

---

# National Patient Information Reporting System: National Data Warehouse

---

## **NDW Production Database**

### **Technical Guide**

Version 3.0

June 2009



Department of Health and  
Human Services

Indian Health Service

Office of Information  
Technology (OIT)

---

# Contents

<b>Version Control</b> .....	<b>iv</b>
<b>Overview</b> .....	<b>1</b>
<b>Design Parameters</b> .....	<b>3</b>
<b>System Environment</b> .....	<b>3</b>
<b>Security</b> .....	<b>5</b>
<b>NDW Design</b> .....	<b>6</b>
Change Management .....	6
DB2 V9 .....	6
AIX 6.1 .....	7
Environment.....	7
Constructs .....	7
Compression .....	8
Federation .....	8
<b>NDW Process Flow</b> .....	<b>8</b>
MatchMaker Process.....	10
Backups.....	10
Data Integrity Process.....	10
Process Locations.....	11
Reporting Process .....	11
<b>NDW Data Flow</b> .....	<b>11</b>
<b>Related Environments</b> .....	<b>13</b>
Overall System Environment.....	13
Test and Quality Assurance Environments.....	16
Tools .....	18
Controls.....	18
Emergency Management Plan (EMP) .....	18
On-Going Plans.....	19
<b>Appendix A: Weekly ETL Process</b> .....	<b>21</b>
Encounter Weekly ETL .....	21
REG Weekly ETL.....	26
ENCTR_HIST Weekly ETL.....	27

MatchMaker Weekly ETL .....	31
ADMIN_INFO Weekly ETL.....	32

## Version Control

Version	Date	Notes
1.0	September 2007	FY07 Contract Deliverable (D1.12.3) Accepted October 15, 2007 Appendix C - NDW Column Detail is a separate document, in PDF format.
2.0	February 2008	Review, update; add/expand: add DB2 V9, Indexes, MQTs, Tablespaces (including Large), Sequences, Nicknames, Compression sections; updated figures, added Topology, Current and Planned Server Configuration figures; Appendix A, B now separate documents; Appendix C now available at NDW Meta Data web site, as noted in text. FY08 Bridge Contract D1.7.3 COTR approved April 10, 2008
3.0	June 2009	Updates to diagrams, glossary information removed, DB2 9 specific issues, and Enterprise level information added. COTR approved June 16, 2009

## Overview

The National Data Warehouse (NDW) environment is comprised of the following:

Three main databases:

(1) Sandia

is the transactional database, the source for most data mart extracts. (One data extract is run from Wildhrse for Userpop Reporting.) Temecula is the sole source of reference tables for Sandia, Wildhrse, and all data marts through federation. This is also the repository of the Export Tracking Mart.

(2) Wildhrse

is a mirror (with a few exceptions) of Sandia, which allows query capability without affecting transactions, and also serves as the computational database for user population information. This database is also the repository for the Data Quality Mart.

(3) Temecula

is the sole source of reference and meta-related tables for all other databases.

Supplemental databases:

(1) Hollywd

is a fully supported reporting database that is the repository of the General Data Mart, utilized primarily by Indian Health Performance Evaluation System (IHPES) and Epidemiology. It contains reportable NDW data (no historical NDW data). It is also a repository for the legacy system data.

(2) Isleta and Laguna

are the sources of NPIRS Internal Test Environment, which allows for complete testing of all changes to the NDW system prior to implementing those changes in the production system.

(3) Nizhoni

is the repository for the Data Integrity Mart.

(4) Tatonka

is a fully supported reporting database utilized primarily to produce reporting for the Workload Reporting Mart.

(5) RTE66

is the NPIRS Data Mart Developer's Test Environment, which allows NPIRS and others outside of NPIRS to develop and test their data mart and the ETL processes from a sample NDW database to their data mart, prior to their implementation in the production environment.

(6) Palms

is a scratchpad database for use in developing meta data prior to implementation.

The NDW utilizes tables, views, sequences, and temporary tables to maintain data received from remote sites and areas in an orderly fashion for use in marts, queries, special requests, and data mining. The design of the NDW maximizes performance for transaction processing. Each logically grouped set of tables shares both a common tablespace (storage area) as well as a common buffer pool (memory area) to minimize disk I/O.

Data is sent to the NDW, typically via FTP, and is received by the Integration Engine (IE). The IE recognizes the file format and sends the file to the NDW directories for loading. The IE is maintained by IHS/Enterprise Tech Services (ETS).

The Acknowledgement (ACK) process recognizes the file, logs the file into the NDW database, and sends an e-mail to the sending site that the file has been received.

The Import Engine (IMP) process subsequently loads the files into the database, where the data is available for transmission and loading to the various marts. An acknowledgement is sent to the sending site with the following information – filename, record count, and date received.

After the file is loaded, a report is emailed to the sending site containing the same information as the acknowledgement and the date loaded in the NDW and data transform/reject issues.

A variety of ETL processes are used to transport data, either transformed or untransformed, from Sandia to all other databases. ETL processes are either automated and scheduled or are done on an on-demand basis.

Process descriptions included in this document refer to the Sandia database, unless otherwise noted.

## Design Parameters

- The NDW environment is a source system for the data marts. Some non-NPIRS data marts use other sources. The NDW can be modified to accommodate future business needs if necessary.
- The various databases and marts have been workload leveled across several servers to maximize performance and to minimize downtime risk due to hardware failure. Data is stored on an enterprise level RAID 10 storage system.
- All data within the NDW is updated continuously from files received from remote sites.
- Availability for transactions is maximized. The NDW is kept online 24/7, except for maintenance. The IMP process is stopped during ETL processes to ensure synchronization and integrity of data..

## System Environment

This is the physical environment of the NDW main database:

Server	FTP Address	AIX Level	Database	Partition Level	DB2 Level	Comments	SLA Related
Gollum	198.45.1.20	6.1+	Sandia	4	V9.5.x	Primary NDW database.	YES
Gollum	198.45.1.20	6.1+	Nizhoni	1	V9.5.x	Data Integrity database. Mart	NO
Smeagol	198.45.1.21	6.1+	Temecula	2	V9.5.x	Reference and Meta table database.	NO
			Wildhrse	4	V9.5.x	Calculation, reporting and research database; Userpop calculation and archive database. Future source of UP/WL Mart	YES

Server	FTP Address	AIX Level	Database	Partition Level	DB2 Level	Comments	SLA Related
Bilbo	198.45.1.8	6.1+	Hollywd	2	V9.5.x	General Data Mart DB; volatile structure based on User demands.	YES
			Rte66	2	V9.5.x	NPIRS Data Mart Developer's Test Environment or "sample" data mart used to determine structure and design of future data marts. Operational on demand only.	YES
Smeagol	198.45.1.21	6.1+	Tatonka	2	V9.5.x	User Population / Workload Reporting Data Mart	YES
Arwen	198.45.1.8	6.1+	Isleta	2	V9.5.x	Primary development database. Small mirror image of NDW primary; can be refreshed or restored to baseline. Used for unit testing.	NO
			Caesar	2	V9.5.x	Secondary development database. Small mirror image of NDW primary; can be refreshed or restored to baseline. Used for unit testing.	NO
			Laguna	2	V9.5.x	Primary QA database. Small - medium mirror image of NDW primary; can be refreshed or restored to baseline. Used for integration testing prior to release to production	NO

For a detailed list of NDW related schemas and tables, see the following documents:

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

See the *NDW Physical Models* document for corresponding models.

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

## System Access

All databases are enterprise compliant to allow various environments to access the database, including ODBC, JDBC, OLE, and CLI.

Native access languages supported are SQL, SQLJ, SQL-Proc, and XML (under UTF-8) and direct.

---

**Note:** ANSI92 SQL statements are not supported under UTF-8 and may cause unreliable results. ANSI98 SQL and above is supported.

---

## Security

The NDW, unless otherwise specified, is restricted from outside access. Access is limited to NDW personnel for maintenance, data refreshing, and report production.

- Only authorized users are allowed access to the NDW or its associated databases/data marts.
- Security controls commensurate with those for a transactional, non-query based database and adhering to IHS standards, as outlined in separate security documents, are enforced. Users may access the NDW and its associated marts in read-only mode. Updates are restricted as follows:
  - Temecula is updatable for tables associated with the SCB (Standard Code Book) only by users authorized to make such changes. Other reference and Meta tables are updatable only by authorized users. All changes in the Temecula database are tracked in a plain English audit table.
  - Security controls associated with individual data marts are specified in their associated technical documents.

- Authorized users may access only data for which they are authorized. This is controlled by multiple level access control. These levels consist of user authentication at both the operating system and database, role based security and Label Based Access Control (LBAC). Additional information on LBAC can be found in *NPIRS Accepted Practices (NAP) 9. LBAC Administration Practices for NPIRS General Data Mart V1.0*.
- Additional security will be added throughout the enterprise including LDAP and more extensive use of LBAC.

## NDW Design

The NDW environment utilizes multiple schemas which are logical groupings of either related data or functions. Additional information on Schemas, Tables, and Columns used in the NDW can be found in *NDW Schemas Tables Views and Nicknames*.

The NDW is a multi-partitioned, multi-node environment designed to maximize performance. A node is a logical and physical partition within the database that enhances performance by promoting CPU parallelism and opportunistic behavior and improves I/O throughput. The NDW is supported by a hybrid SQL/XML engine under UTF8 to support future enhancement and capabilities in anticipation of the growing capability of XML/XQuery.

## Change Management

Change Management of the database or processes follows IHS standards and *NAP 4. Change Management Process V1.0*.

## DB2 V9

DB2 V9 provides strong and comprehensive capabilities to enhance the Enterprise Information Integration (EII), ETL, and the Enterprise Application Integration (EAI) capabilities of the NDW. A Meta data layer is currently being expanded to enhance the ETL layers, and a number of tools are being used to enhance the EAI paradigm. Under V9, multiple applications can be more tightly linked, using a single repository such as Workbench, and reduce the number of stovepipe applications.

Upgrades to the current Fixpack level are anticipated and will be performed as needed.

## AIX 6.1

AIX 6.1 provides the base operating system and controls I/O, backup storage, communications, and first level of security. Features used in the NDW include scheduling, Simultaneous Multithreading, Micro partitioning and intrusion and event logging. Workload management is in the process of being implemented. Workload management will allow the system to automatically assign/restrict resources to achieve maximized performance. It will also ‘borrow’ under utilized resources from other servers during periods of heavy demand.

## Environment

The NDW exists primarily on disparate AIX servers with connection to Windows Servers for I/O interfaces. The specific configuration and interconnectivity is portrayed in a diagram below. Due to the expansion requirements, new technology, and security modifications, the connectivity diagram is subject to change without notice. Additional connections within the IHS cloud or via a VPN connection into the cloud are supported.

## Constructs

These are the main construct types associated with the NDW:

- Table
- Indexes
- Views
- MQTs
- Temporary Tables
- Tablespaces
- Sequences
- Nicknames

Information on these constructs can be found in the *NDW Data Mart DB2 Glossary* located on the NDW Informational web site:

<http://www.ihs.gov/CIO/DataQuality/warehouse/what-if-I-have-other-questions.asp>

Information on the naming conventions for the database constructs can be found in *NAP 5. Naming Conventions User Objects*.

## Compression

Information on Compression can be found in the *NDW Data Mart DB2 Glossary* located on the NDW Informational web site:

<http://www.ihs.gov/CIO/DataQuality/warehouse/what-if-I-have-other-questions.asp>

## Federation

Information on Federation can be found in the *NDW Data Mart DB2 Glossary* located on the NDW Informational web site:

<http://www.ihs.gov/CIO/DataQuality/warehouse/what-if-I-have-other-questions.asp>

## NDW Process Flow

The following figure illustrates the NDW process flow.

