National Patient Information Reporting System: National Data Warehouse

NDW Production Database Technical Guide

Current Version:7.0

Created: 9/16/2011 1:39:00 PM
Published:2/14/2012
Author: NPIRS/NDW

Department of Health and Human Services

Indian Health Service

Office of Information Technology (OIT)

Contents

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 Metadata website, 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
4.0	November 2010	Updates to diagrams; added tools and descriptions; reordered several sections to improve readability; merged several sections to eliminate redundancy; added sample ACK and IMP Acknowledgement Email Messages and corresponding description table; added History_Sweeper information; added Federation details, updated server/database table. New HACMP/HA, DI, and CMP sections. New Data Integrity Process Flow diagram.
4.1	January 2011	Added new diagram to Data Integrity section; renumbered all Figures accordingly; regenerated TOC
5.0	March 2011	COTR approved March 21, 2011
5.1	April 2011	Updated with new diagrams since completion of data center relocation to BIA.
6.0	June 2011	Annual Review/Update complete.
7.0	February 2012	Updated with modified diagrams of Server Racks, etc. (Appendix A)

Overview

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

Three main databases:

(1) SANDIA

The transactional database, the source for most data mart extracts. This is also the repository of the Export Tracking Mart (information about the status of a sending site's export files which is generated when those files are processed into the NDW).

(2) WILDHRSE

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, User Population/Workload Mart.

(3) TEMECULA

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

Supplemental databases:

(1) HOLLYWD

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

(2) ISLETA and LAGUNA

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

The repository for the Data Integrity Mart which allows for comparisons between data stored and files received to ensure accuracy.

(4) RTE66

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 extract, transform, and load (ETL) processes from a sample NDW database to their data mart, prior to their implementation in the production environment. RTEE66 is currently being used as the LDI database pending design changes to another location.

(4) OLDLEGCY

Contains legacy data from the old NPIRS (Pre-2006) data that is generally available through federation to WILDHRSE.

Constructs

The NDW utilizes a number of constructs 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.

The main construct types associated with the NDW include:

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

Information on these constructs can be found in the *NDW Data Mart DB2 Glossary* located on the NDW Informational website.

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

Tools

The NDW utilizes a number of tools used for management or operation. These tools and a brief description of their purpose appear below.

TOOL	PURPOSE
IBM High Performance Expert	Consolidates, analyzes, and recommends changes on DB2 performance-related information.
IBM High Performance Unload	Unloads and extracts data for movement across enterprise systems.
IBM InfoSphere QualityStage	Performs survivorship (un-duplication) as well as metadata analysis, data profiling, data monitoring, data enrichment and entity resolution.
Rational Data Architect	Supports the management, impact analysis, search, and reporting across all forms of metadata.
Rational Software Architect	Provides a modeling and development environment that uses the Unified Modeling Language (UML) for designing architecture for C++ and Java 2 Enterprise Edition (J2EE) applications and web services
Rational Clear Case	Provides a software configuration management tool (SCM) as well as compilation, deployment and repository.
SAP Business Objects	Complete set of Business Intelligence is the primary web support for reporting and provides additional tools and applications which support reporting, data querying and data analysis, interactive data visualization, data management, planning, budgeting and forecasting.
SAP Business Objects Metadata Manager	Consolidates and audits metadata from disparate tools and sources, including business intelligence (BI). Extracts, transforms, and loads (ETL) relational databases, modeling tools, and third-party metadata into a single repository.

SAP Business Objects Universe	Allows users to create, modify, deploy and read definitions of the Business Objects stored in the Business Objects repository. This is the primary source for Workload and Userpop reporting.
SAP Business Objects Intelligence	Facilitates reporting, analysis, visualization, and exploration of data to provide immediate answers to business questions.
SAP Crystal Reports	Reporting application which allows users to design and generate reports from a selected data sources.
Quest Central	Provides monitoring, database administration, SQL tuning, and space management for DB2.
Quest Toad	Comprehensive database and SQL development and administration tool.
RazorSQL	A DB2 database browser, editor, and administration tool.
Edit Rocket	A source code editor, text editor, and software development tool.
X-Win32	An X-windows software utilized for software installation and updates on the AIX servers.
Uptime	Simplifies the management of virtual, physical and cloud environments; monitors services, applications, servers, and platforms from a unified dashboard.
SharePoint 2010	Collaboration, file sharing, and web publishing tools.
VMWare Vsphere	Provides virtualization capability to the Web and Winserver portion of the NDW allowing rapid response to user reporting requirements.
Rational Team Concert	Provides both a repository and deployment tool for code.

Reference Documents

There are several external documents that further detail 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) Recommended Security Controls for Federal Information Systems and NIST's Federal Information Security Management Act (FISMA) processing standards and guidelines*

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

Design Parameters

- The NDW environment is a source system and data repository for the data marts. Some non-NPIRS data marts use other sources. The NDW can be modified to accommodate future business needs.
- 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 XIV grid 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 24/5, except for maintenance. The Import Engine process (IMP) is stopped during ETL processes to ensure synchronization and integrity of data.

System Environment

Specifications related to the physical environment of the NDW main database appear below:

Server	FTP Address	Operating System Name/Version	Database	Instance Name	Partition Level	DB2 Level (where applicagle)	Comments
Gollum	198.45.1.31	AIX 6.1+	Sandia	DB2SAN	4	V9.7.x	Primary NDW DB2 database. Also the repository of the Export Tracking Mart.
Gollum	198.45.1.31	AIX 6.1+	Nizhoni	DB2NIZ	1	V9.7.x	Data Integrity Data Mart
Smeagol	198.45.1.33	AIX 6.1+	Temecula	DB2TEM	2	V9.7.x	Reference and Meta table database.
			Wildhrse	DB2WIL	4	V9.7.x	Calculation, reporting and research database containing the Data Quality, Userpop, and Workload Data Marts.
			Tatonka	DB2TAT	2	V9.7.x	Change Control Database
			QSMGR	DB2QSMGR	1	V9.7.x	Quality Stage Database
			Oldlgcy	DB2OLD	1	V9.7.x	Legacy Data (Pre-NDW) (Static) 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.
Bilbo	198.45.1.38	AIX 6.1+	Hollywd	DB2HOL	2	V9.7.x	General Data Mart DB; volatile structure based on User demands. Also contains a copy of the Legacy Data Mart.

Server	FTP Address	Operating System Name/Version	Database	Instance Name	Partition Level	DB2 Level (where applicagle)	Comments
			Rte66	DB2RTE	2	V9.7.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.
Arwen	198.45.1.10	AIX 6.1+	Isleta	DB2ISL	2	V9.7.x	Primary data warehouse development environment. Small mirror image of NDW primary; can be refreshed or restored to baseline. Used for unit testing.
			Laguna	DB2LAG	2	V9.7.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 and file test Data Mart.
Rohan	198.45.1.39	Windows Server 2003	N/A	N/A	N/A	N/A	The Business Objects web reports server.
Shire	198.45.1.52	Windows Server 2008 R2	N/A	N/A	N/A	N/A	The Windows SharePoint Services server for both the 3.0 and 2010 versions.
MidlErth	198.45.1.54	Vsphere 4.1	N/A	N/A	N/A	N/A	For VMware Vsphere 4.1
Fanghorn	198.45.1.51	Windows Server 2008 R2	N/A	N/A	N/A	N/A	For SQL Server 2008 STD.

Server	FTP Address	Operating System Name/Version	Database	Instance Name	Partition Level	DB2 Level (where applicagle)	Comments
Mordor	198.45.1.41	Windows Server 2003	N/A	N/A	N/A	N/A	The web development server.
IRCSRV6	198.45.1.42	6.1+	TSM	N/A	1	N/A	The Tivoli Storage Manager Server Version 6.2.

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 Metadata internet website: http://www.ihs.gov/scb/metadata/.

Server Configuration

The NDW exists primarily on disparate AIX servers with connection to Windows Servers for I/O interfaces and reporting. Internal Logical Partitions (LPAR) separate discrete servers within the main server. Additional connections within the IHS core or via a VPN connection into the core are also supported. The inter-relationship of the IHS National Data Warehouse, data marts, and associated hardware is shown below.

The Power HA high availability failover configuration (explained in the "NDW Architecture" section of this document) is shown in the figure below, followed by a detailed Fibre Switch diagram for the McData and Brocade devices. A text-only version of the NPIRS Fibre Switch mapping appears in Appendix C.

