
National Patient Information Reporting System: National Data Warehouse

NDW General Data Mart

Technical Guide

Version 4.0

June 2009



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
NDW Source System	3
Extract, Transform and Load (ETL)	3
Federation.....	3
General Data Mart.....	4
Security.....	4
End-User Access	4
General Data Mart Design	5
General Data Mart Schemas and Tables.....	5
Legacy Schema	6
Security	6
Area Access	6
Partially Cleansed Data.....	7
Security Access Levels	8
Extract, Transform and Load (ETL)	9
Extract.....	9
Transforms	9
Loads.....	9
Baseline Loads	10
Incremental Loads	10
Backups	10
Types	10
Appendix A: New Structure – March 2009	11
Process.....	11
Appendix B: Split Tables	12
Appendix C: Label-based access control (LBAC)	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>.</p> <p>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>

Overview

The General Data Mart (GDM) is representative of the Encounter and Registration data as found in the National Data Warehouse (NDW) modified for query optimization. Comprised of a set of tables, views and Materialized Query Tables (MQTs), this data mart provides access to the following data:

- Complete set of all current registration data
- Current and historic user population data
- Complete set of all current encounter data
- Subset of encounter history data
- Complete set of reference table data
- Meta data, which contains information about tables and columns included in the database
- A snapshot of Historic Legacy data tables (not updated)
- Administration tables
- Modified set of registration data that has been partially cleansed of Personally Identifiable Information (PII)

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.
- HOLLYWD is the data base associated with the General Data Mart and is used for connections.
- All data within the General Data Mart is refreshed as stipulated in the *Service Level Agreement – General Data Mart*.
- The data in the General Data Mart is a copy of NDW production data. Tables are available based on the level of a user's security access. 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.
- Access controls are administered that allow users to query data that is appropriate to their authorized level of access.

- Enhanced security controls adhering 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 adhering to security restrictions.
- The General Data Mart is enterprise compliant to allow various environments to access the database, including ODBC, JDBC, OLE, 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.x
AIX Version:	5.3 or above
FTP Address:	198.45.1.8
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, loading it into the GDM, applying security and end-user access.

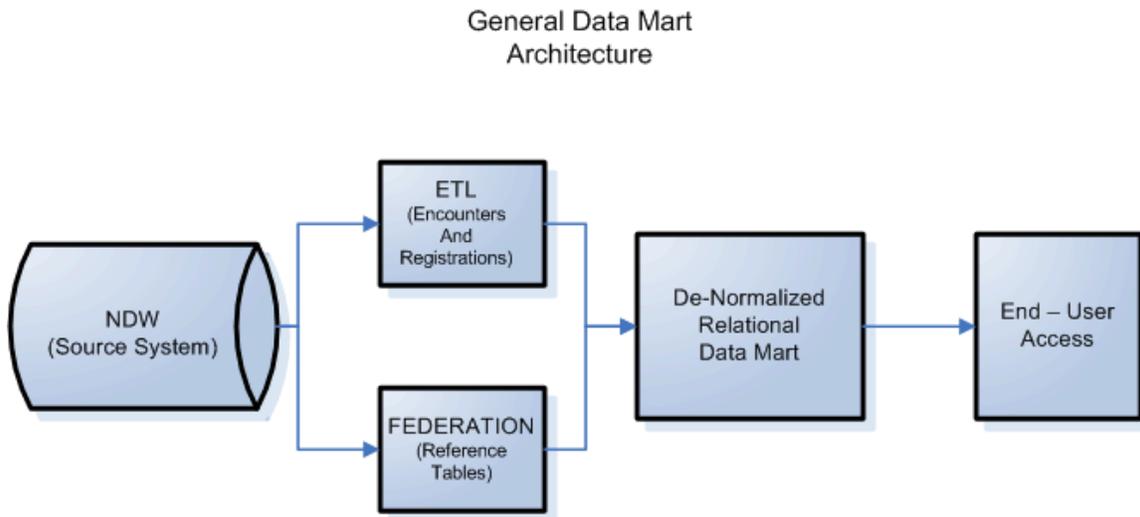


Figure 1. GDM architecture

A general description of the architectural diagram is provided below. For more information about architecture see the *NDW Production Data Base Technical Guide* located at http://www.ndw.ihs.gov/documents/NDW_Production_Datbase_TG_V2.0.pdf.

NDW Source System

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.

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. A detailed explanation of the ETL process is given in the “Extract, Transform and Load (ETL)” section of this document.

Federation

The reference tables are sent to the GDM using a federation process. For more information about federation see the *NDW Data Mart DB2 Glossary* located at <http://www.ndw.ihs.gov/what-if-I-have-other-questions.asp>

General Data Mart

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.