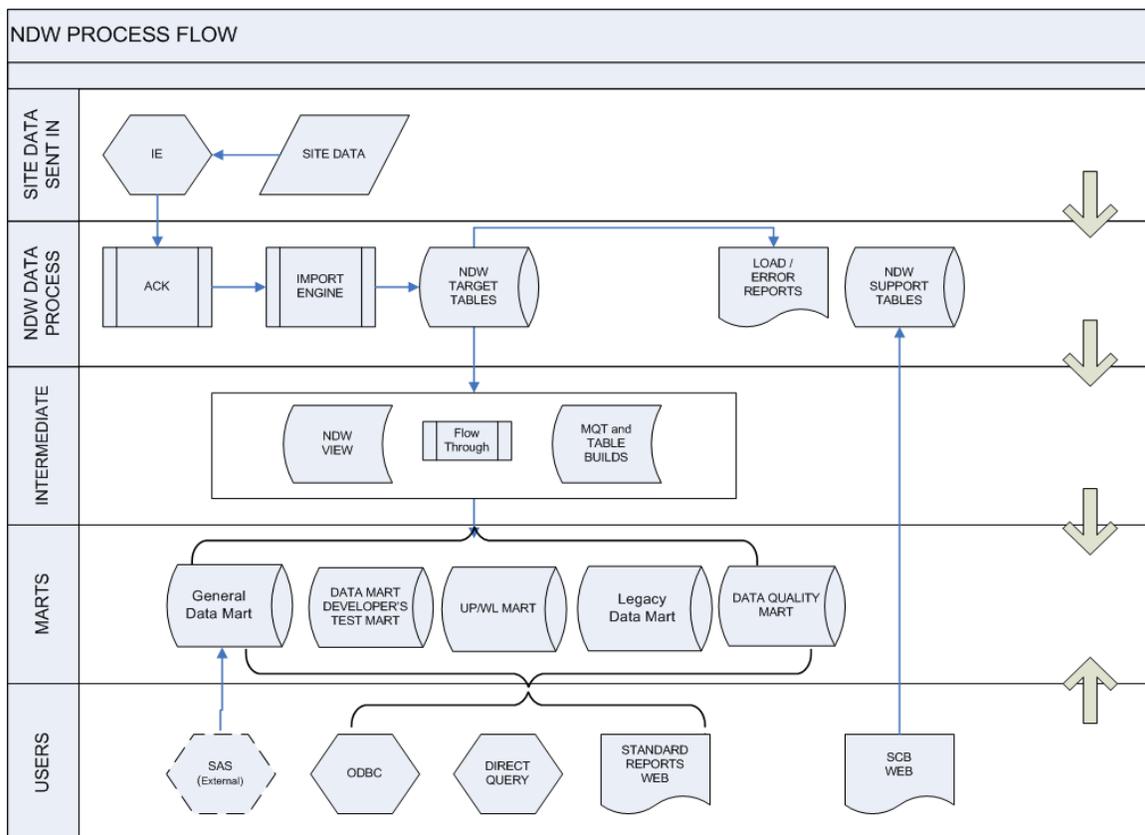


Figure 1. NDW Process Flow

A sequence of diagrams outlining the ETL process can be found in Appendix A.

All extract processes are SQL based processes with simple AIX shell script wrappers.

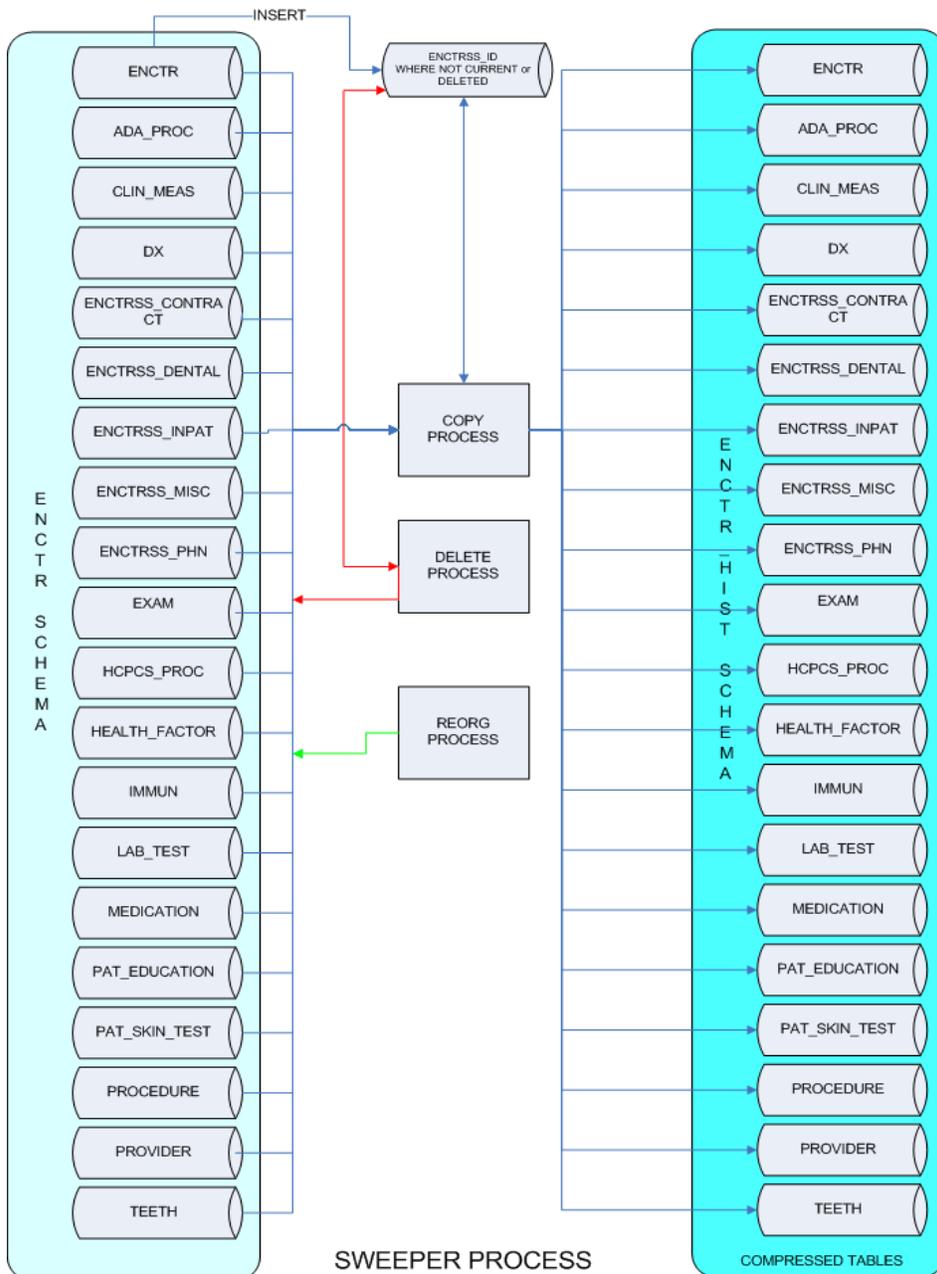


Figure 2. Sweeper Process

## MatchMaker Process

The purpose of the Matchmaker process is to correlate or link previously unlinked data in the ENCTR tables with data in the REG tables. Data is normally linked during the IMP process; however, there are times that ENCTR data is loaded before there is corresponding REG data. The MatchMaker process typically runs as part of the ETL process tree. Additional information on the MatchMaker process may be found in the document *SPC\_327 Promoter Elimination Specifications V1.6*.

## Backups

Backups of the NDW database are performed daily. Several types of backups are available: full offline, full online, incremental, delta, Mobility on Demand (MOD), and tablespace. The type of backup is parameter driven in the backup script, depending on the day of the week.

The NDW has multiple nodes, each backed up individually. In the event of a restore, the backup information of each node is required. Backups are controlled and managed by the Tivoli system and the DB2/Tivoli interface. DB2 tracks

- Backups performed
- Type of backup
- Nodes affected
- Associated logs
- Time window performed

## Data Integrity Process

The Data Integrity Process compares data physically stored within the NDW with Post-IE data files to ensure data received is stored properly. Business rule transformations are addressed within the process. This process occurs on a periodic basis and utilizes both the Sandia and Nizhoni databases.

## Process Locations

Processes associated with Backups, Matchmaker, ETL (except Userpop/Workload initial extract), and monitoring are located within the Instance owner userid directories. Other processes, such as IMP, ACK, reporting are stored in other directories. Security access is applied to the process locations to prevent inadvertent changes to production code and to prevent source code vulnerability scans as defined by NIST 800-53.

## Reporting Process

Reports, except one time reports, are generally driven by a SQL stored procedure utilizing one or more defined views. The stored procedures are typically defined utilizing the IBM Workbench.

## NDW Data Flow

The NDW receives and processes data from multiple sources. The NDW loads data irrespective of the source as long as the data are exported in one of the published standard formats.

Sites transmit data export files through FTP to IHS/OIT. On arrival, the data export file goes first to the IE, where it is reformatted and output as a post IE data export file. After the post-IE data export moves the file to the incoming NDW directory, the ACK process assigns an internally generated unique export\_id to the data export and then moves the file to the appropriate directory. The data export is then processed and loaded into the NDW (SANDIA database).

Once loaded, selected and transformed data is copied to WILDHRSE and HOLLYWD, which are synchronized with SANDIA on a periodic basis. The design paradigm allows maximum throughput transaction processing for the NDW environment. WILDHRSE is used for special data requests, data investigation, data archive, and user population calculations. For more information, see the *User Population/Workload Data Mart Technical Guide Version 1.2*. Users external to NPIRS use HOLLYWD to satisfy ongoing needs to query and report on data. For more information, see the *General Data Mart Technical Guide Version 4.0*.

Current plans will move the Tatonka Userpop/Workload data mart to be self contained on WILDHRSE. The purposes of this change is to maximize the transactional functions of Sandia with little or no impact to data loads, to minimize ETL and calculation time and impact for Userpop/Workload processes, and to improve data research capability.

The following figure illustrates the general data flow through the NDW environment.