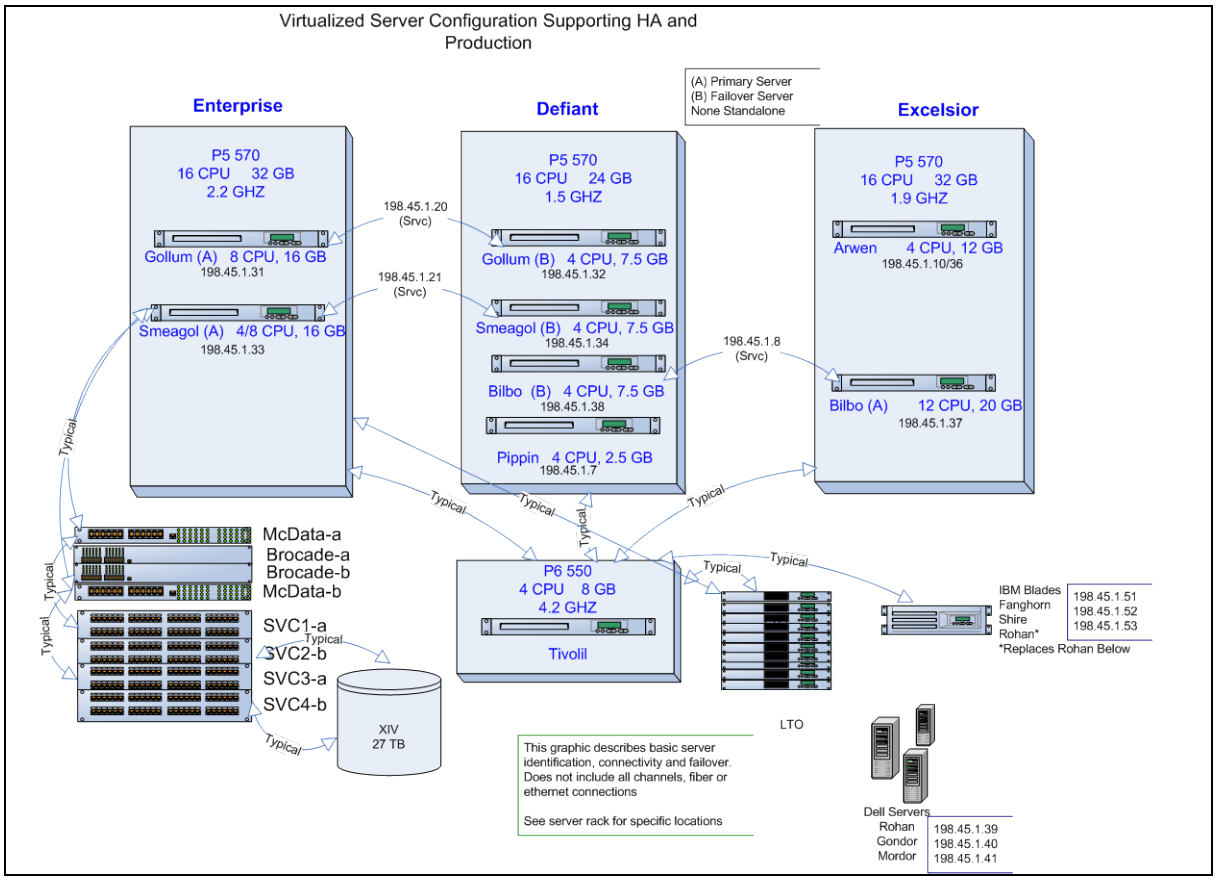


Figure 1 - Current IHS Data Warehouse/Data Marts Server Configuration

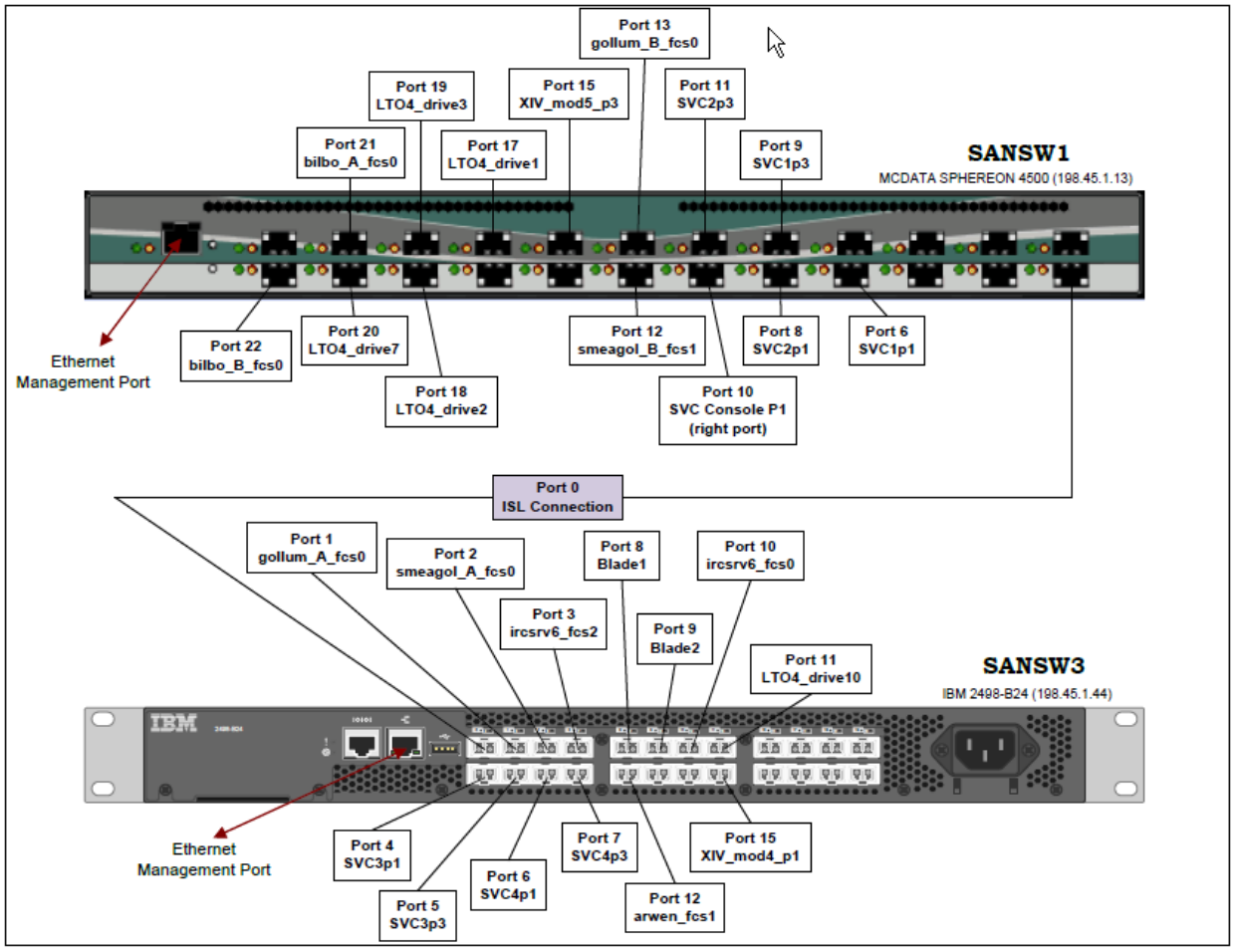


Figure 2 – FABRIC "A" Fibre Switch Diagram

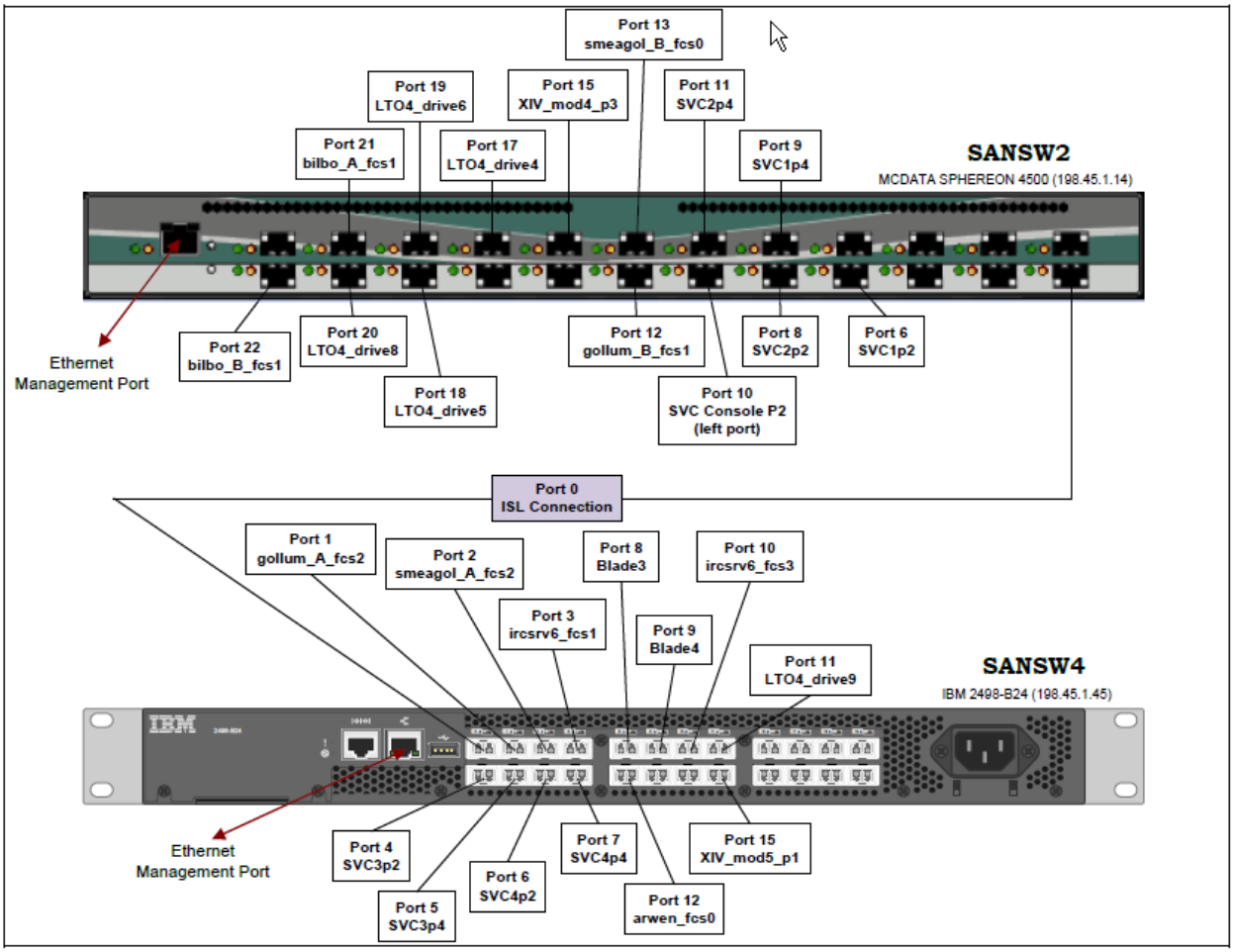


Figure 3 – FABRIC "B" Fibre Switch Diagram

Data Integrity Environment

The Data Integrity (DI) is a mandated process that verifies the data received is the data stored. The NDW does no cleaning or verification of data, but stores the data received. There are minor transformations that occur. The DI process operates with the files received from the Integration Engine (IE) and compares the data contained in these file to the data stored in the NDW.

This process occurs on a periodic basis and utilizes both the SANDIA and randomized input data files (as the sources) and NIZHONI database (as the destination) where they are compared using a tertiary process. The process loads data files into staging tables with regard to any business rule and in a what-you-see-is-what-you-get format where it then compared to the NDW target tables.

Change Management Process Environment

The Change Management Process (CMP) database is utilized to track changes to the NPIRS system. The change management database is in the TATONKA database on the SMEAGOL server. The CMP schema contains the data tables, and the CMP_REF schema contains the reference/validation tables that are used for various columns within the CMP tables.

Details on the Change Management Process Database, including its data dictionary, are contained in "*NAP_7_Change Management Process Database User Guide_V1.0"*. Details on the Change Management Process as a whole can be found in the document entitled "*NAP_4_Change Management Plan"*.

Test and Quality Assurance Environments

Two additional database environments, ISLETA and LAGUNA, support changes to the NDW database or associated applications. These two environments are used for development (ISLETA database) 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. The QA environment is also used to test data files from sources for completeness and adherence to import guidelines. LAGUNA also serves a secondary QA role in testing input files from new sending sites to ensure the compliance of the sending site to the transmission standards established for the NDW.

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

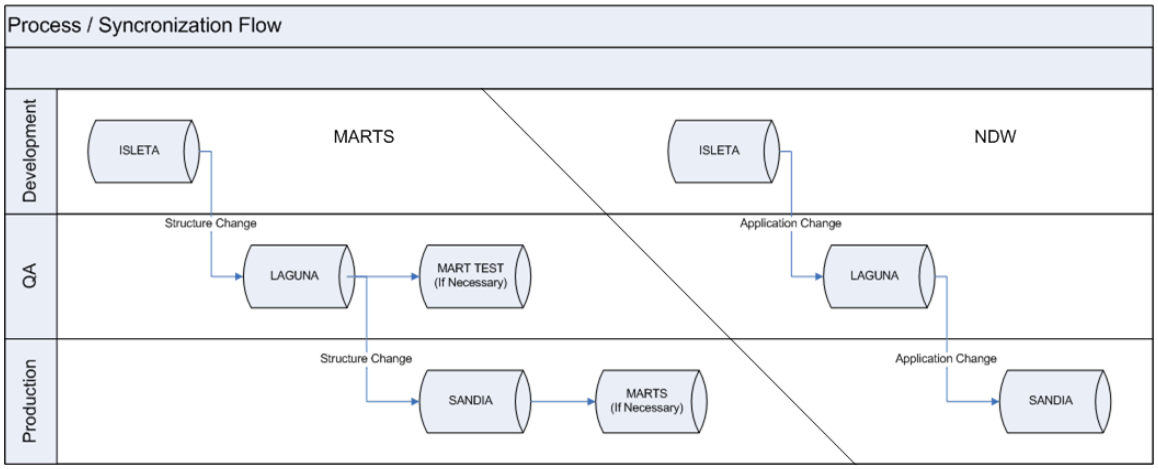


Figure 4 - Process/Synchronization Flow

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.

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 and direct.

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 controlled by NPIRS in read-only mode. It should be noted that security on 3rd party marts are typically controlled by MOU between the mart owner and IHS. Updates are restricted as follows:
    - TEMECULA is updatable for tables associated with the Standard Code Book (SCB) only by users authorized to make such changes. The SCB contains vital code set information for exporting data to the NDW. 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 and/or MOU.
    - Individual 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 levels, role based security and Label Based Access Control (LBAC). LBAC is currently implemented only on NPIRS controlled external marts such as the General Data Mart. Additional information on LBAC can be found in "*NPIRS Accepted Practices (NAP) 9. LBAC Administration Practices for NPIRS General Data Mart"*.
    - Additional security will be added throughout the enterprise including LDAP and more extensive use of LBAC.

NDW Architecture

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 to support future enhancement and capabilities in anticipation of the growing capability of XML/XQuery.

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 metadata 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 Fix Pack 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’ underutilized resources from other servers during periods of heavy demand.

Compression

Both column and row compression is utilized in the NDW. The compression algorithms maximize utilization of the mass storage paradigm while maintaining, or optimizing, retrieval I/O. Detailed information on Compression can be found in the *NDW Data Mart DB2 Glossary* located on the NDW Informational website.

Federation

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.

Due to the dynamic nature of the data in the NDW, and the frequency of both manual and automated updates to the REF tables on TEMECULA, federation is necessary to keep the data between the NDW and all of its data marts synchronized. Within the NDW, the federation process is applied to the ADMIN.EXPORT_INFO table (in SANDIA) and all of the REF tables (in TEMECULA). Currently the federation scripts are scheduled to run daily on HOLLYWD, and five times per week on both SANDIA and WILDHRSE. Only Reference Tables are loaded by federated nicknames (a local name for a remote object in a federated system) from TEMECULA.

All REF tables in TEMECULA are updated in the WILDHRSE and General Data Mart environments, while only the following tables are updated in SANDIA: AREA, CLINIC, ERROR, FACILITY, INJURY_PLACE_CD_XREF, PROVIDER_DISC, and REGION. The development and QA environments are updated on demand.

Detailed information on the use of federation for integrating heterogeneous sources of data can be found in the IBM Redbooks publication at http://www.redbooks.ibm.com/redbooks/pdfs/sg247032.pdf . More details on the federation process as it applies to the NDW can be found in the "*NDW Data Mart DB2 Glossary*" located on the NDW Informational website.

HACMP/HA/Redundancy

High Availability Cluster Multi-Processing (previously known as Power-HA) is the AIX driven capability of ensuring a fail-over in the event of a server or primary component failure. DB2-High Availability (HA) is the database component of the same feature. These components work together to ensure continuous run capability in the event of most hardware failures.

Three of the existing NDW servers are currently set up to fail-over to each other. The current IP addresses are set up as SERVICE IP addresses that also failover to each server, as shown below:

Main Server Name	Fail-Over Server Name	Service IP Address
SMEAGOL_A (ENTERPRISE)	SMEAGOL_B (DEFIANT)	198.45.1.21
GOLLUM_A (ENTERPRISE)	GOLLUM_B (DEFIANT)	198.45.1.20
BILBO_A (EXCELSIOR)	BILBO_B (DEFIANT)	198.45.1.8

All servers share access to the same hard disks in the XIV Storage Subsystem (ORCA) through the SAN Volume Controllers (SVC). Shared hard disks are used as “heartbeat” disks between the servers (i.e., as long as a regular "pulse" or "heartbeat" continues between the main server and the second fail-over server, the second server will not initiate its systems). The second server will immediately take over the work of the first as soon as it detects an alteration in the "heartbeat" of the first machine. In addition, the DB2 databases are set up as shared applications that failover to each server. The service address is the address used by the users and applications vs. the actual IP address of the server.

Further redundancy is achieved by utilizing dual-SAN fabrics between the server and the mass storage devices. This is achieved through two channels through discrete SAN fabrics requiring dual SAN and Fibre controllers. The mass storage device (XIV) utilizes GRID storage which further enhances redundancy and reliability.

Crystal Interface

The NDW utilizes Crystal Reports, a business intelligence application, to design and generate reports for the majority of its marts (Userpop/Workload Mart, Data Quality Mart, Export Tracking Mart, and Data Integrity Mart). The data is pulled from dedicated views or universes and published using Business Objects Web application on the IHS National Data Warehouse web site in separate folders designated for each of these marts.

Data originates on the appropriate mart and the Crystal Report is run on a Winserver (ROHAN), distributed via a web page, and made accessible to the user either directly from the web or via export into multiple formats (.PDF, .XLS, .RTF).

NDW Processes

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 the Division of Information Technology Operations (DITO).