Security

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

- Multi-layer authentication (AIX server security)
- Data base security
- Label-based access control (LBAC) – see Appendix C for more information.

A detailed explanation of the security levels granted to GDM users is given in the “Security” section of this document.

End-User Access

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.

General Data Mart Design

General Data Mart Schemas and Tables

The following table contains a summary 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 NDW ADMIN.EXPORT_INFO table; includes all columns and rows for files that were successfully loaded into the NDW.
REF	Reference	The complete set of NDW reference tables.
META	Meta Data	Subset of the NDW META tables; it includes all columns and rows from the DW_INFO tables.
LEGACY	Legacy	Snapshot of historic data from the NPIRS Legacy tables; used for historic reports only.

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

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

These documents are located at: <http://www.ndw.ihs.gov/what-if-I-have-other-questions.asp>

Detailed descriptions of data elements are available at the IHS Meta Data internet web site: <http://www.ihs.gov/CIO/scb/metadata/>

Legacy Schema

The Legacy schema was established so that historical data from the NPIRS legacy system would be available for query only purposes. Selected legacy tables will be available for query and are being retained online primarily for authorized users needing access to historical data. For more information, see the *Legacy Data Mart Getting Started Guide Version 1.0* at:

http://www.ihs.gov/CIO/DataQuality/warehouse/documents/LegacyDataMart_GettingStartedGuide_V1.0.pdf

Note: The legacy tables are not intended to be utilized to recreate official reports for past fiscal years.

Security

- The NPIRS Program Manager, working in collaboration with the NPIRS Investment Owner, will advise the contractor on how it will be determined who will be granted access to the General Data Mart, as well as determine the time period during which each user will be granted access.
- 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.
- Only authorized users will be allowed access to the General Data Mart.
- Security controls commensurate with those for a transactional, non-query based database and adhering to IHS standards are enforced.

Area Access

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

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 *Restricted PII Data - Requirements*.

Tables are identified by their schemas. The REG_NP schema relates to data that is not generally PII sensitive (partially cleansed). An explanation of the new schema is given in Appendix A. The ENCTR schema tables generally do not contain PII data.

The following columns are considered PII data and are removed 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_SUFEX
- 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 will be 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.

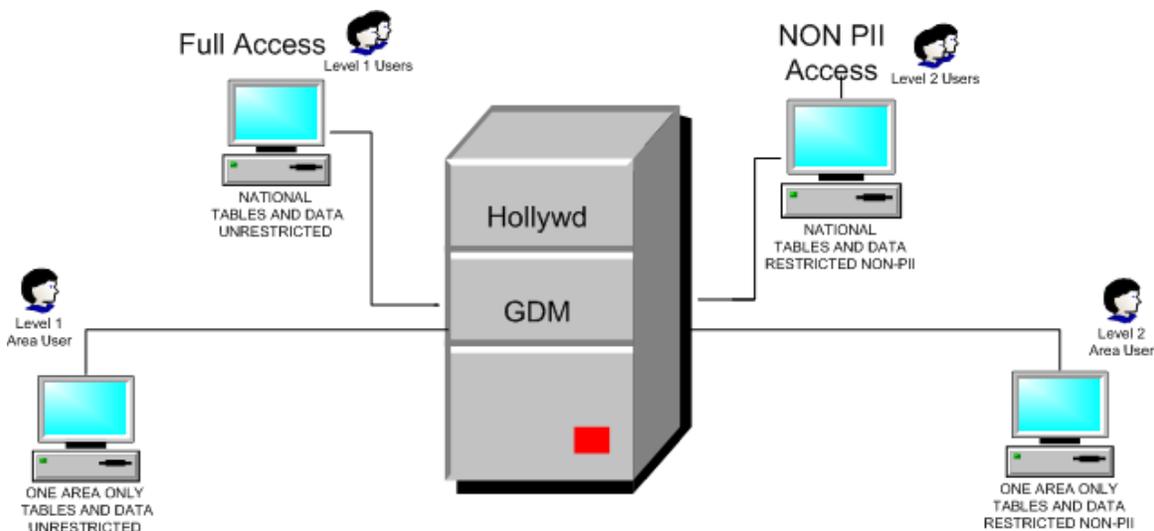


Figure 2. Depiction of different levels of user access.

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

Extract, Transform and Load (ETL)

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 Data Mart DB2 Glossary* located at <http://www.ndw.ihs.gov/what-if-I-have-other-questions.asp>

Extract

The ETL process first extracts the data from the NDW source system. The data is then converted to a flat file, transformed, and then loaded into the GDM tables.

Transforms

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.
- 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.

Loads

There are two types of load processes, incremental and baseline. The tables are refreshed per the *Service Level Agreement – General Data Mart* located at <http://www.ndw.ihs.gov/what-if-I-have-other-questions.asp>.