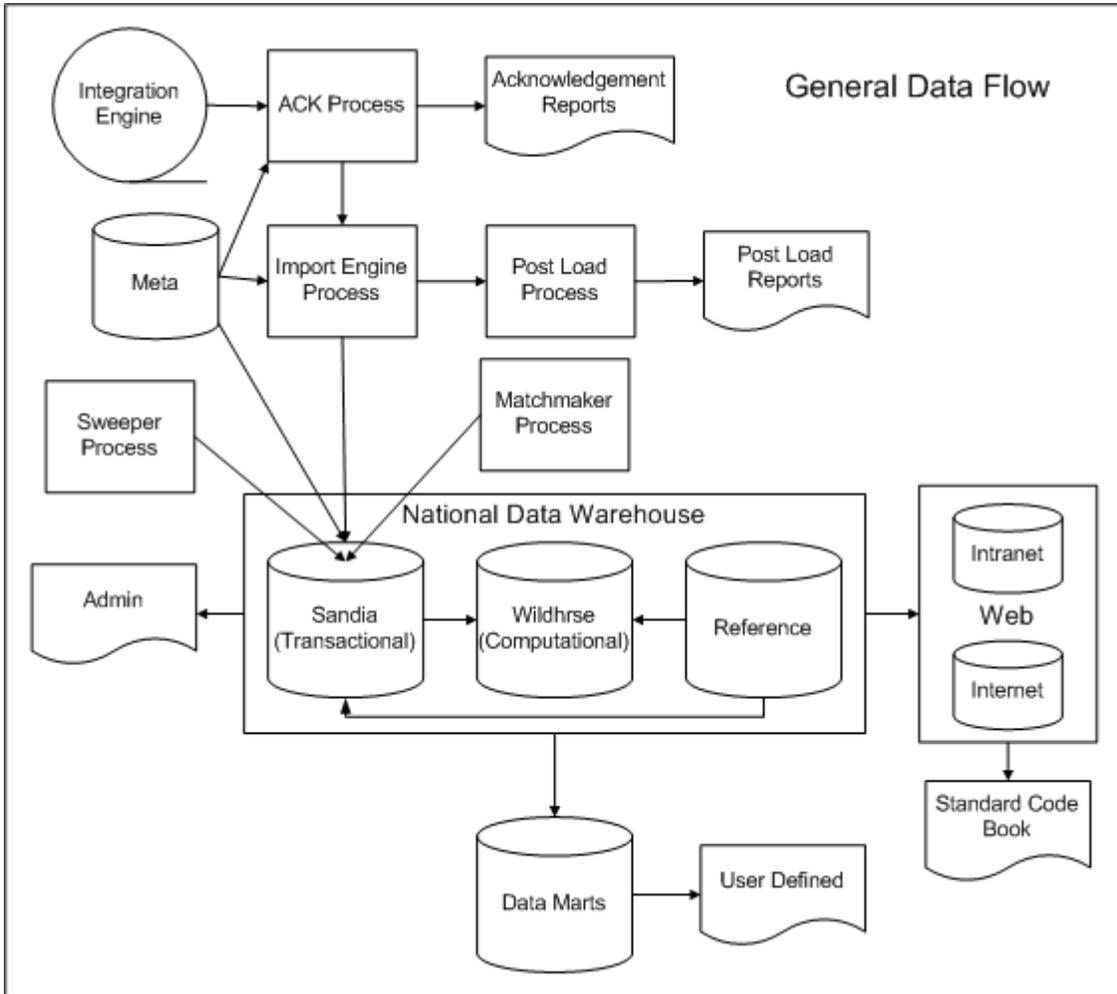


Figure 3. NDW Data Flow

## Related Environments

### Overall System Environment

The following figure illustrates the role of the NDW environment within the overall system environment.

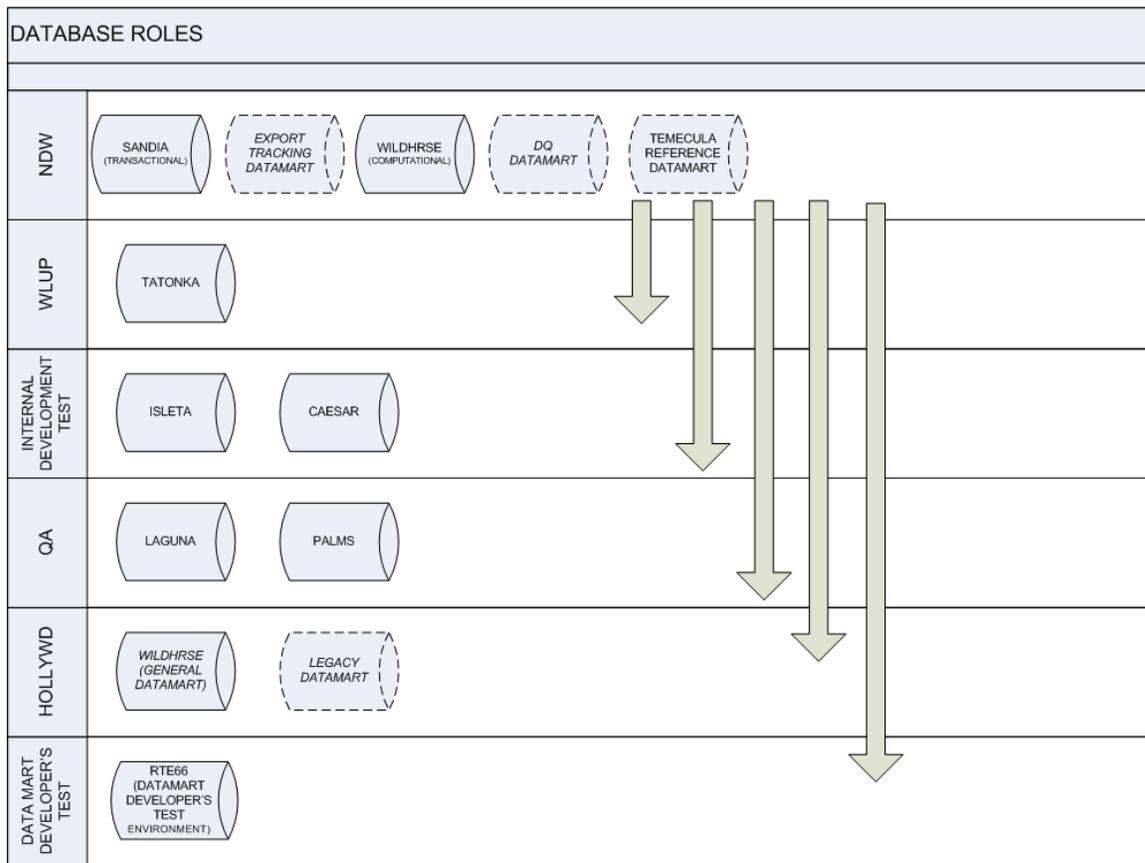


Figure 4. Database Roles

Note that the Legacy Data Mart was imported from the Legacy NPIRS database and contains pre-2006 data. For more information, see the *Legacy Data Mart Getting Started Guide Version 1.0*.

The following diagram illustrates the interrelationship of the NDW, related marts and associated hardware.

Current Server Configuration for IHS Data Warehouse and DataMarts

February 28, 2009

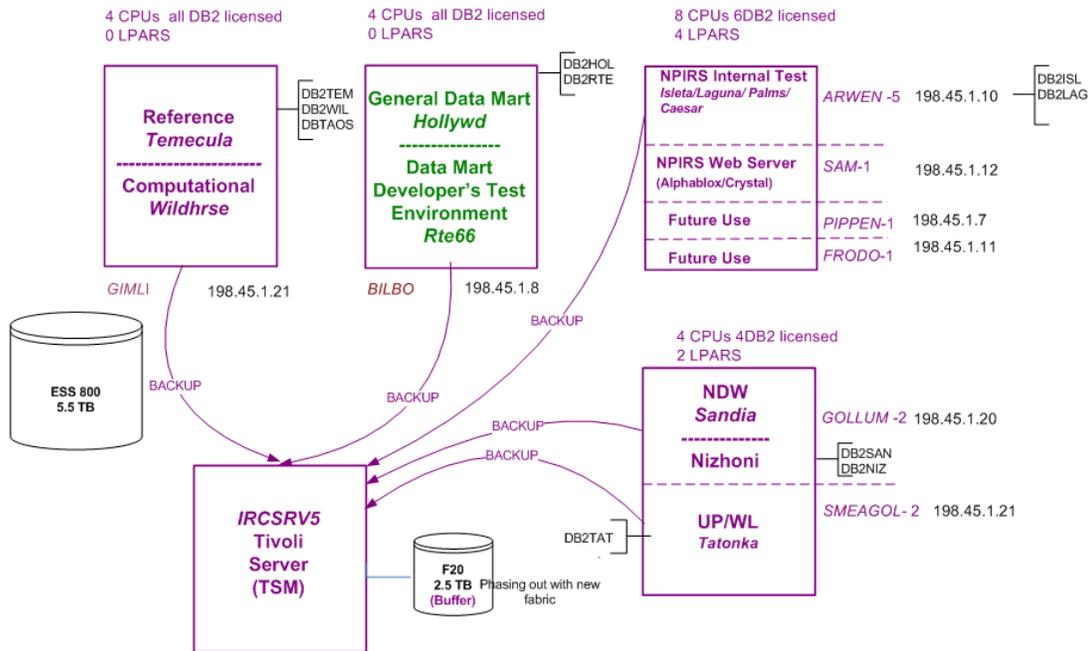


Figure 5. Current IHS Data Warehouse/Data Marts Server Configuration

The following diagram illustrates the physical relationship of the NDW and associated hardware.

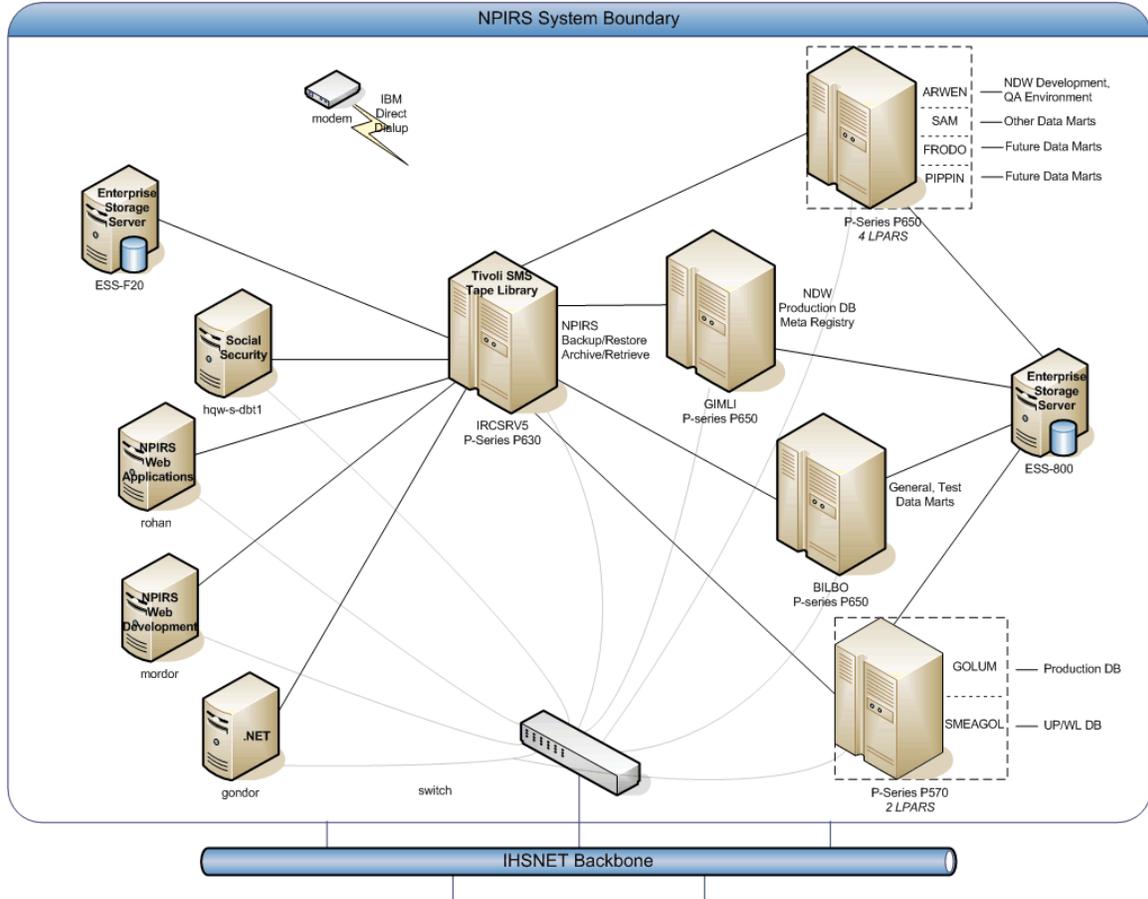


Figure 6. NPIRS System

The following diagram illustrates the high level topography of the NDW and related marts.