ACK Process

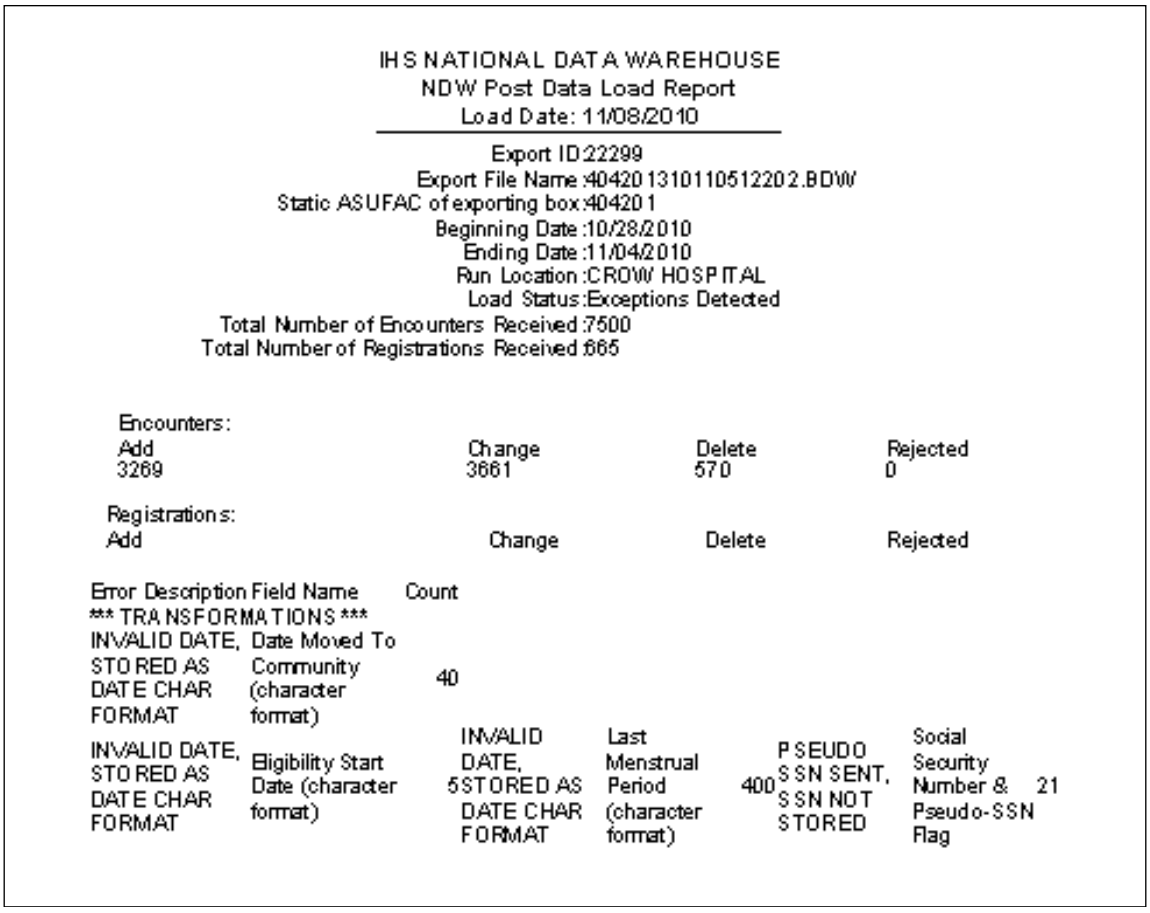
The Acknowledgement (ACK) process recognizes the file, logs the file into the NDW database, and sends an e-mail to the sending site that the file has been received. The ACK process is a "java wrapper" application that applies specific data and file rules, and utilizes embedded SQL.

```
                        * DATA WAREHOUSE *
                  NDW Data Transmission Log Report
                  Data Received Date: Nov 05, 2010

For more information please send an email to: NDW-ACKREPLY@ihs.gov

                                Export ID: 22299
                          Export file name: 40420131011051220 2.BDW
              Static ASUFAC of exporting box: 404201
                           Beginning Date: Oct 28, 2010
                              Ending Date: Nov 04, 2010
                        Received by IE Date: Nov 05, 2010
                               Run Location: CROW HOSPITAL
                        Transmission Status: No errors detected
       Total Number of Registrations Received: 665
          Total Number of Encounters Received: 7500
```

Figure 5- Sample ACK Transmission Email Message

IMP Process

The Import Engine process (IMP) subsequently loads the files into the database, where the data is available for transmission and loading to the various marts. A detailed acknowledgement is sent to the sending site, an example of which appears below. If applicable, detailed information on any errors that may have occurred during the IMP process will also be included. The IMP process is a java wrapper application that applies business rules and utilizes embedded SQL.



IHS NATIONAL DATA WAREHOUSE
NDW Post Data Load Report
Load Date: 11/08/2010

Export ID:22299
Export File Name:404201310110512202.BDW
Static ASUFAC of exporting box:404201
Beginning Date:10/28/2010
Ending Date:11/04/2010
Run Location:CROW HOSPITAL
Load Status:Exceptions Detected
Total Number of Encounters Received:7500
Total Number of Registrations Received:665

Encounters:

Add	Change	Delete	Rejected
3269	3661	570	0

Registrations:

Add	Change	Delete	Rejected

Error Description	Field Name	Count					
*** TRANSFORMATIONS ***							
INVALID DATE, STORED AS DATE CHAR FORMAT	Date Moved To Community (character format)	40					
INVALID DATE, STORED AS DATE CHAR FORMAT	Eligibility Start Date (character format)	5	INVALID DATE, STORED AS DATE CHAR FORMAT	Last Menstrual Period (character format)	400	PSEUDO SSN SENT, SSN NOT STORED	Social Security Number & Pseudo-SSN Flag 21

Figure 6- Sample Post Data Load Email Message

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.

The table below briefly describes the information contained in the Post Data Load report:

Report Item	Description
Load Date	The date the data export file was loaded via the process.
Export ID	The unique ID assigned by the NDW to this file.
Export file name	The file name assigned to the data export by the source system.
Static ASUFAC of exporting box	The 6-character code that identifies the site that sent the file.

Report Item	Description
Beginning Date	The Date Last Modified start date for this data export. **Note**: This date is determined by the default delay settings at the source data site.
Ending Date	This date represents the Date Last Modified ending date for this data export.
Run Location	The name of the site that sent the file.
Load Status	No errors detected (pass), Exceptions Detected, or Errors Detected in file (fail). A failure status can be • Record count discrepancy of 10 or more records • No records found in file • File was previously loaded • Area Code associated with the site cannot be determined • Too many sub-records of a specific type in a record. **Note:** NPIRS staff will contact the Site/Area for any failure other than “no records” or “previously loaded.”
Received by IE Date	The date the NDW Integration Engine received the file, where the file is checked initially to make sure there are no errors serious enough to prevent the file from going on to the Acknowledgement process.
Total Number of Encounter Records Received	The number of Encounter records counted by the ACK process.
Total Number of Registration Records Received	The number of Registration records counted by the ACK process.
Encounters (Add, Change, Delete, Rejected)	Individual Encounter record counts per category.
Registration (Add, Change, Delete, Rejected)	Individual Registration record counts per category.

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.

The following figure illustrates the NDW process flow.

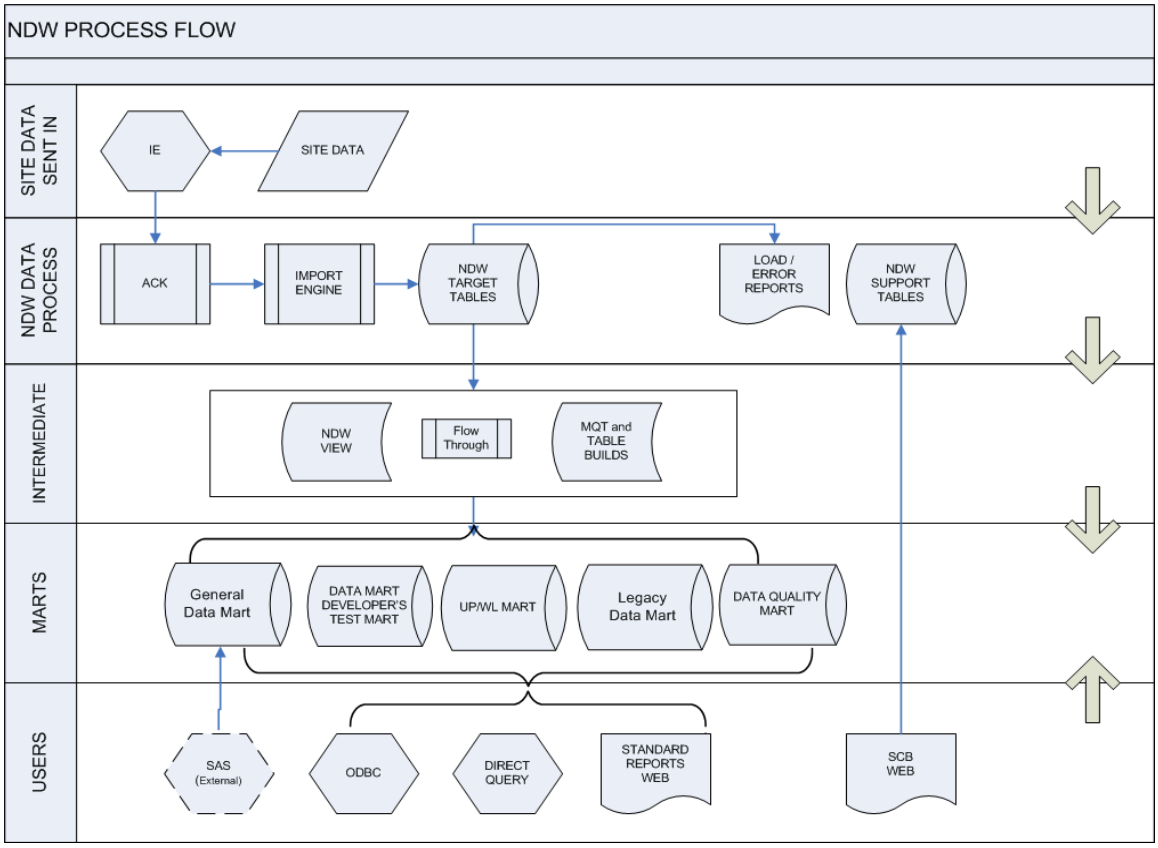


Figure 7 - NDW Process Flow

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.

Sweeper Process

The NDW environment utilizes a weekly automatic Sweeper process to reduce storage requirements. This process moves data which is either duplicated or no longer useful in day-to-day reporting but must be retained for both historical and data quality reporting, from ENCTR.ENCTRSS and related ENCTR schema tables to the ENCTR_HIST.ENCTR and ENCTR_HIST related tables.

After a successful move, the original target (ENCTR schema) tables are reorganized to remove empty space left by the deleted records. This is necessary because the primary key, ENCTRSS_ID, is a sequential, unique value, and back space cannot be utilized.

The weekly Sweeper process, which is usually a precursor for the weekly ETL process, is illustrated in *Figure 8* - Weekly Sweeper Process.

An additional sweeper process – the History Sweeper – is also utilized. Whereas the weekly Sweeper moves data from the ENCOUNTER schema to the ENCOUNTER_HIST schema as soon as the data becomes non-current, the History Sweeper, which is run quarterly, moves data from ENCOUNTER_HIST into ENCOUNTER_HIST_ARCHIVE when the data is more than 2 years old (as of the most recent September 30th). The History Sweeper process is illustrated in *Figure 8* - Weekly Sweeper Process