Baseline Loads

Baseline loads are loads that replace all the records in the data. The Admin, Registration, Encounter History, and Meta data tables are refreshed using baseline loads.

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 regular refresh schedule. During the incremental load new records are inserted and older records are updated or deleted. Updates and deletions are based on the encounter history tables.

A separate update process is also run to ensure synchronization of data changed in the NDW by the Matchmaker process.

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.

Types

Online backups are run weekly at the completion of the Extract Transformation and Load (ETL) refresh and are part of the ETL schedule jobs.

Offline backups are run monthly. During offline backups, no other processing within the database is supported. Offline backups are run on an independent schedule.

Additional detailed information on backup types can be found in the *NDW Glossary* located at <http://www.ndw.ihs.gov/what-if-I-have-other-questions.asp>

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)* located at <http://www.ndw.ihs.gov/what-if-I-have-other-questions.asp>.

Appendix A: New Structure – March 2009

The tables below have been changed (March 2009) to accommodate the new security requirements for the General Data Mart.

The following tables and schemas have been added:

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

The REG_NP schema is for all the tables that are partially cleansed of PII data. These tables are for National Level 2 and Area Level 2 users.

The REG.PAT_REG_SSN and the ENCTR.ENCTRSS_SSN tables were created to provide additional patient confidentiality. These tables will have limited user access based on the security approval obtained by the user.

Process

Views are used to pull the data from the NDW source tables. A security tag is created based on the region abbreviation codes. The data is put into a flat file, 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* located at <http://www.ndw.ihs.gov/what-if-I-have-other-questions.asp>.