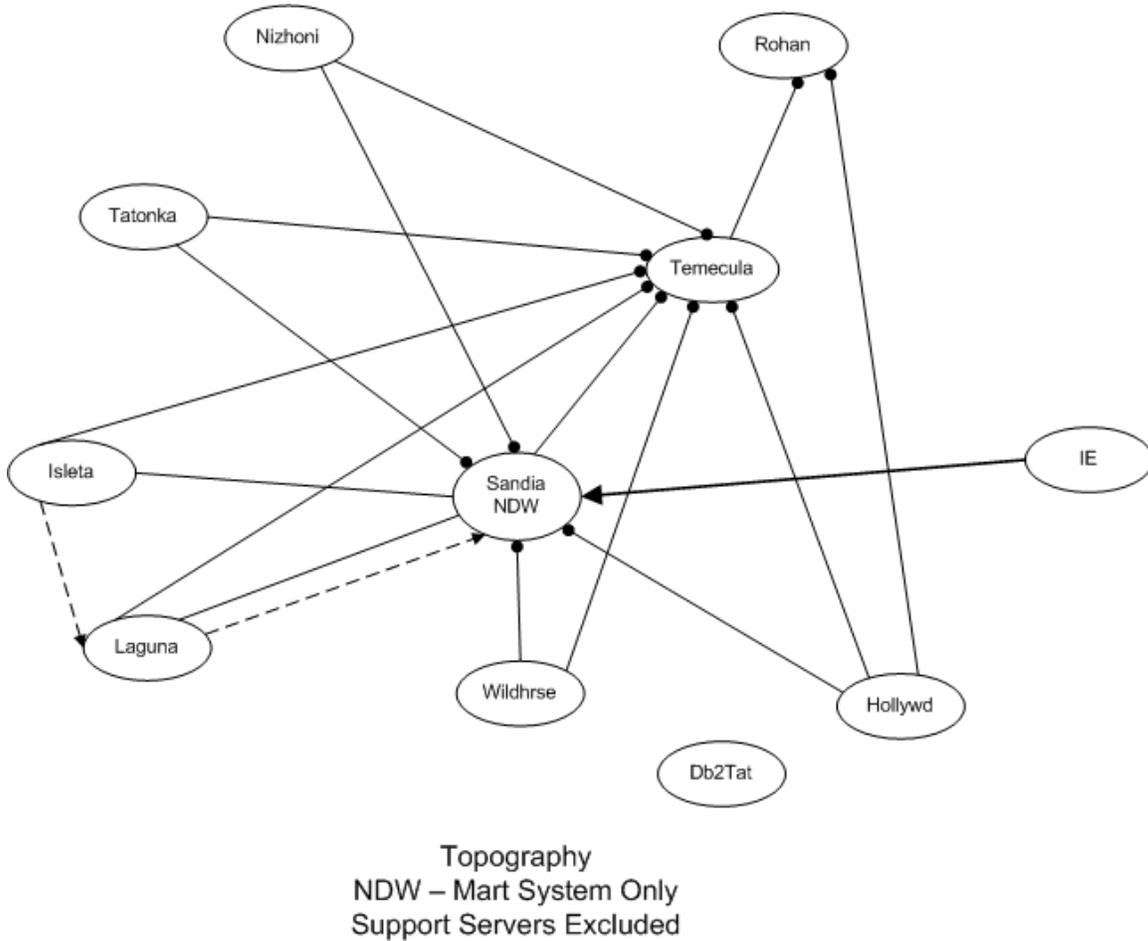


Figure 7. NDW/Data Mart System Topography

## Test and Quality Assurance Environments

Two additional database environments, Isleta/Caesar and Laguna, support changes to the NDW database or associated applications. These two environments are used for development (ISLETA and CAESAR databases) and QA (LAGUNA database). Both environments reside on a server (ARWEN) separate from the NDW environment and are occasionally refreshed from the NDW or from a predetermined baseline on an as needed basis.

These environments are typically kept in synchronization with NDW in regards to structure, except when testing structure changes, as illustrated in the following figure:

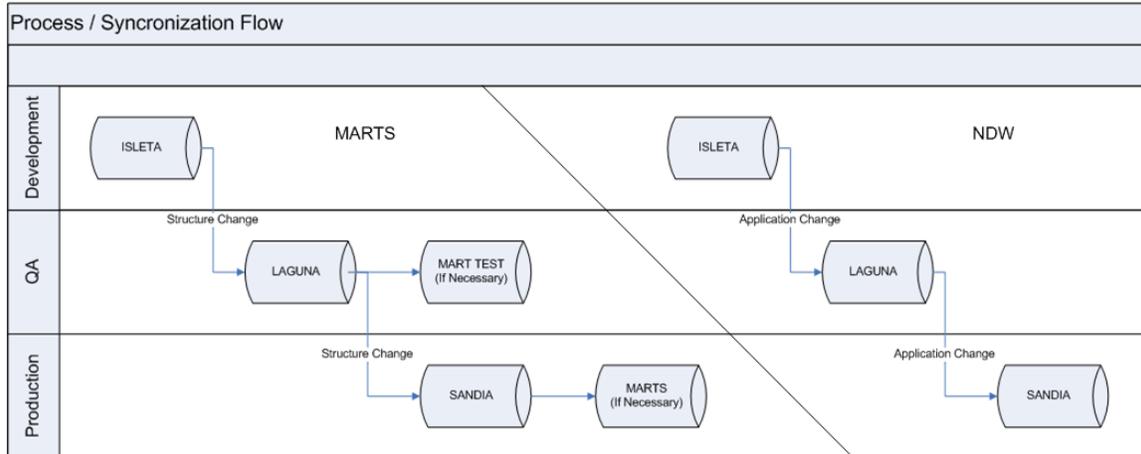


Figure 8. Process/Synchronization Flow

All changes, including application changes, are tested first in the development (Isleta) database environment and then in the QA (Laguna) database environment prior to production release and movement (promotion) to production in NDW.

## Tools

The NDW utilizes a number of tools used for management or operation. These include:

- IBM High Performance Expert (HPE)
- IBM High Performance Unload (HPU)
- Dataflux (In development/QA)
- Rational Data Architect (DA)
- Rational Software Architect (SA)
- Crystal Reports
- Business Objects

## Controls

There are several controlling documents that control the operation, modification, and data flow of the NDW. Among these are:

- NPIRS Accepted Practices (NAPS)
- NPIRS Operating Procedures (NOPS)
- IHS and HHS SOPs
- National Institute of Standards and Technology (NIST):  
<http://csrc.nist.gov/publications/nistpubs/800-53-Rev2/sp800-53-rev2-final.pdf>

## Emergency Management Plan (EMP)

The NDW is identified as a mission critical database in the EMP. A copy of the latest version of the *Emergency Management Plan (EMP) for the National Patient Information Reporting System (NPIRS) SOP-09-01i* document is available at the following location to authorized individuals:

P:\NPIRS\11. Security\Standard Operating Procedures\Emergency Management Plan\Current\

Each critical data mart/database within the NDW environment is referenced with the EMP plan.

An abridged version is available for review on the NDW Informational web site:

<http://www.ihs.gov/CIO/DataQuality/warehouse/what-if-I-have-other-questions.asp>

The abridged version contains a general description of the EMP plan sans proprietary information.

## On-Going Plans

At the time of the release of this document several changes are under way. Among these are:

- 1) Re-tasking of servers and movement of databases. The following diagram shows the proposed plan, which will maximize the government's investment in current storage and servers while improving risk mitigation.

Planned Server Configuration for IHS Data Warehouse and DataMarts

Second Quarter, 2009

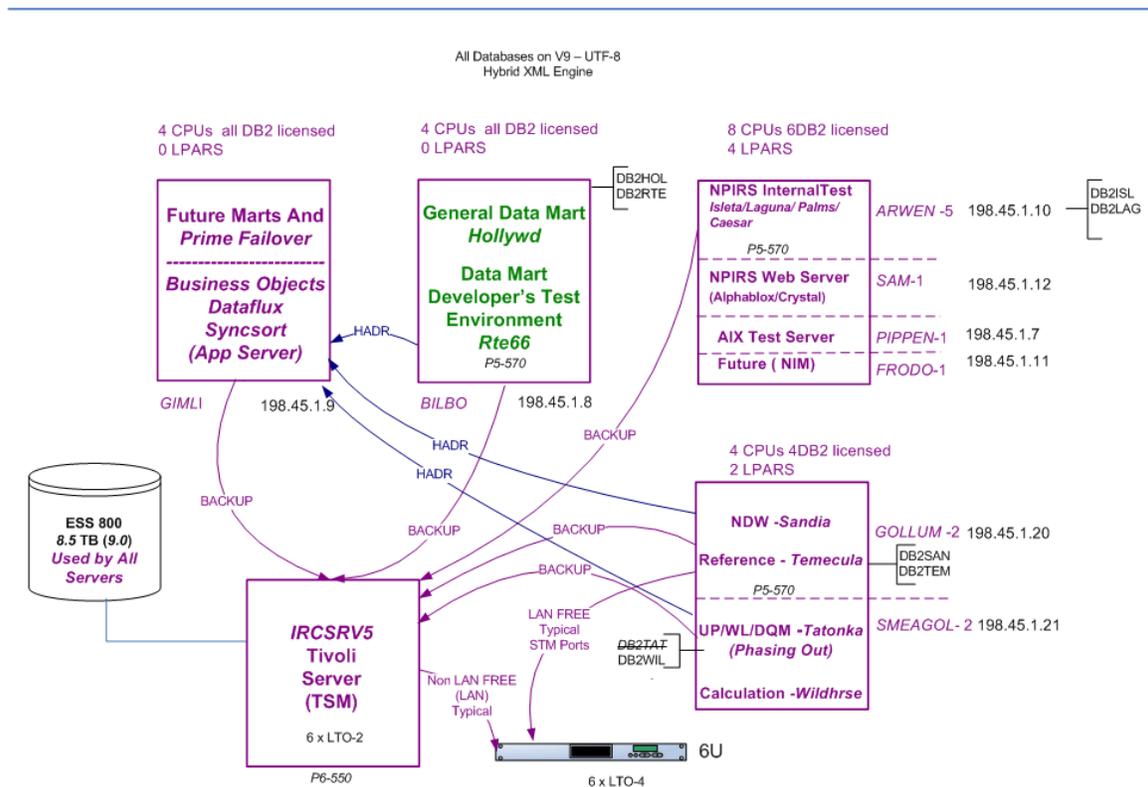


Figure 9. Proposed IHS Data Warehouse/Data Marts Server Configuration

- 2) More tightly managed applications using both Rational and Workbench. This will improve the EAI index and the manageability and flexibility of the system to satisfy the government's current and future needs.
- 3) Under DB2 V9, a robust Meta data layer that is largely automated will manage more of the ETL processes, eliminating some scripting and also improving the flexibility of the NDW.
- 4) The Change Management paradigm is being enhanced in a top-down model which will contain the following elements:
  - a) Requirements are derived.
  - b) Models are tested and evaluated against business processes and rules.
  - c) Changes in applications or DB structures are defined.
  - d) Documentation is updated.
  - e) Models are updated.
  - f) Changes are implemented.
- 6) Implementation of HACMP (High Availability Cluster Multi Processing) and DB2-HA (High Availability).

HACMP is the AIX driven capability of ensuring a fail-over in the event of a server or primary component failure. DB2-HA is the database component of the same feature. This will help ensure continuous run capability in the event of most hardware failures.
- 7) Other hardware and software improvements are planned to ensure reliability through redundancy, and improve overall performance of the entire NDW Enterprise.
- 8) Implementation of AIX and DB2 workload management. This will allow better control of IHS resources across multiple data bases and servers. More information can be found in the *Workload Management White Paper VI.0*.