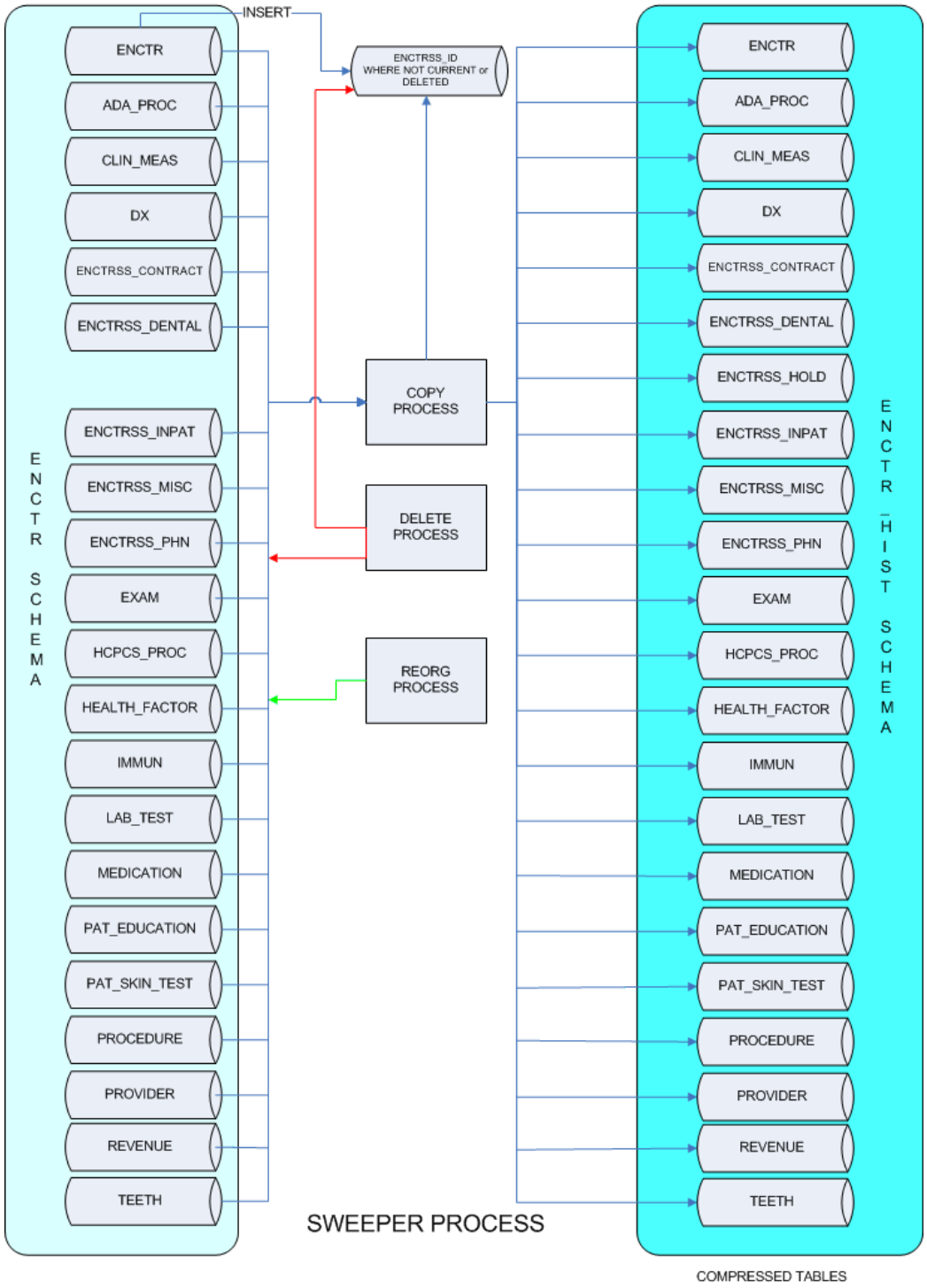


Figure 8 - Weekly Sweeper Process

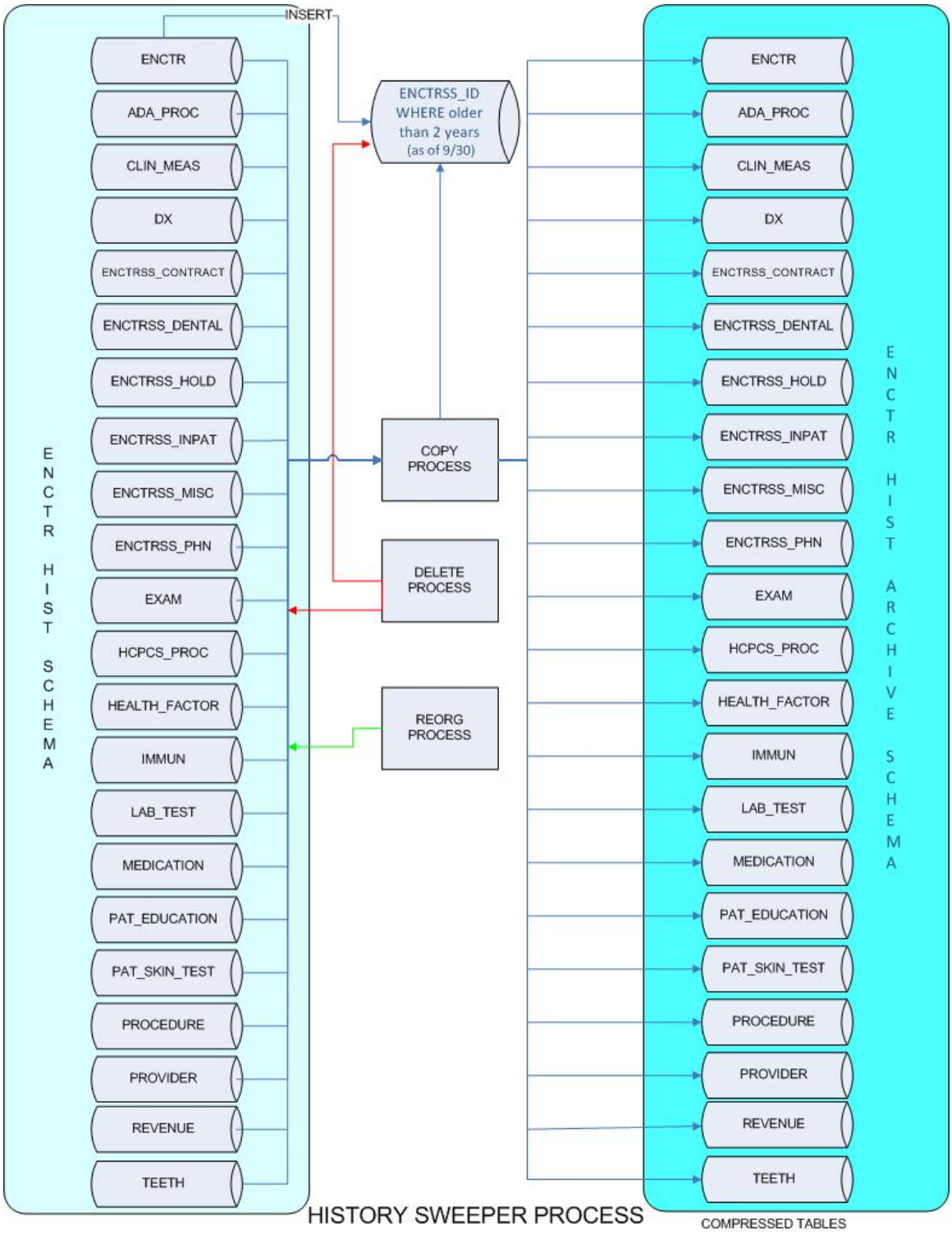


Figure 9 – History Sweeper Process

MatchMaker Process

Since it is possible for encounter data to be received before registration data, it is important to match encounter data to registration data when the registration data is received. 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.*

Backup Process

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. The process moves the data, untransformed, from the IE into 'staging' tables from which it may electronically be compared to the data in the NDW. The data movement and comparison is done largely by views. The DI process identifies the file type by the directory in which files are stored. External data, such as filename, must be appended to the data stream of the data loads. Some basic NPIRS business rule transformations are addressed within the process. The DI process consists of a set of drivers, scripts and SQL calls. The diagram below illustrates the relationship between the various DI modules. For a detailed description of the current data integrity procedures as well as the mechanism used to initiate the process, see the document entitled *NPIRS NDW Data Integrity Process*.

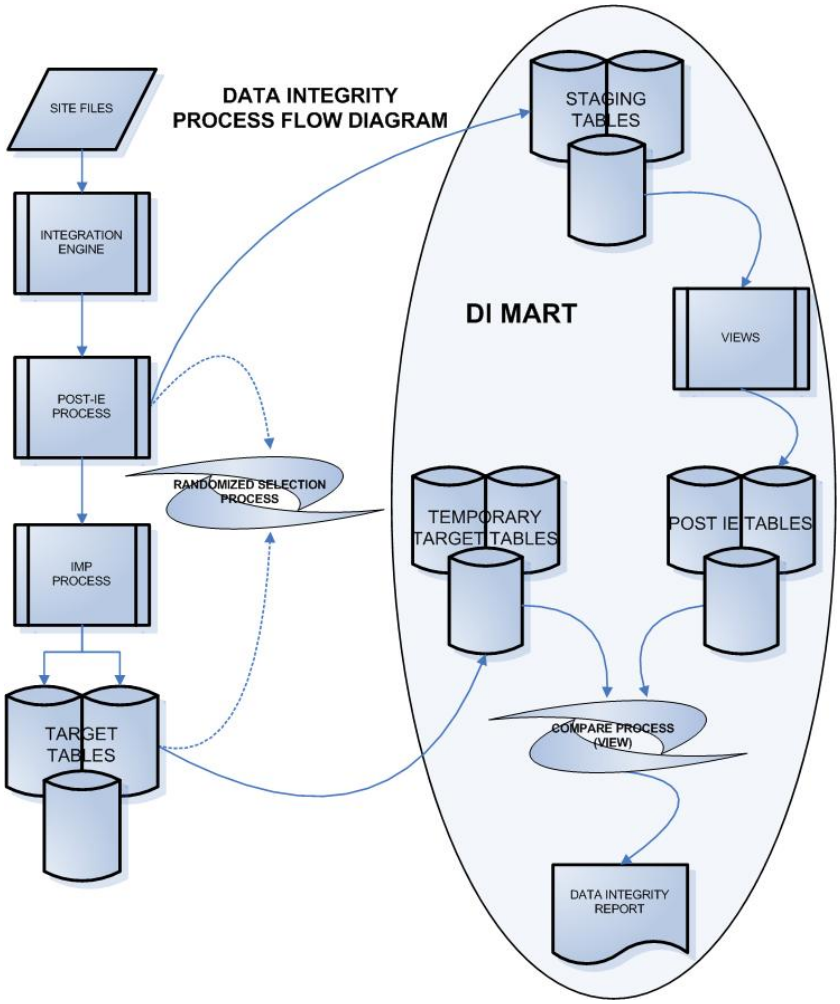


Figure 10 – Data Integrity Process

Reporting Process

Reports, except one-time reports, are generally:

1) Driven by a SQL stored procedure utilizing one or more defined views. The stored procedures are typically defined utilizing the IBM Workbench.

2) Utilize a Business Objects Universe for reporting with appropriate data security filters.

Un-Duplication Process

Registration Data

The NDW can receive multiple registrations for a given person if they visit more than one facility. In order to count a person only once in each IHS Area, a pre-established set of business rules are applied to un-duplicate the registration records which are used for User Population reporting purposes and have been reviewed and approved by HHS Headquarters (HQ), Division of Program Statistics (DPS).

Encounter Data

Similarly, the same encounter record can be received from one source multiple times, or the same record received from multiple sources. Since the NDW stores all encounter records which are received (duplicates or not), the un-duplication process must be applied to encounter records which are used for Workload reporting purposes, using a pre-established set of business rules.

For **registration data**, un-duplication is accomplished using a special software package, IBM's Infosphere QualityStage. QualityStage is a data cleansing component that provides a parallel processing engine to process large stores of source data which provides the basis for accomplishing the following data cleansing concepts:

- Resolving data conflicts and ambiguities
- Conforming data by transforming data types into a standard format
- Creating one unique result.

Un-duplication of **encounter data** is performed in two steps. The first step is an initial un-duplication of records that have been sent more than once from the same source due to modifications by the site. A unique identifier is sent to determine whether a record already exists in the database. The unique identifier combined with the date last modified

is used to determine if the incoming record is newer than the existing record. If it is, the existing record is marked as non-current (Current Encounter Flag = 'No') and the incoming record is marked as current (Current Encounter Flag = 'Yes').

The second step takes place after the encounter data Extract, Transform and Load (ETL) process and before report generation. An official un-duplication is performed against the entire NDW database and encompasses all export file formats. Depending on the value of the most recent Row Create Date/Timestamp, a Duplicate Flag is set according to whether a record is a duplicate or not. Once a group of records is identified as being multiple iterations of the same encounter, the Duplicate Flag (DUP_FG) of the record with the most recent Row Create Date/Timestamp is set to 'No' and the remaining records' Duplicate Flags set to 'Yes.'

Additional details on the un-duplication process are located in the document "*NPIRS Basic Business Rule"s at http://www.ihs.gov/NDW/documents/WUPS/NPIRS_Basic_Business_Rules_v2.0.pdf*.

Process Locations

Processes associated with Backups, MatchMaker, monitoring, and ETL (other than Userpop/Workload) are located within the Instance owner directories. The ETL processes for Userpop/Workload initial extract are run within the NDW Production Data Mart's NDWMART schema. Other processes, such as IMP, ACK, and reporting are stored in production owner 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.

NDW Data Flow

The NDW receives and processes data from multiple sources. The NDW loads data regardless 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 is first routed to the IE, which reformats the file into a post-IE data export file. The post-IE data file is then transported 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 it has been loaded, the data is copied to WILDHRSE and HOLLYWD, which are synchronized with SANDIA on a periodic basis. This 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*". 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*".

The Userpop/Workload data mart, which contains all data needed for workload and user population reporting, is embedded in WILDHRSE. This arrangement better maximizes the transactional functions of SANDIA with little or no impact to data loads, minimizes ETL and calculation time and impact for Userpop/Workload processes, and improves data research capability.

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

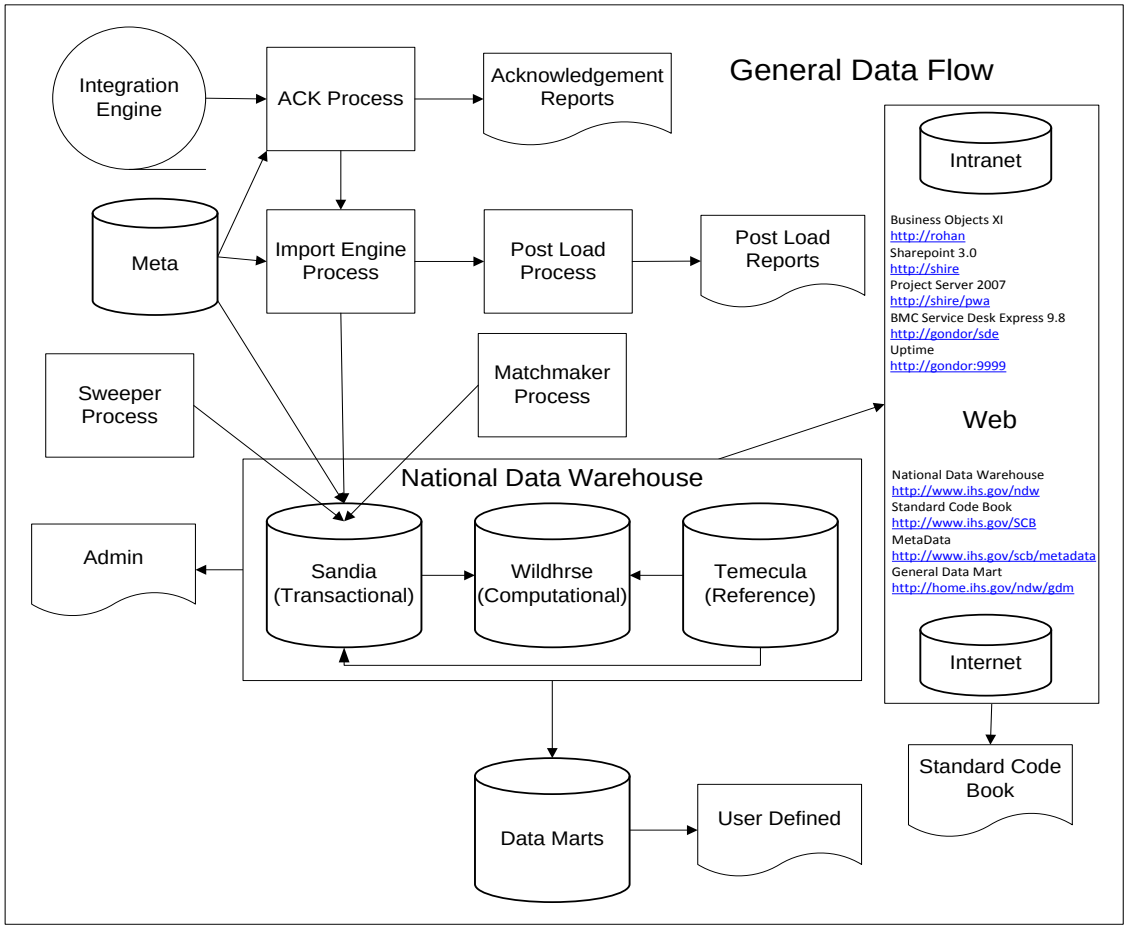


Figure 11 - NDW Data Flow

Change Management Plan

Change Management of the database or processes follows IHS standards and "*NAP_4_Change Management Plan"*.

