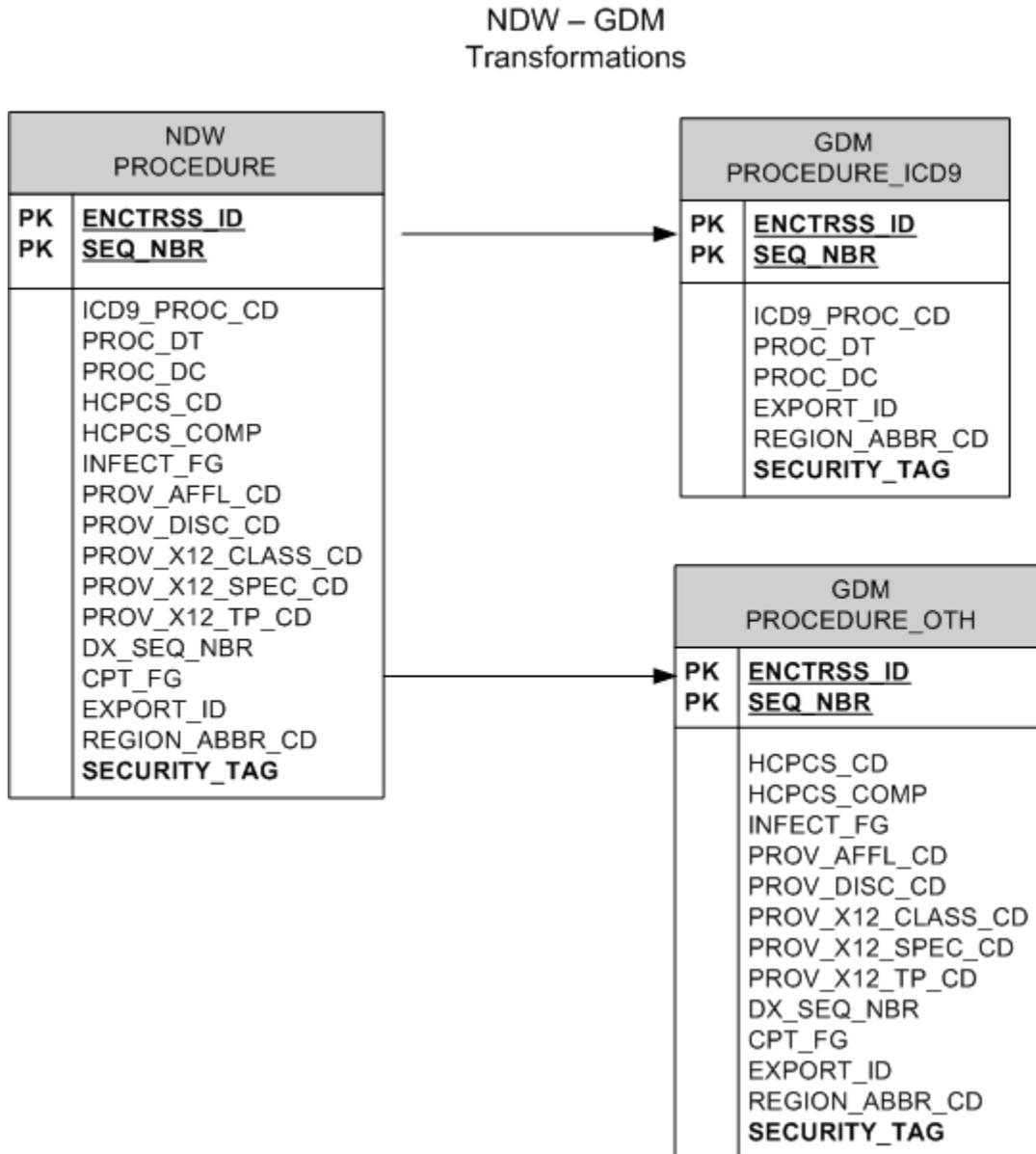


Figure 3 - Procedure Transformations

NDW – GDM
Transformations

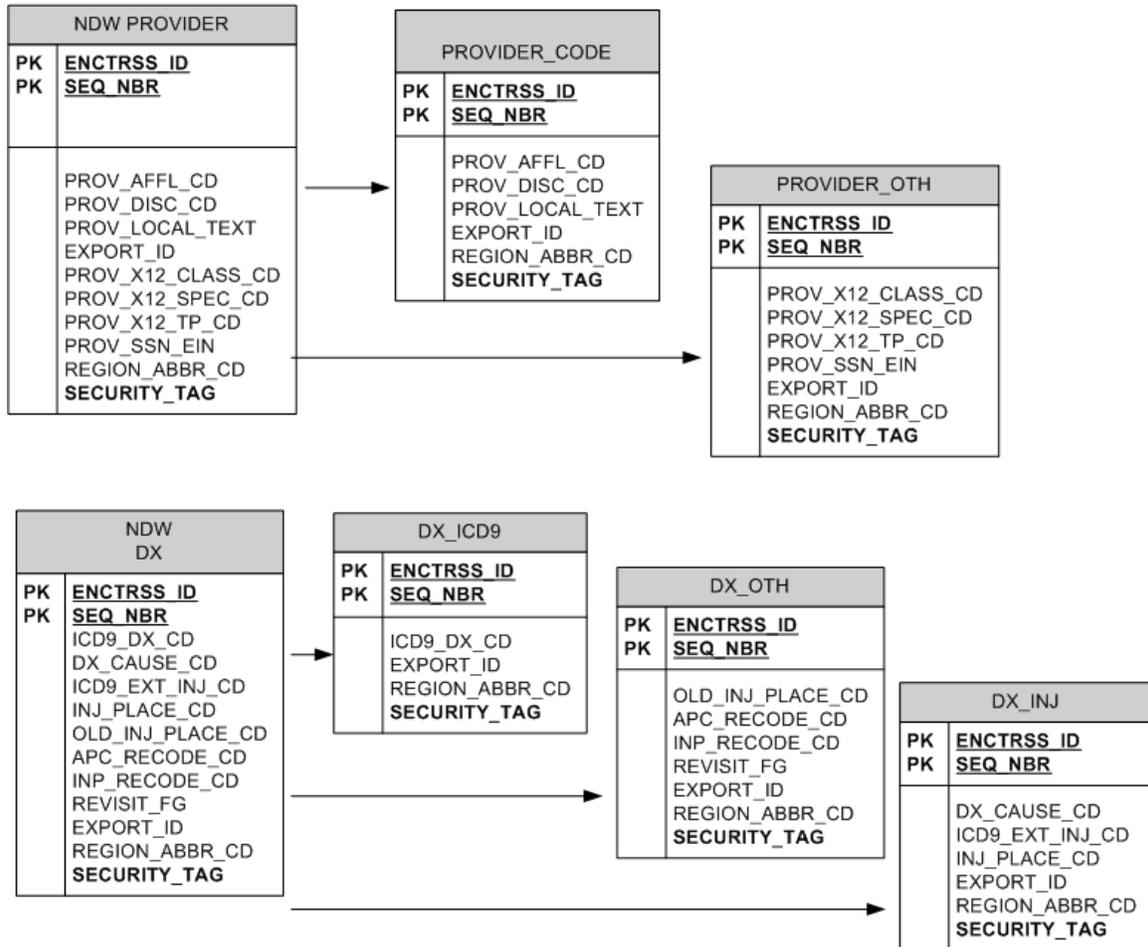


Figure 4 – Provider Details and Diagnostic Code (DX) Transformations

Appendix C: Label-Based Access Control

Label-based access control (LBAC) allows the security administrator to choose whether a user has write or read access to individual rows and individual columns in a table or view.

LBAC Implementation

The security administrator configures the LBAC system based on the security policy of IHS. This is done by creating security label components. A security label component is a database object based on the criterion determined by the level of access a user is allowed.

Below are the security policies for the GDM users:

Authorized users who are granted access to the General Data Mart are assigned one of the following security access levels:

1. National Level 1 Access – PII Data

This level allows the user to read all the data stored in the General Data Mart from all IHS Regions.

2. National Level 2 Access – Non-PII Data

This level allows the user to read partially cleansed data stored in the General Data Mart from all IHS Regions.

3. Area Level 1 Access – PII Data

This level allows the user to read all the data stored in the General Data Mart from a specific IHS Region.

4. Area Level 2 Access – Non-PII Data

This level allows the user to read partially cleansed data stored in the General Data Mart from a specific IHS Region.

After creating the security policies, the administrator creates security labels (objects) for each policy. The security labels are associated with the specific tables or views for an individual user.

The security administrator can grant exceptions when necessary. These exceptions allow a user to access protected data.