## Appendix A: Weekly ETL Process

### Encounter Weekly ETL

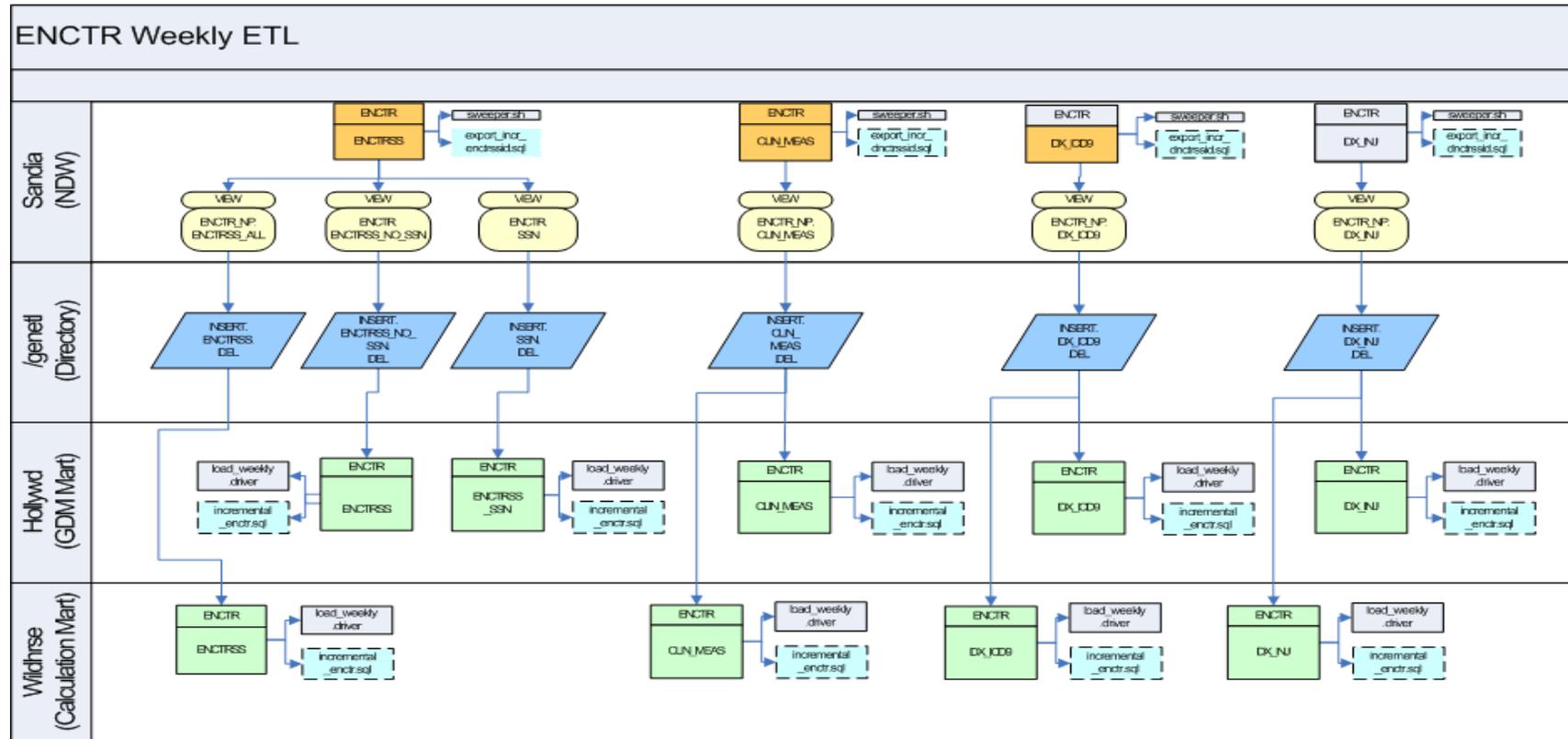


Figure 10. This diagram displays the ENCTR weekly ETL

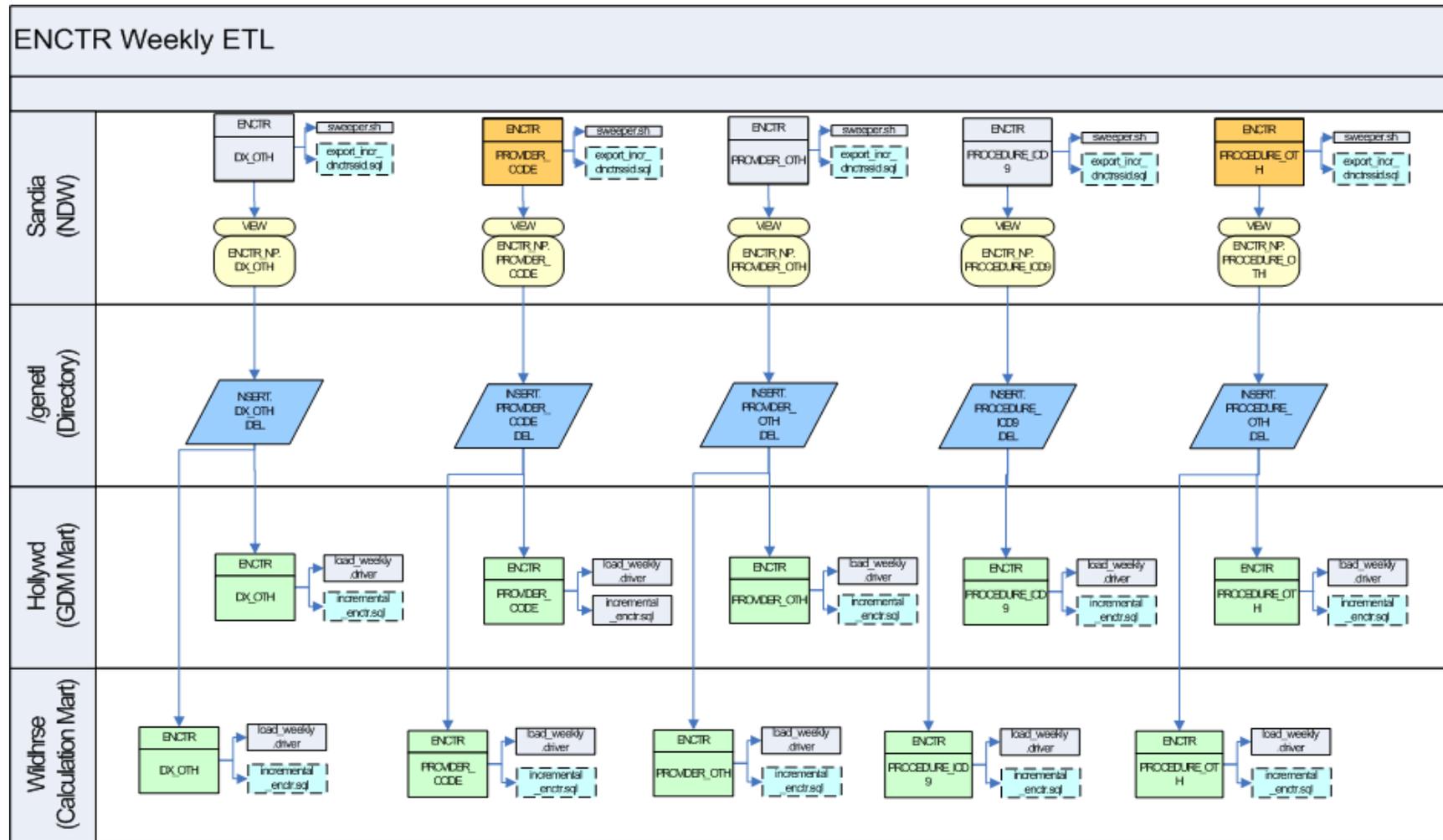


Figure 11. This diagram displays the ENCTR weekly ETL

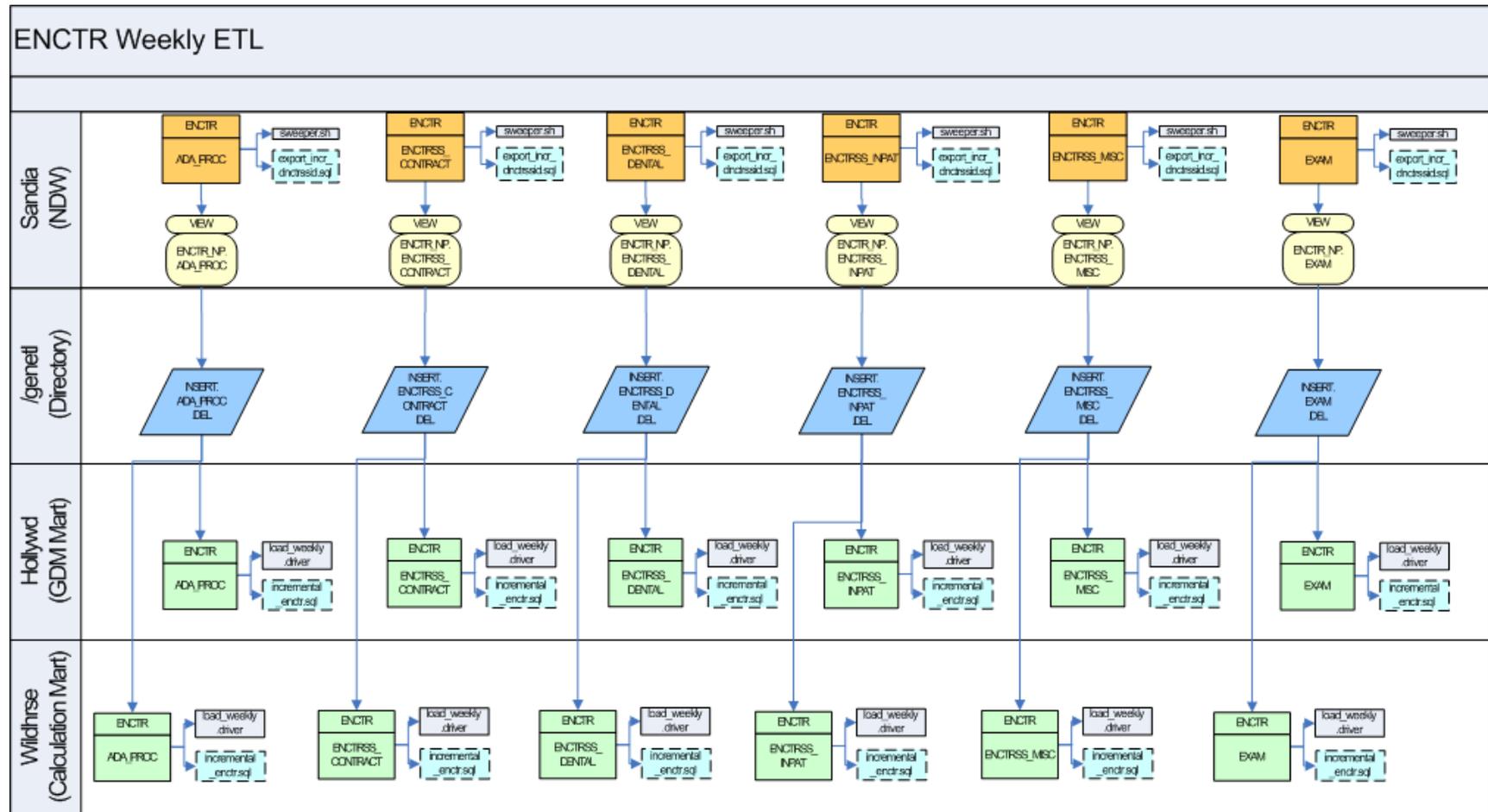


Figure 12. This diagram displays the ENCTR weekly ETL