Emergency Management Plan

The NDW is identified as a mission critical database in the Emergency Management Plan (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 to authorized individuals. Each critical data mart/database within the NDW environment is also referenced with the EMP plan.

An abridged version, which contains a general description of the EMP plan with proprietary information removed, is available for review on the NDW Informational website.

Proposed Enhancements

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

1) A semantic layer between the NDW and the user is being developed and enhanced to allow easier report management, data provenance, reusability, and some BI functions.

2) More tightly managed applications using both Rational and Workbench. This will improve the Enterprise Application Integration (EAI) index and the manageability and flexibility of the system to satisfy the government's current and future needs.

3) Implementation of AIX and DB2 workload management. This will allow better control of IHS resources across multiple data bases and servers.

Appendix A: Rack Mapping at BIA

As of April, 2011, Indian Health Service (IHS) has completed the process of relocating the NPIRS section of the data center from 5300 Homestead Road NE, Albuquerque, NM to the BIA Data Center at 1000 Indian School Road NW, Albuquerque, NM. The diagrams on the following pages illustrate the configuration of the current hardware components in the new location.

Figure 12 - NPIRS Server Rack DR

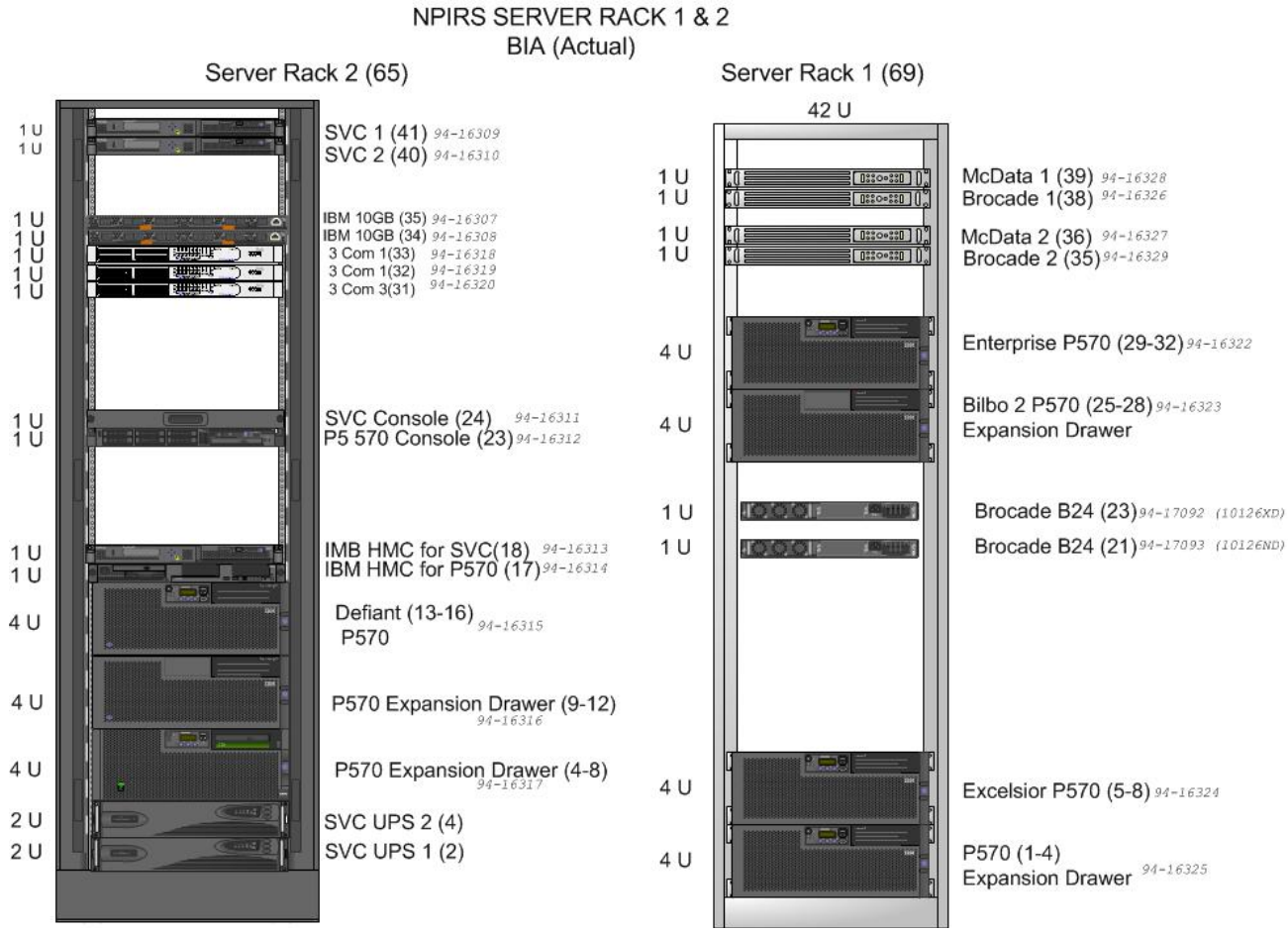


Figure 13 - NPIRS Server Racks 1 and 2

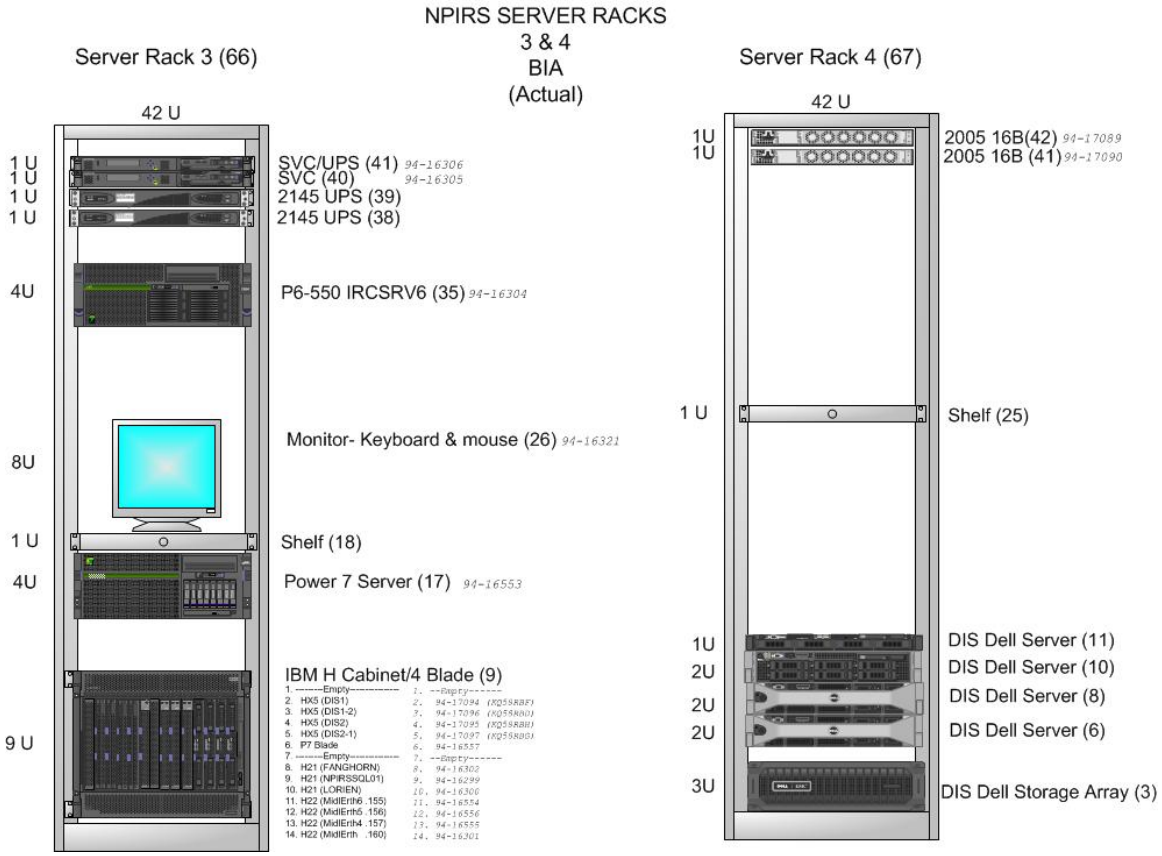


Figure 14 - NPIRS Server Racks 3 and 4

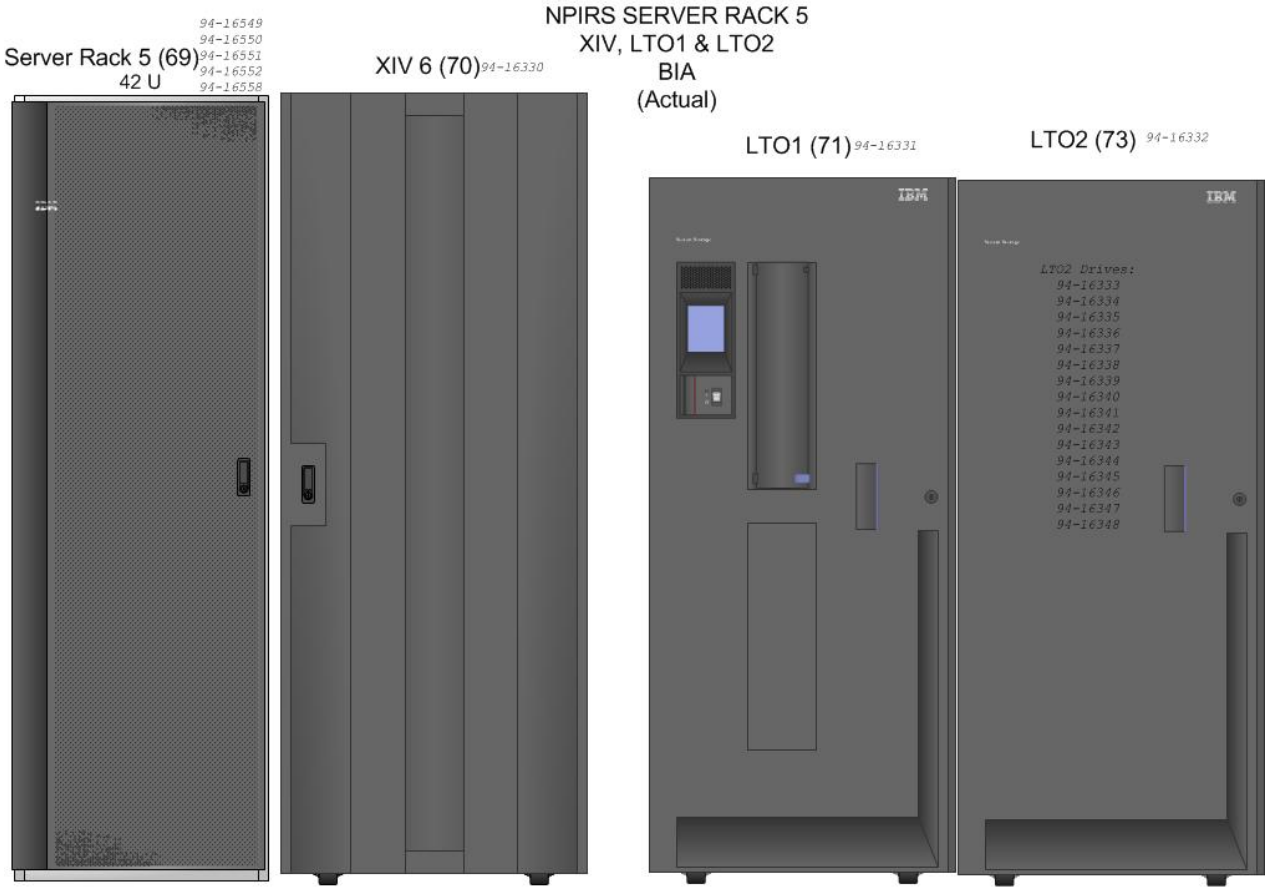


Figure 15 - NPIRS Server Rack 5

Appendix B: Weekly ETL Process

Encounter Weekly ETL

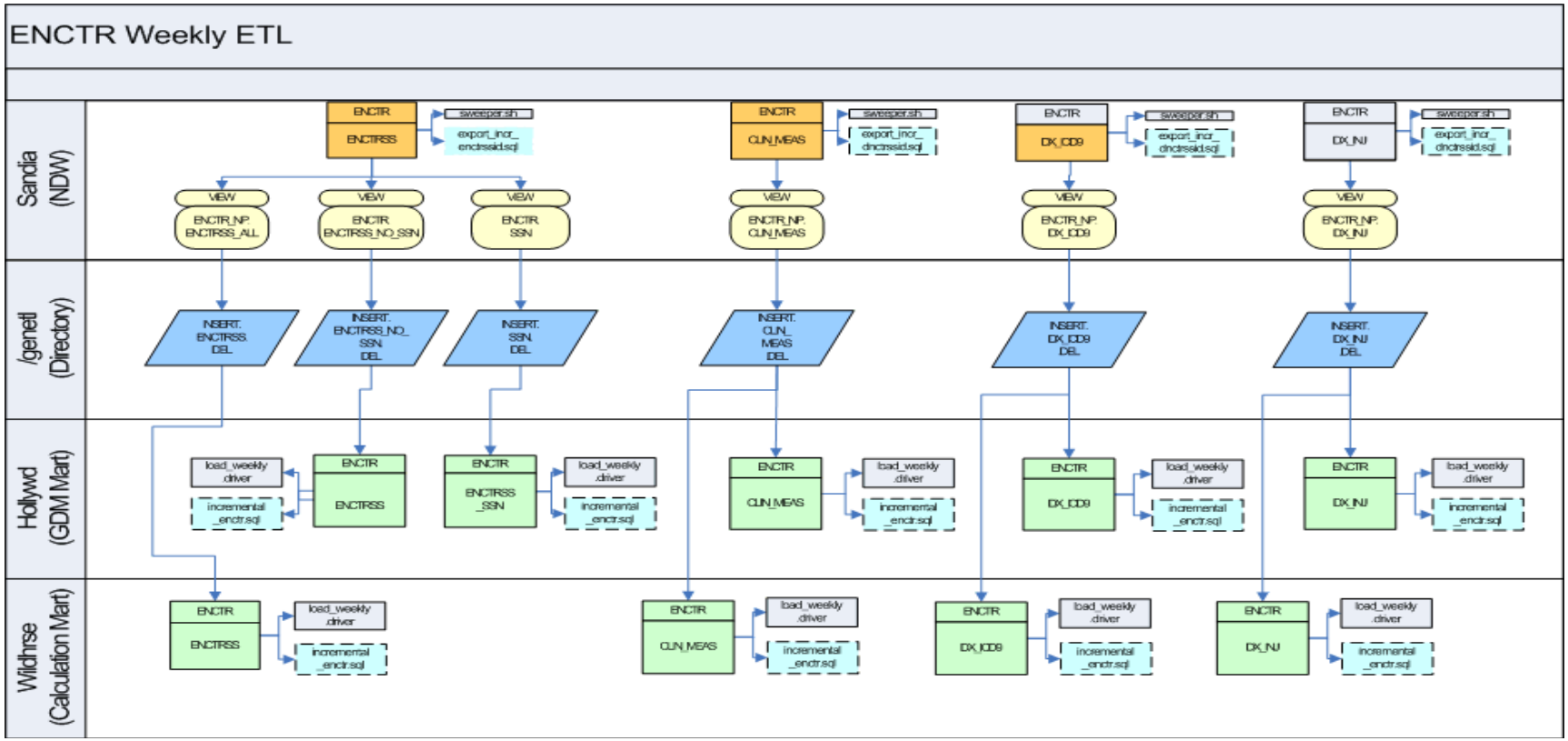


Figure 16 - ENCTR weekly ETL (1)

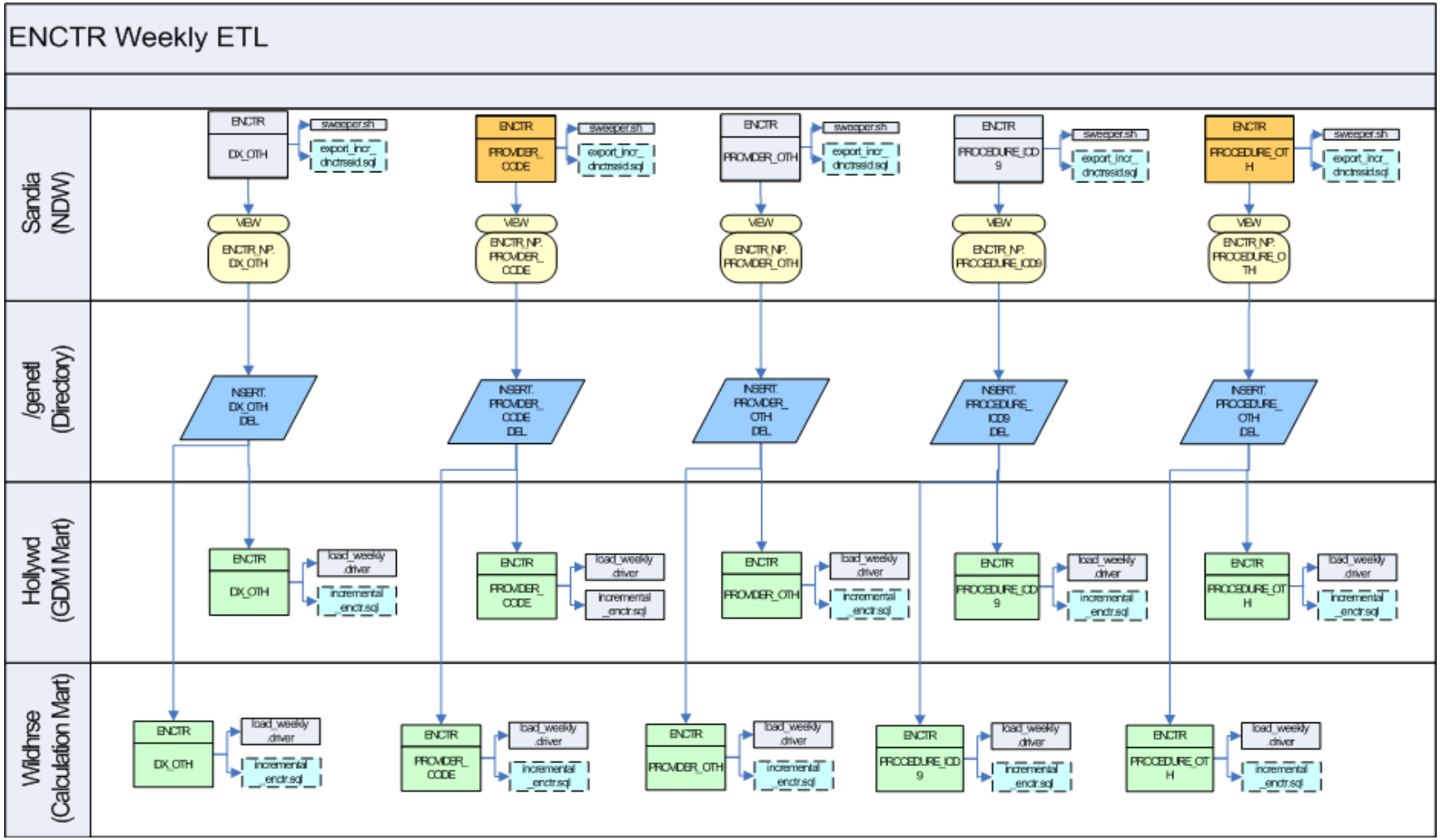


Figure 17 - ENCTR weekly ETL (2)

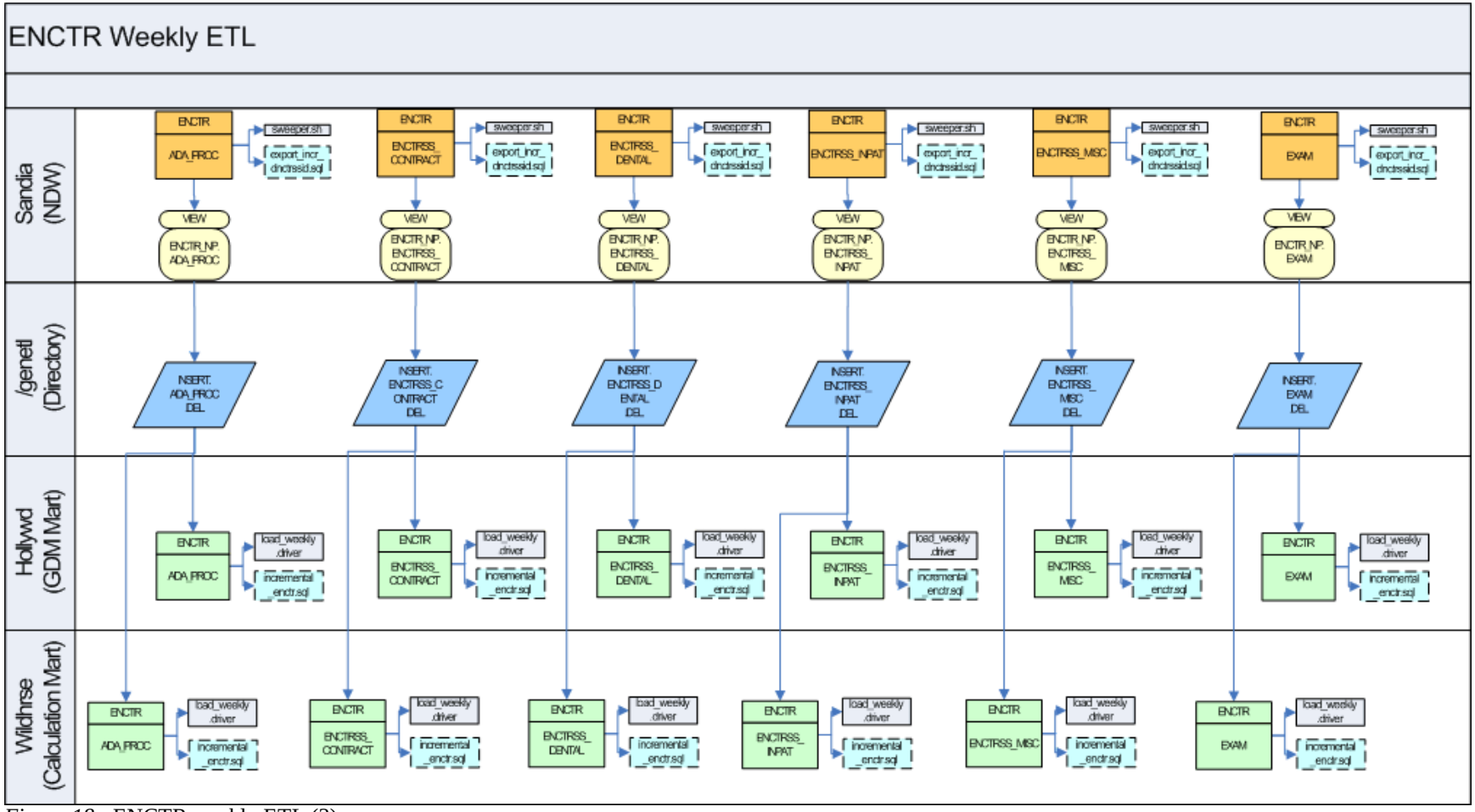


Figure 18 - ENCTR weekly ETL (3)

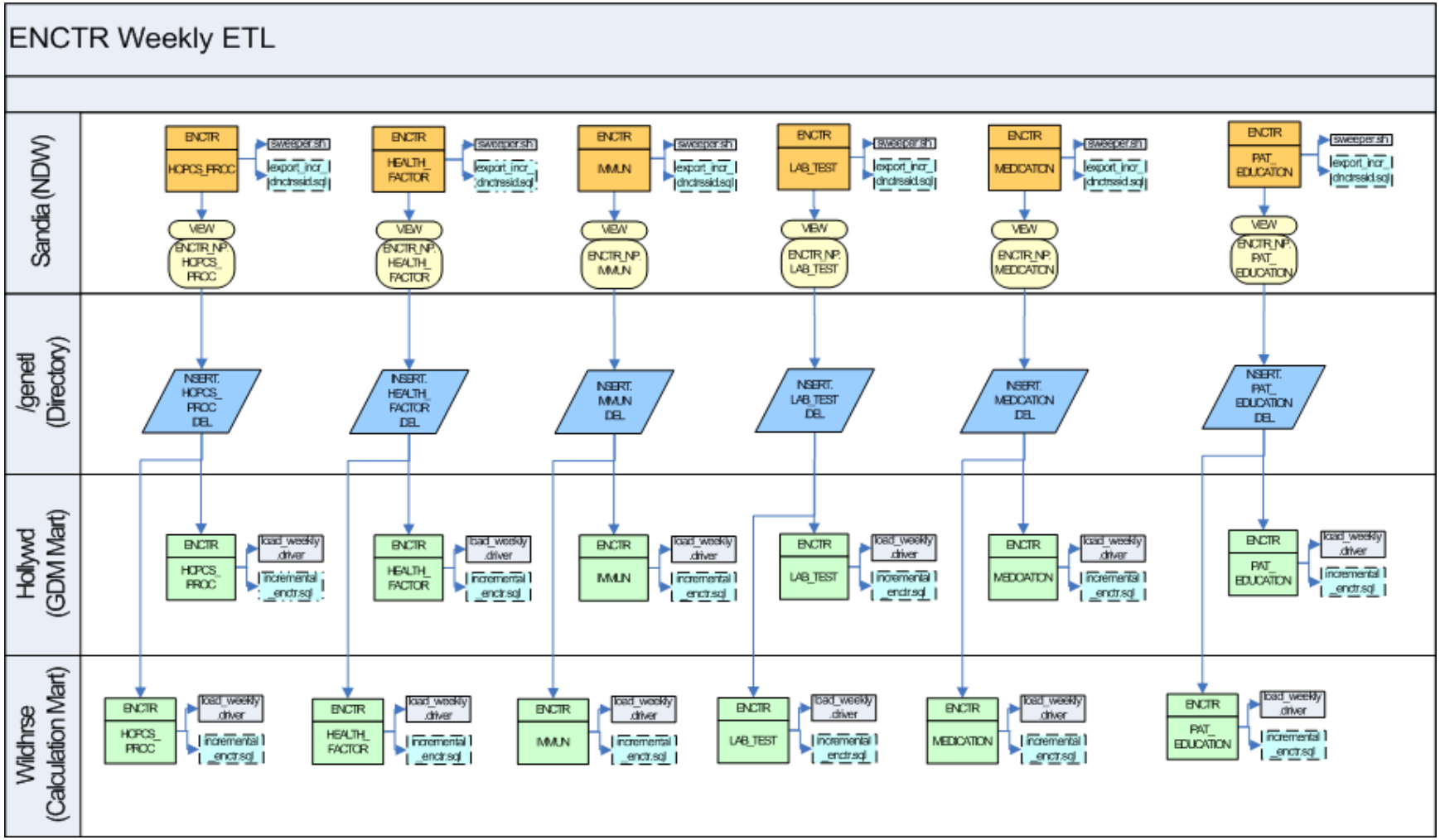


Figure 19 - ENCTR weekly ETL (4)