Appendix B: Split Tables

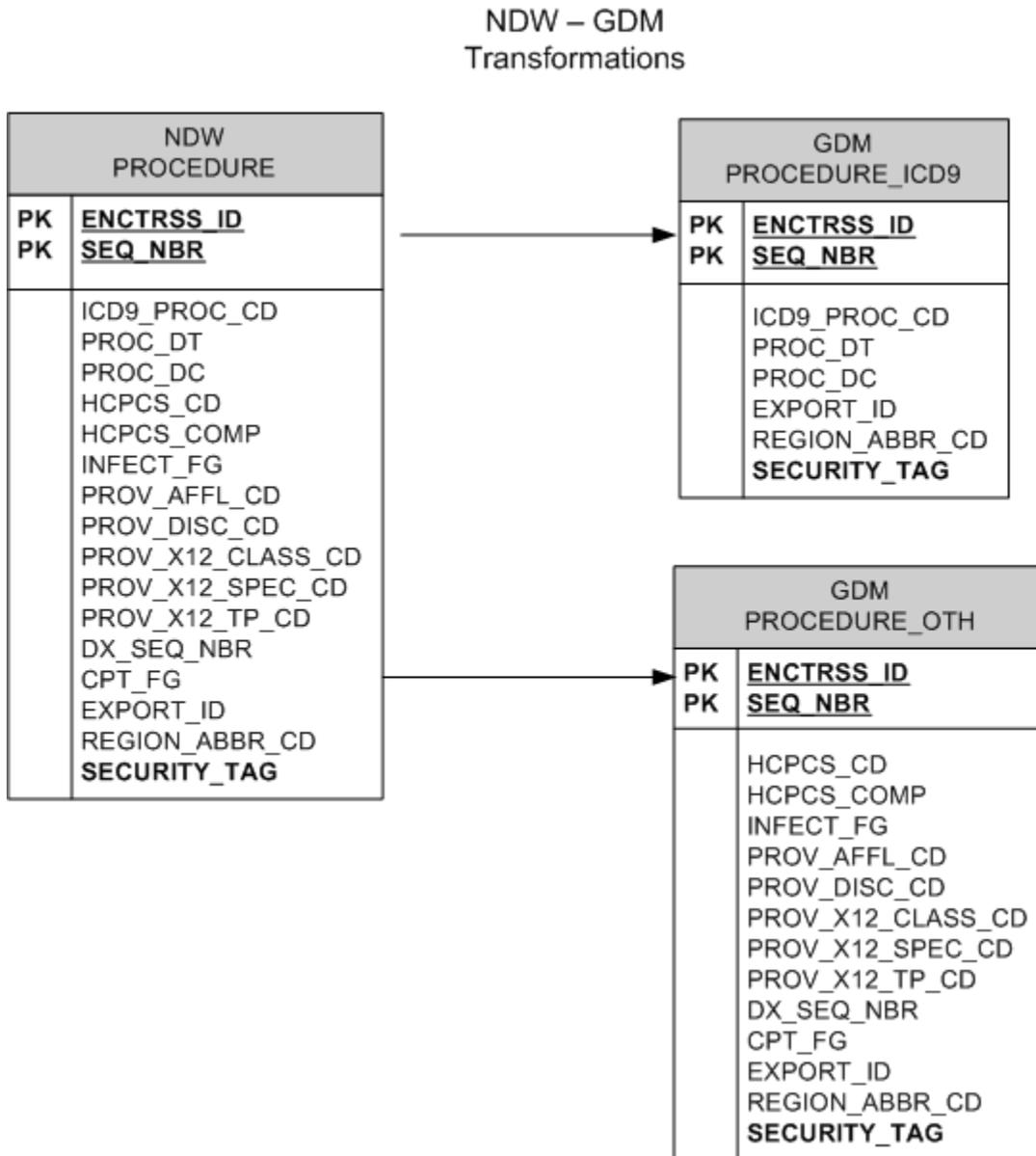


Figure 3. Procedure transformations

NDW – GDM
Transformations

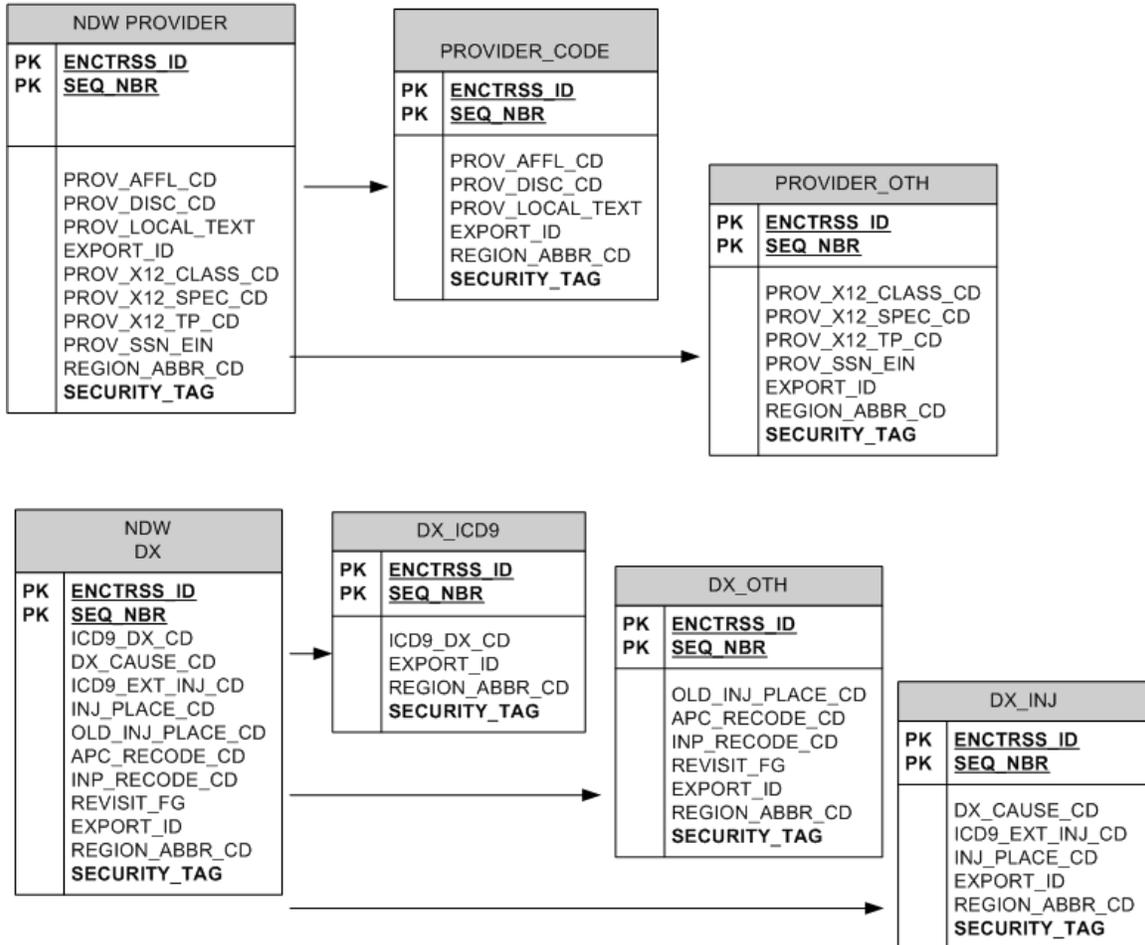


Figure 4. Provider and DX transformations

Appendix C: Label-based access control (LBAC)

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 data base object based on the criterion determined by the level of access a user is allowed.

Below are the security policies for the GDM users:

The General Data Mart users have four levels of security for users:

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 will allow a user access to protected data.