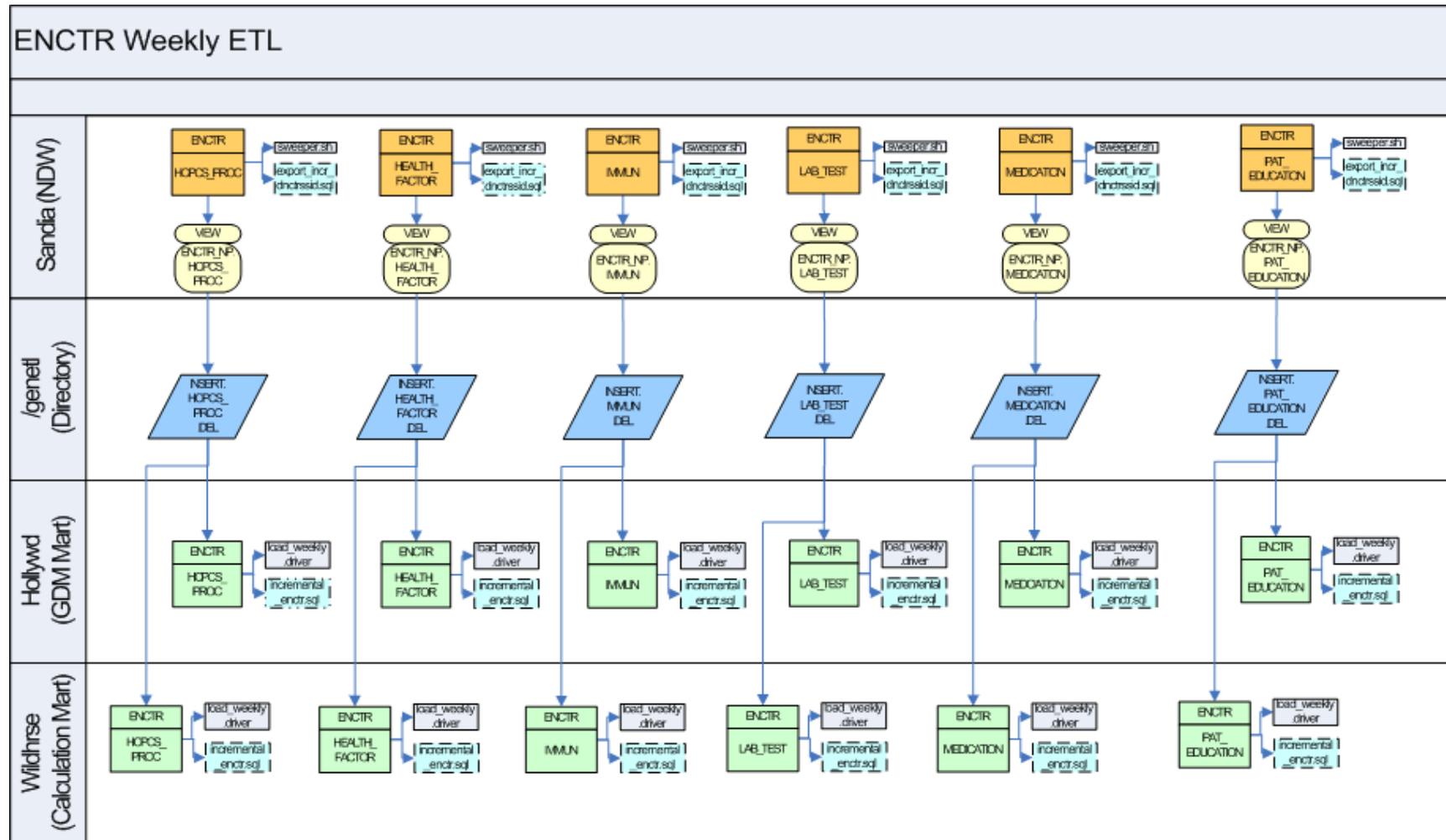


Figure 13. This diagram displays the ENCTR weekly ETL

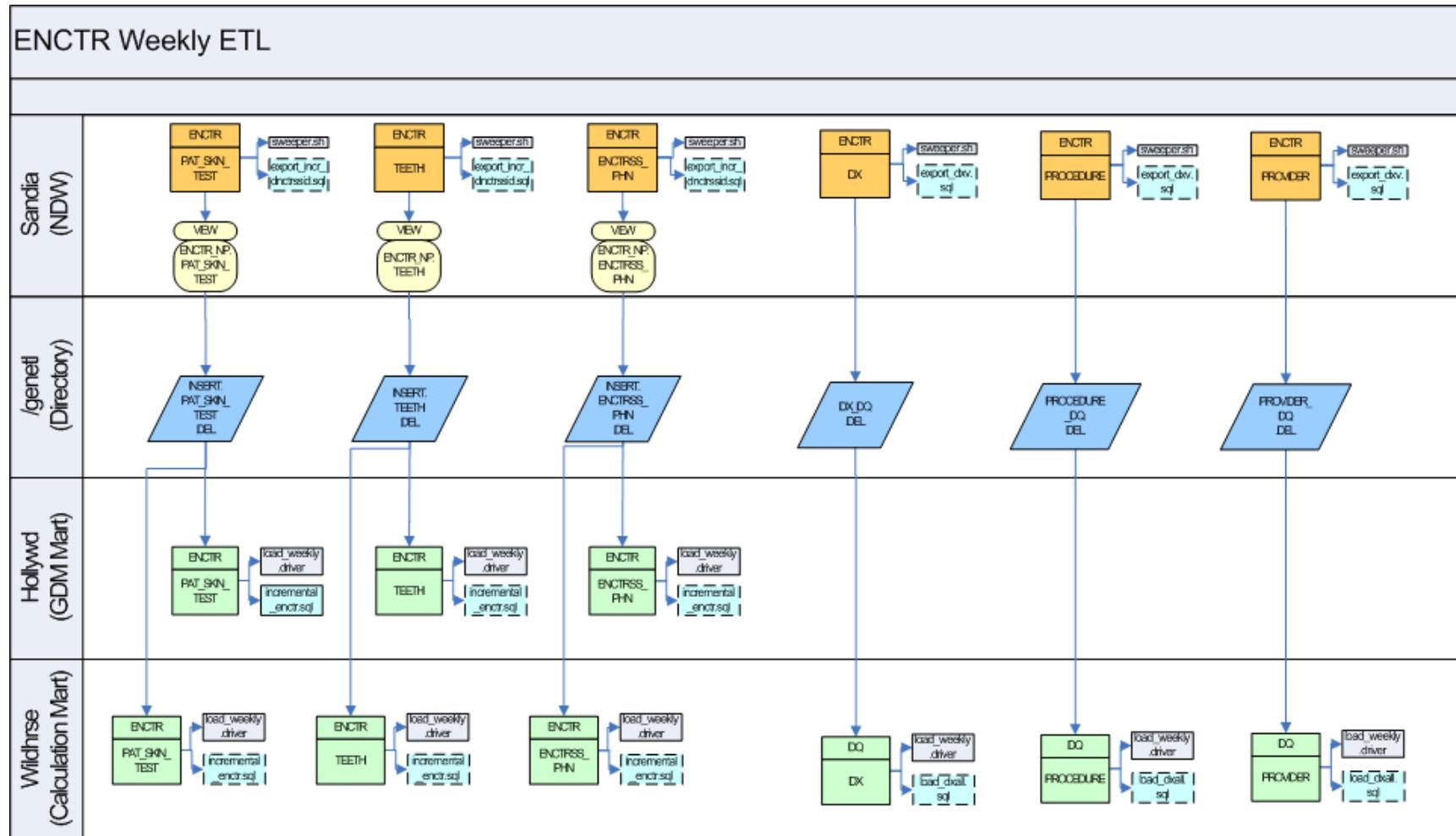


Figure 14. This diagram displays the ENCTR weekly ETL

## REG Weekly ETL

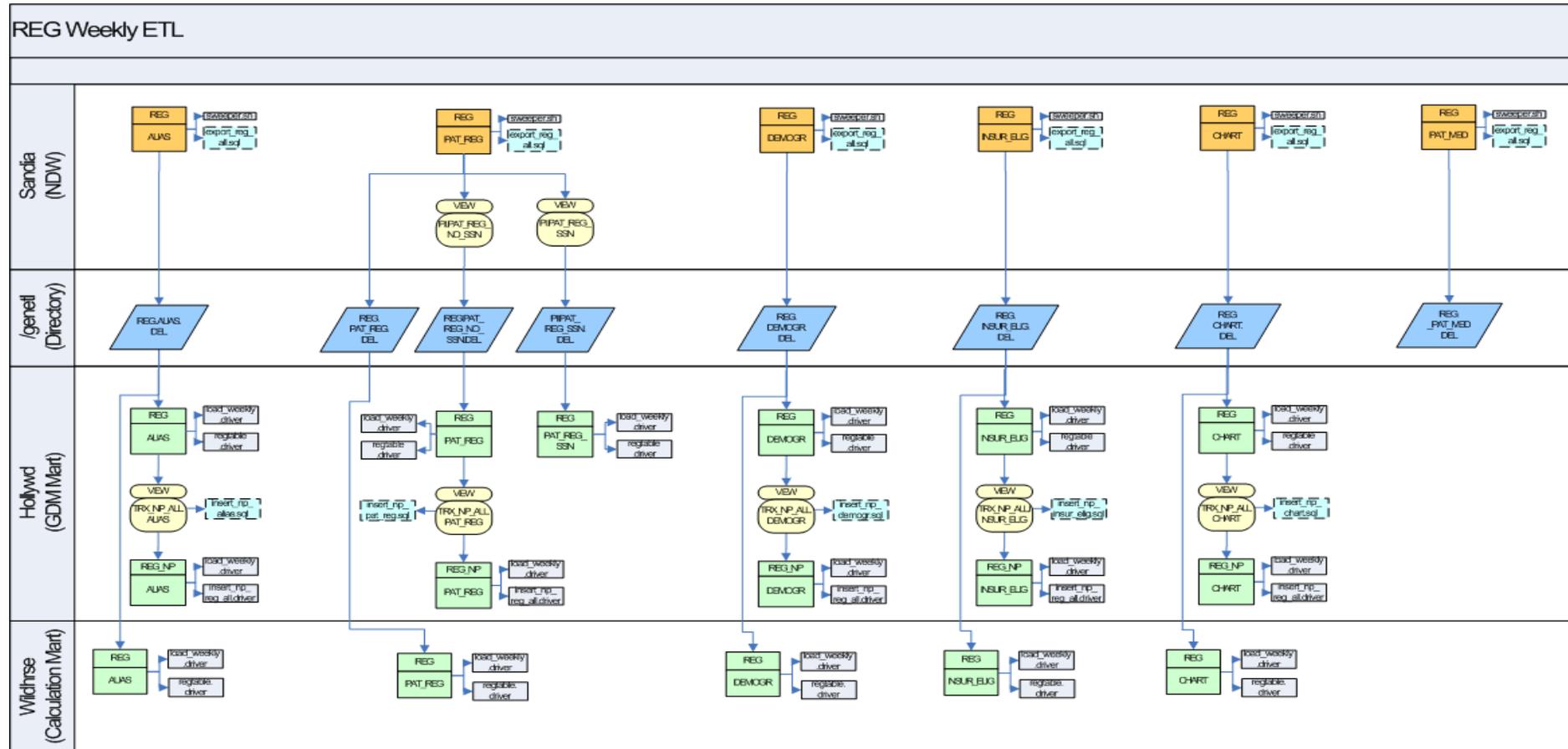


Figure 15. This diagram displays the REG weekly ETL