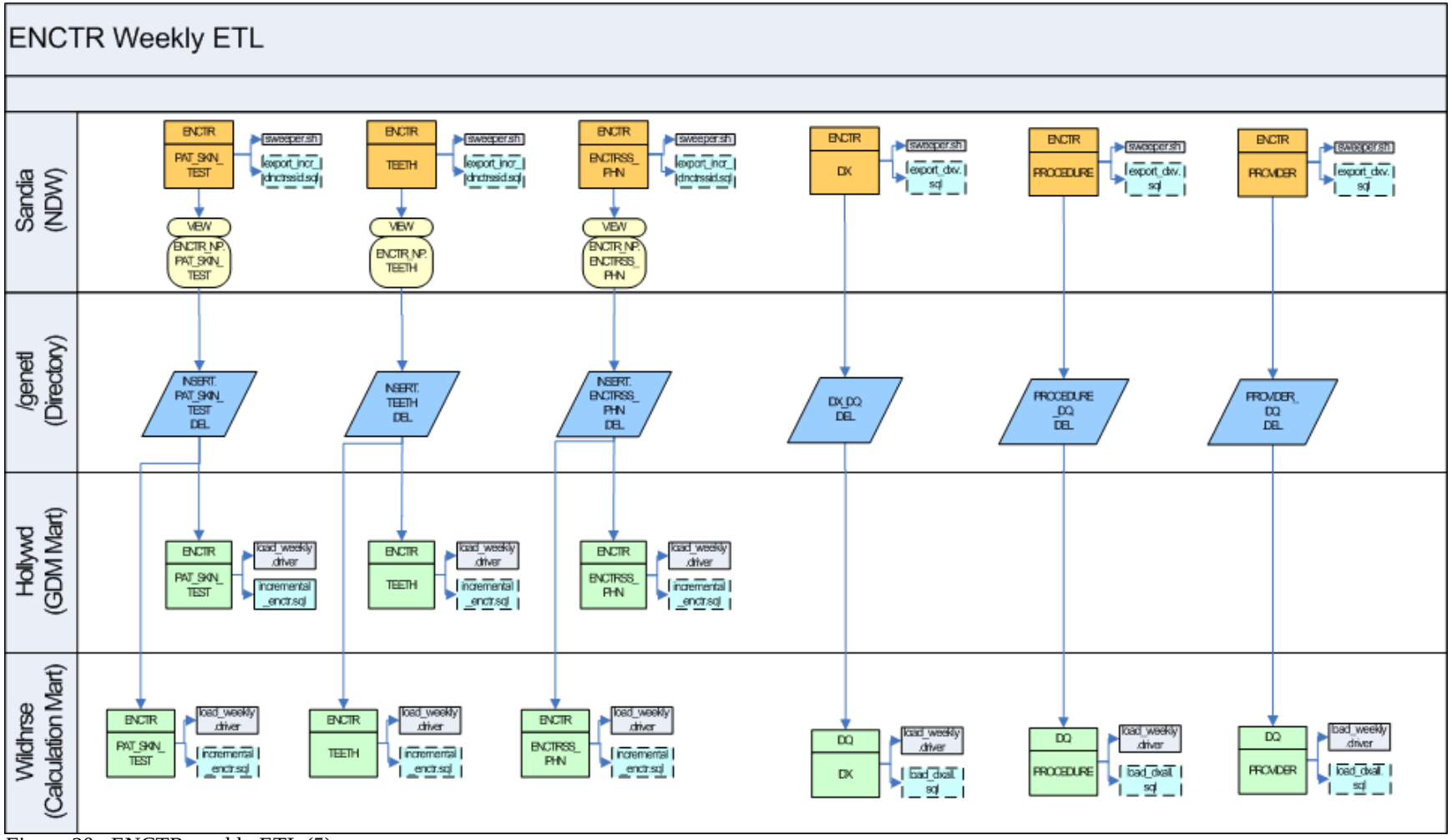


Figure 20 - ENCTR weekly ETL (5)

REG Weekly ETL

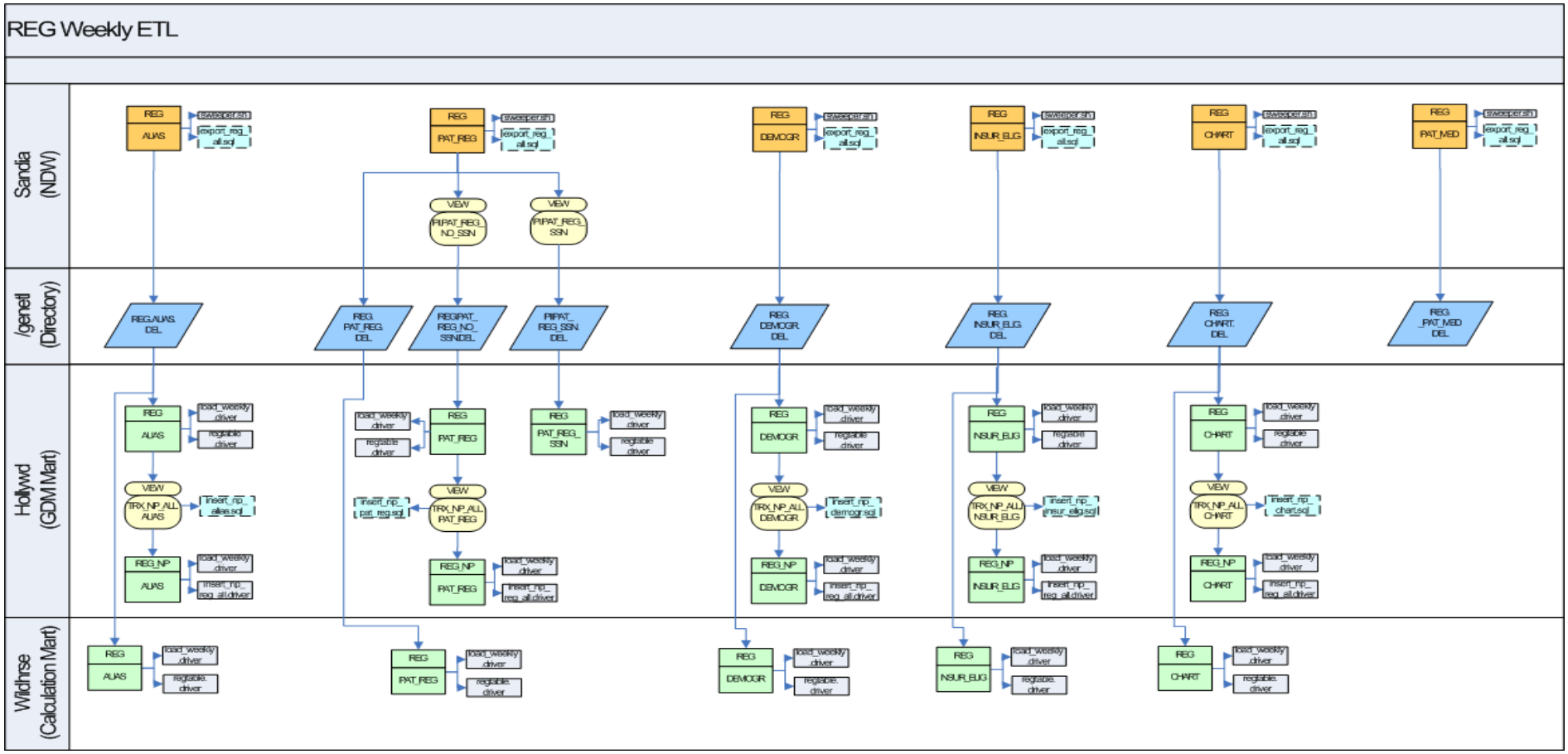


Figure 21 - REG weekly ETL

ENCTR_HIST Weekly ETL

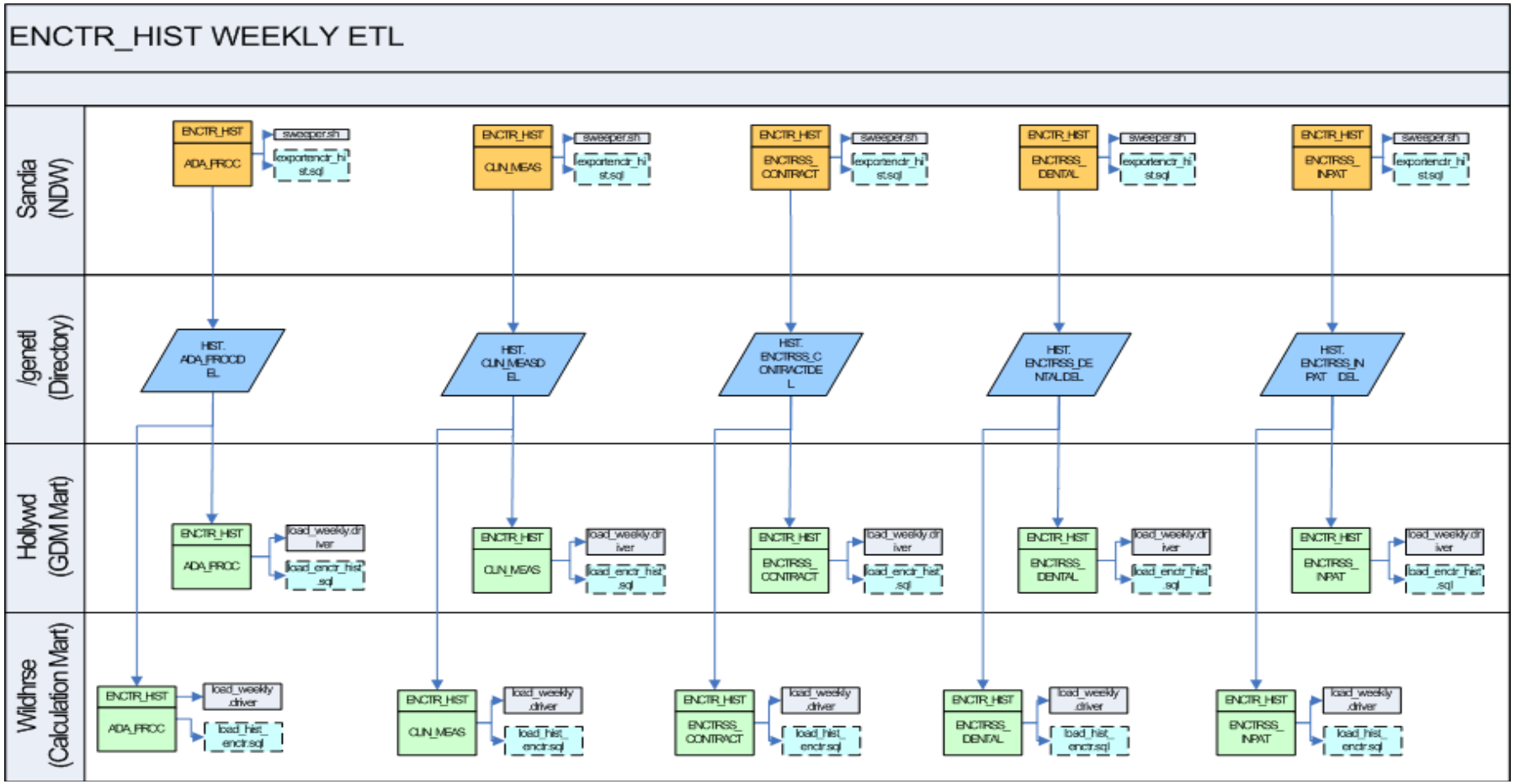


Figure 22 - ENCTR_HIST weekly ETL (1)