## ENCTR\_HIST Weekly ETL

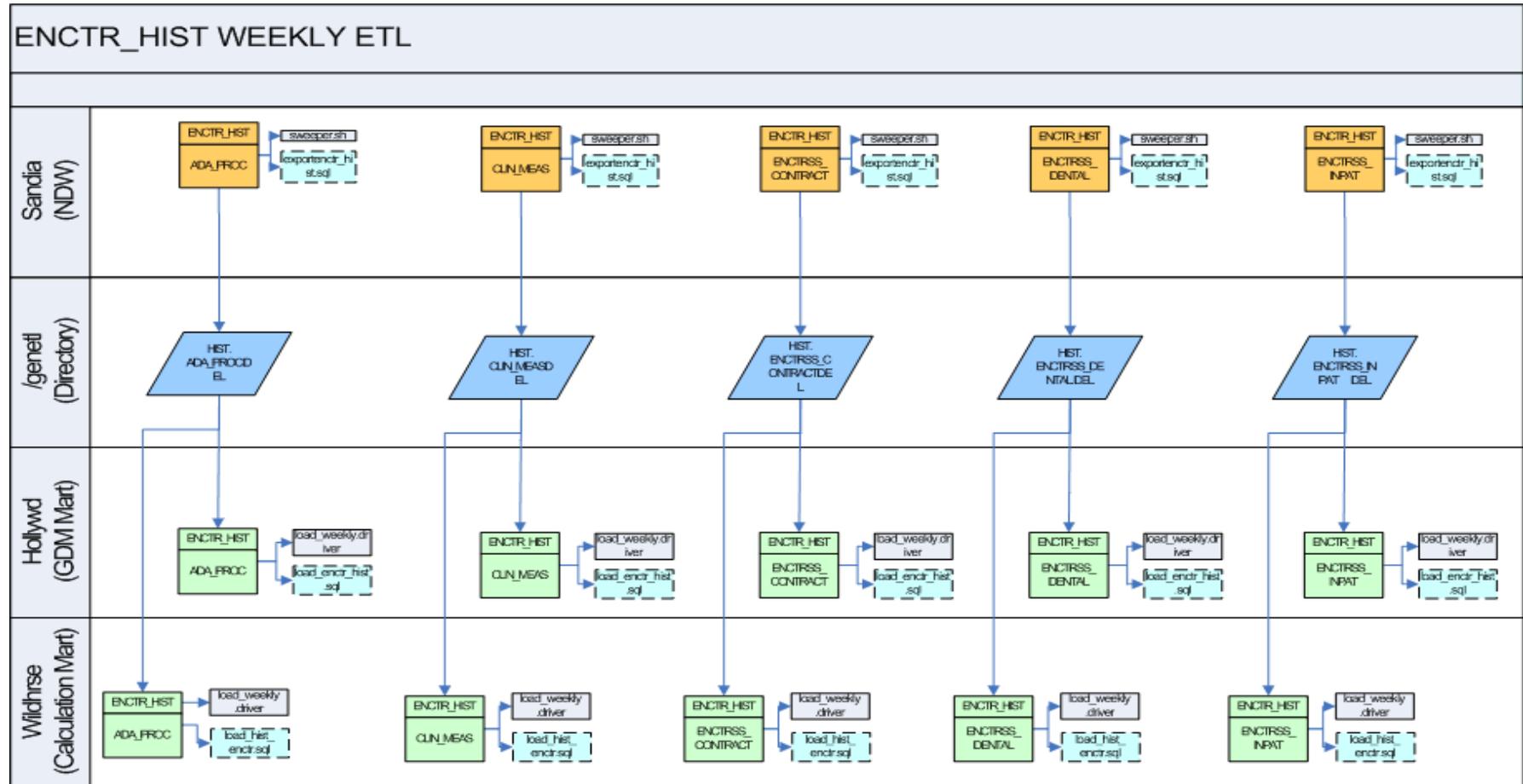


Figure 16. This diagram displays the ENCTR\_HIST weekly ETL

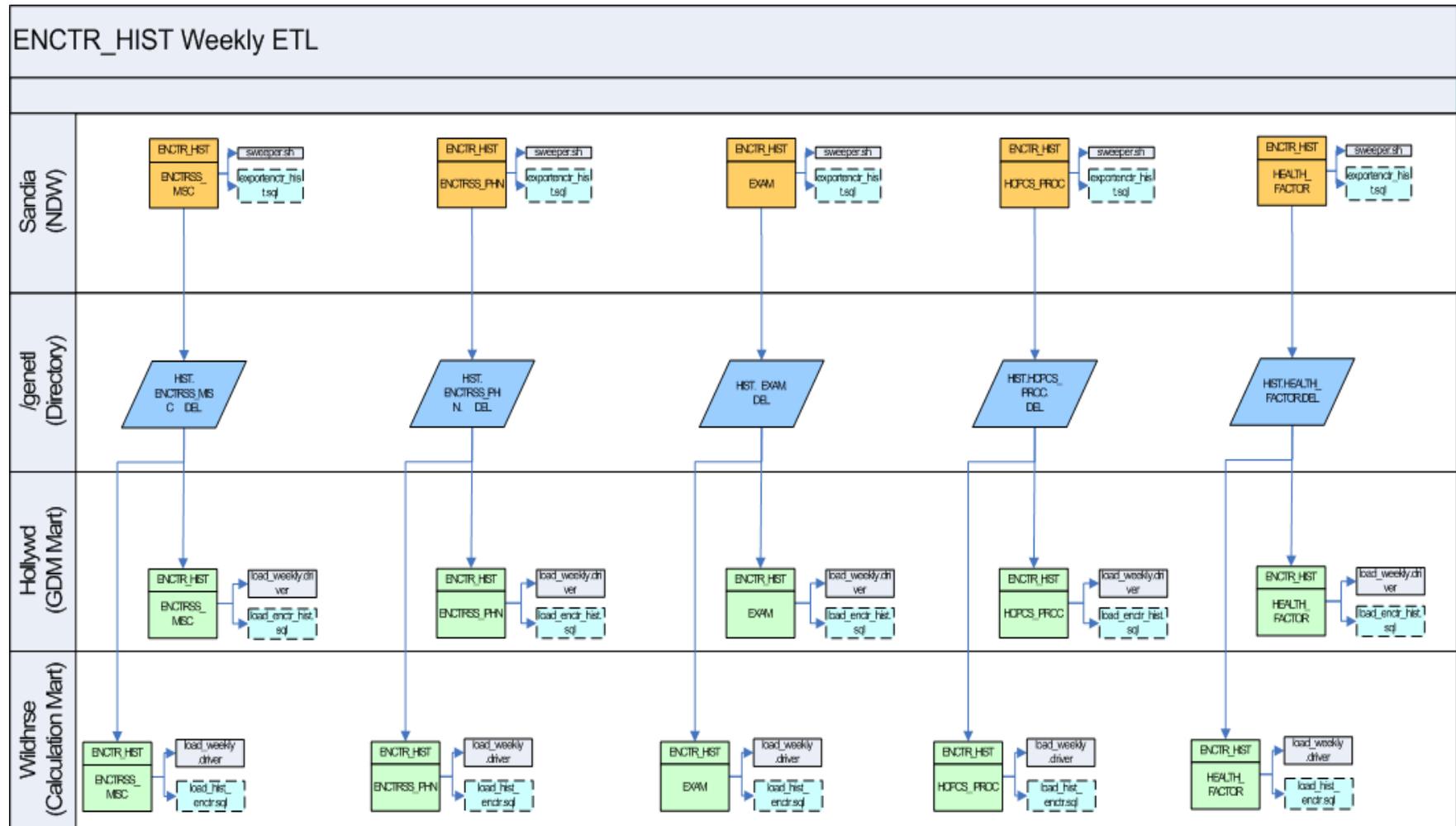


Figure 17. This diagram displays the ENCTR\_HIST weekly ETL

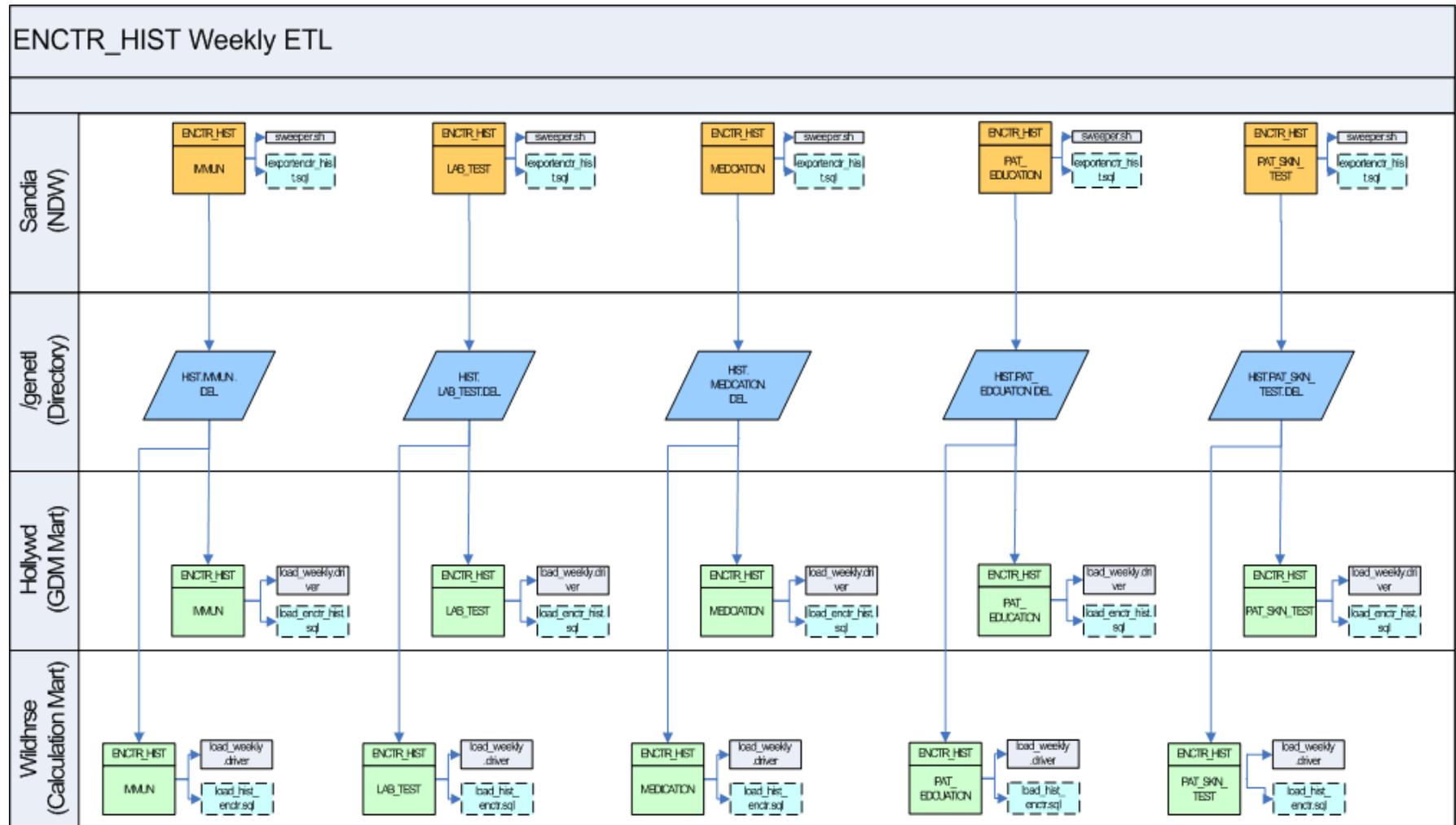


Figure 18. This diagram displays the ENCTR\_HIST weekly ETL

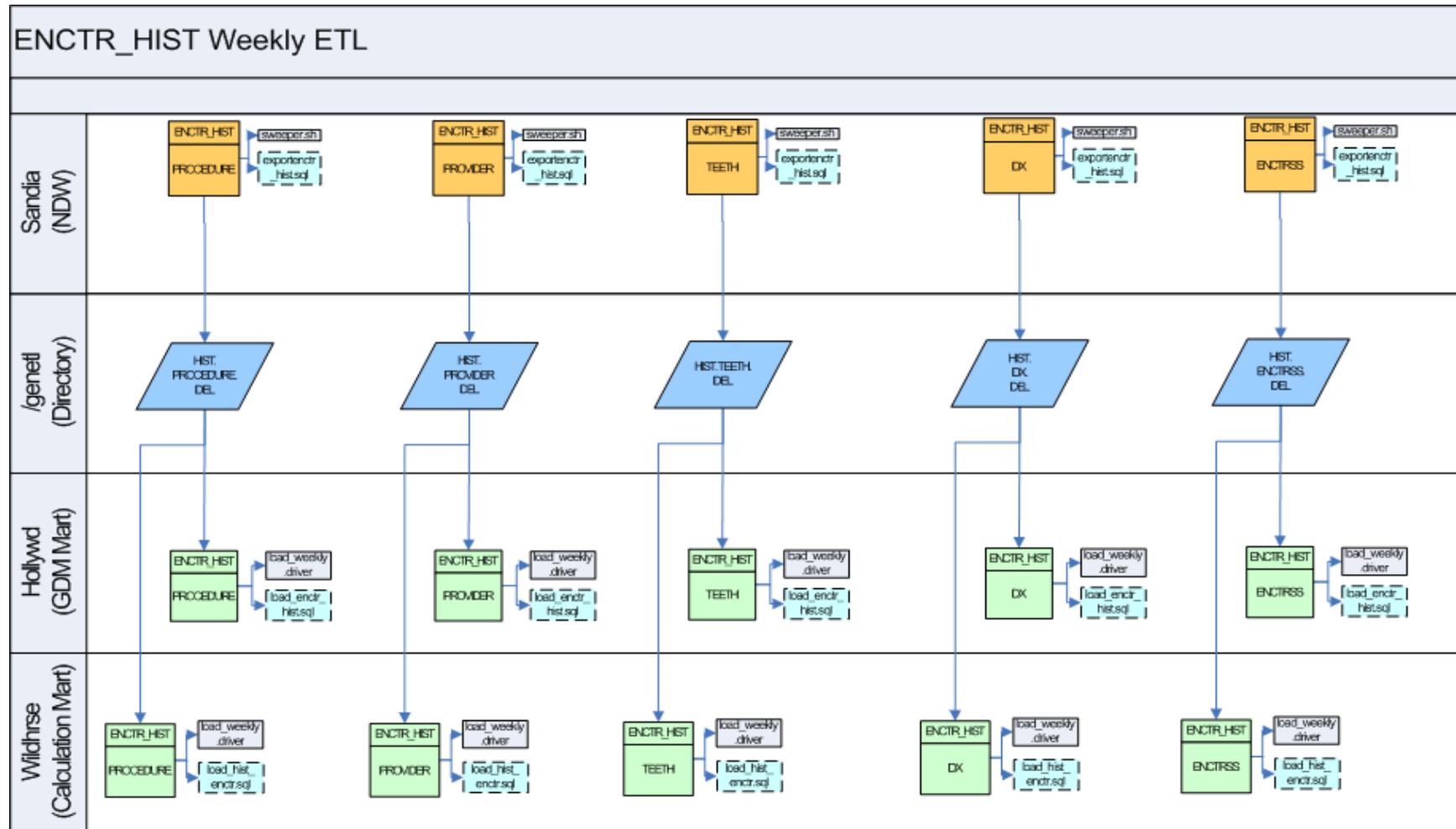


Figure 19. This diagram displays the ENCTR\_HIST weekly ETL

## MatchMaker Weekly ETL

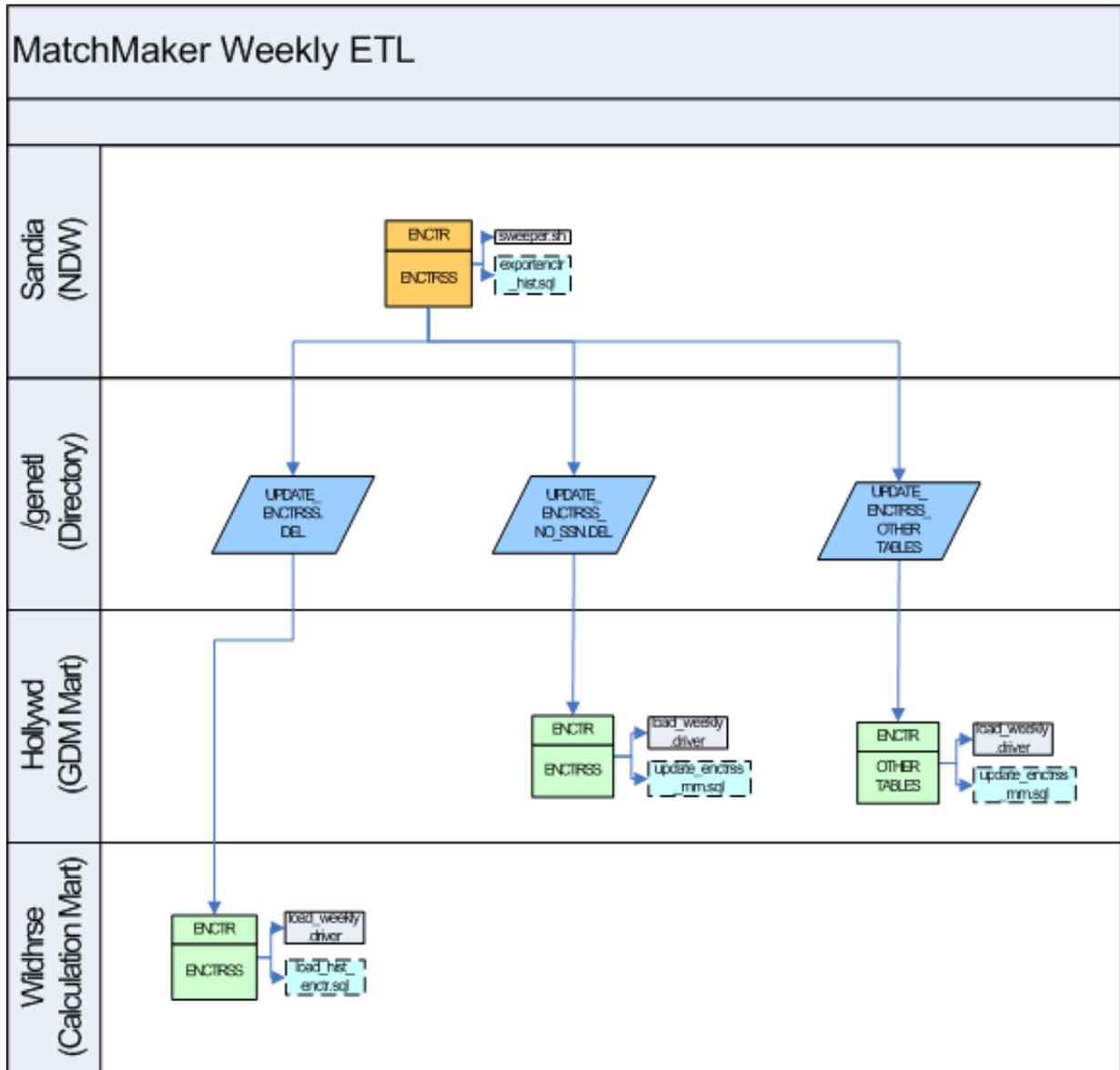


Figure 20. This diagram displays the MatchMaker weekly ETL

## ADMIN\_INFO Weekly ETL

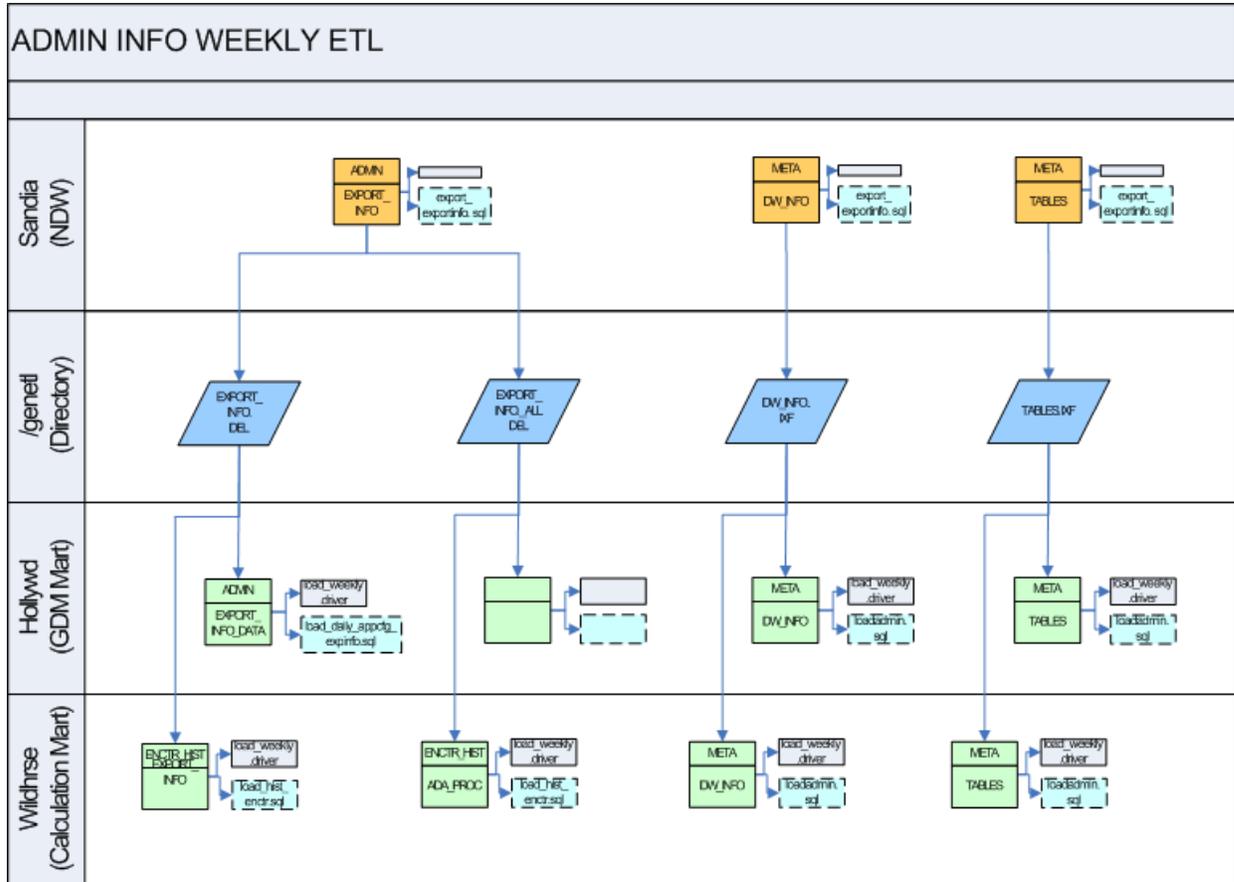


Figure 21. This diagram displays the ADMIN\_INFO weekly ETL