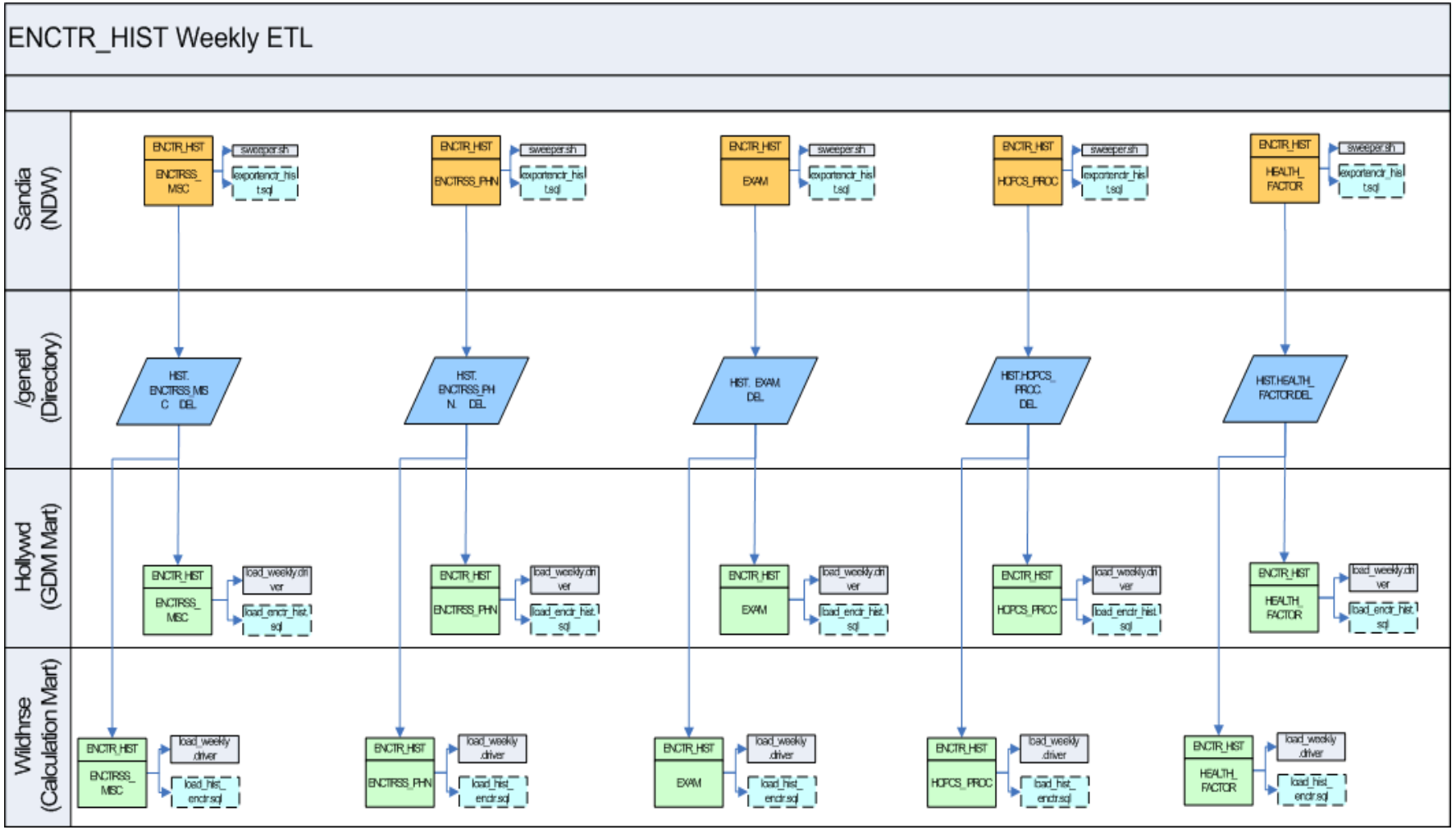


Figure 23 - ENCTR_HIST weekly ETL (2)

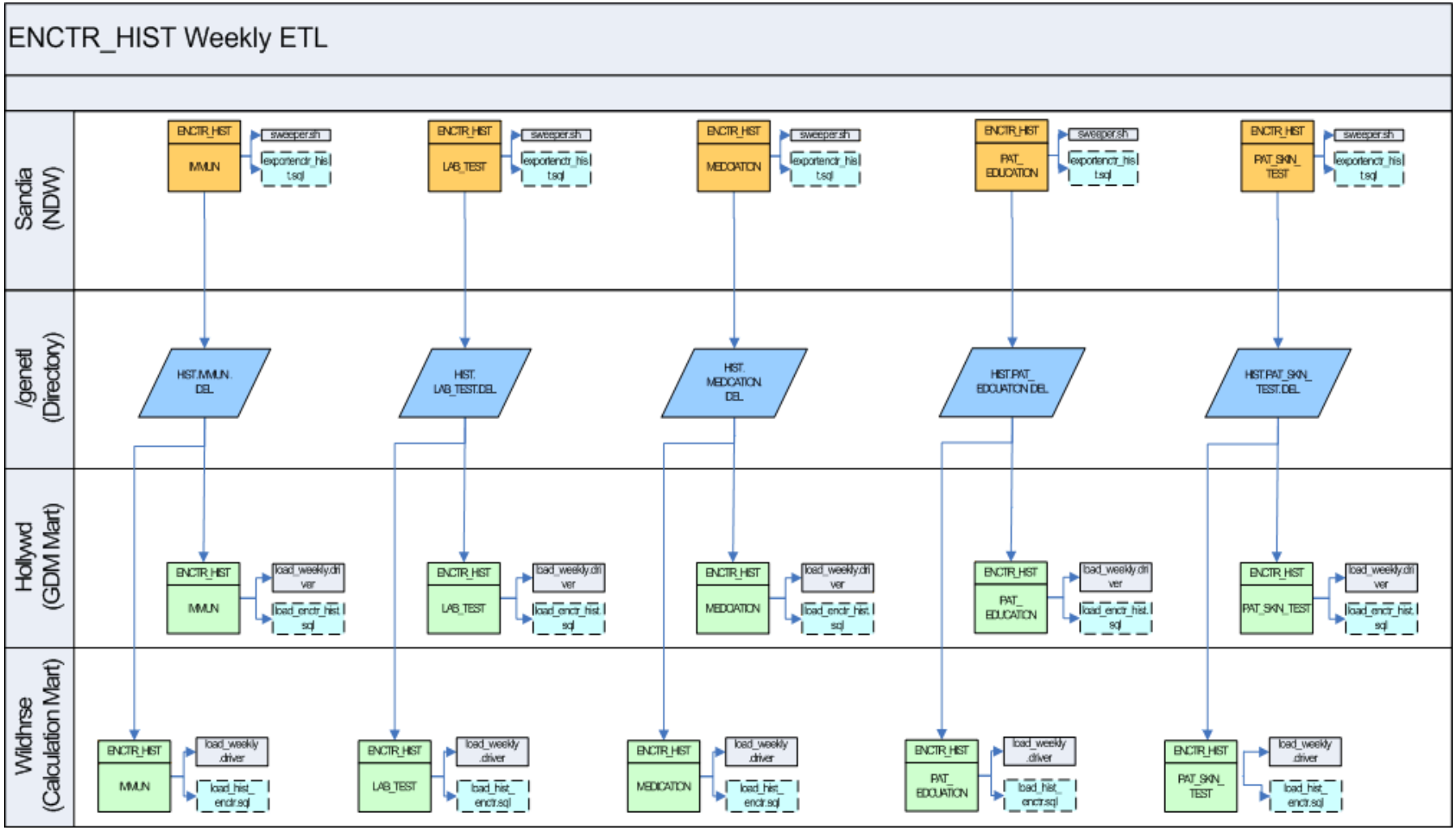


Figure 24 - ENCTR_HIST weekly ETL (3)

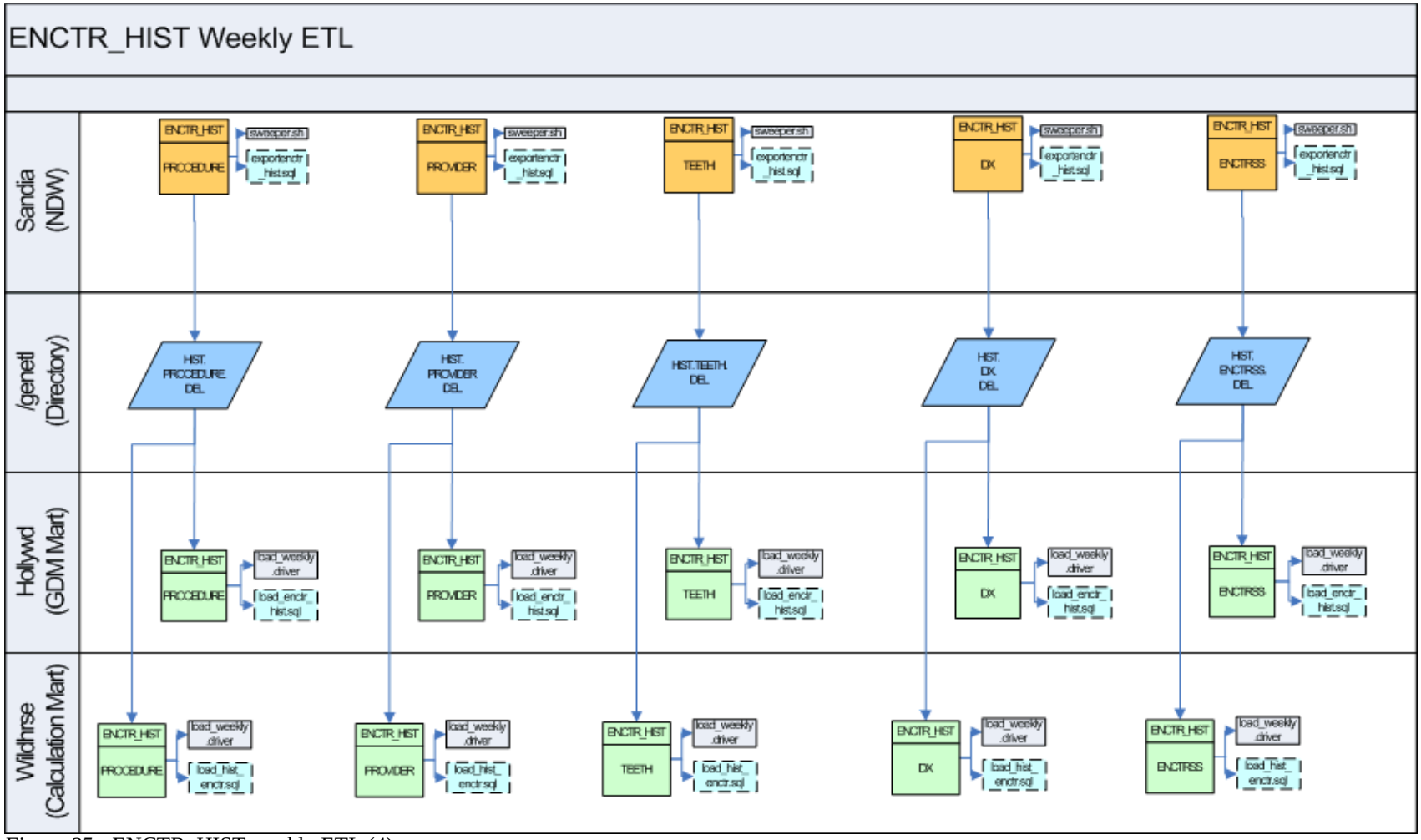


Figure 25 - ENCTR_HIST weekly ETL (4)

MatchMaker Weekly ETL

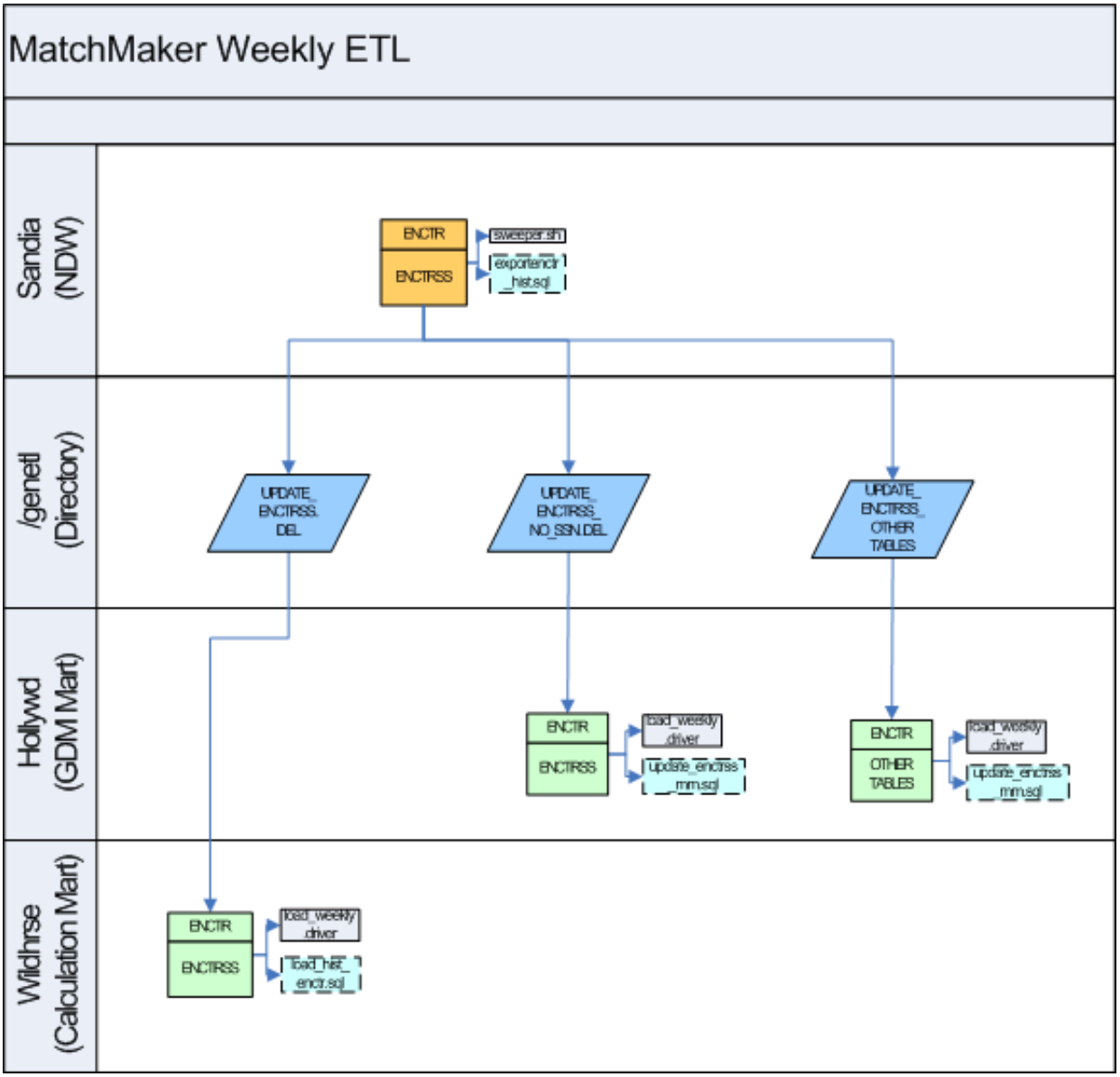


Figure 26 - MatchMaker weekly ETL

ADMIN_INFO Weekly ETL

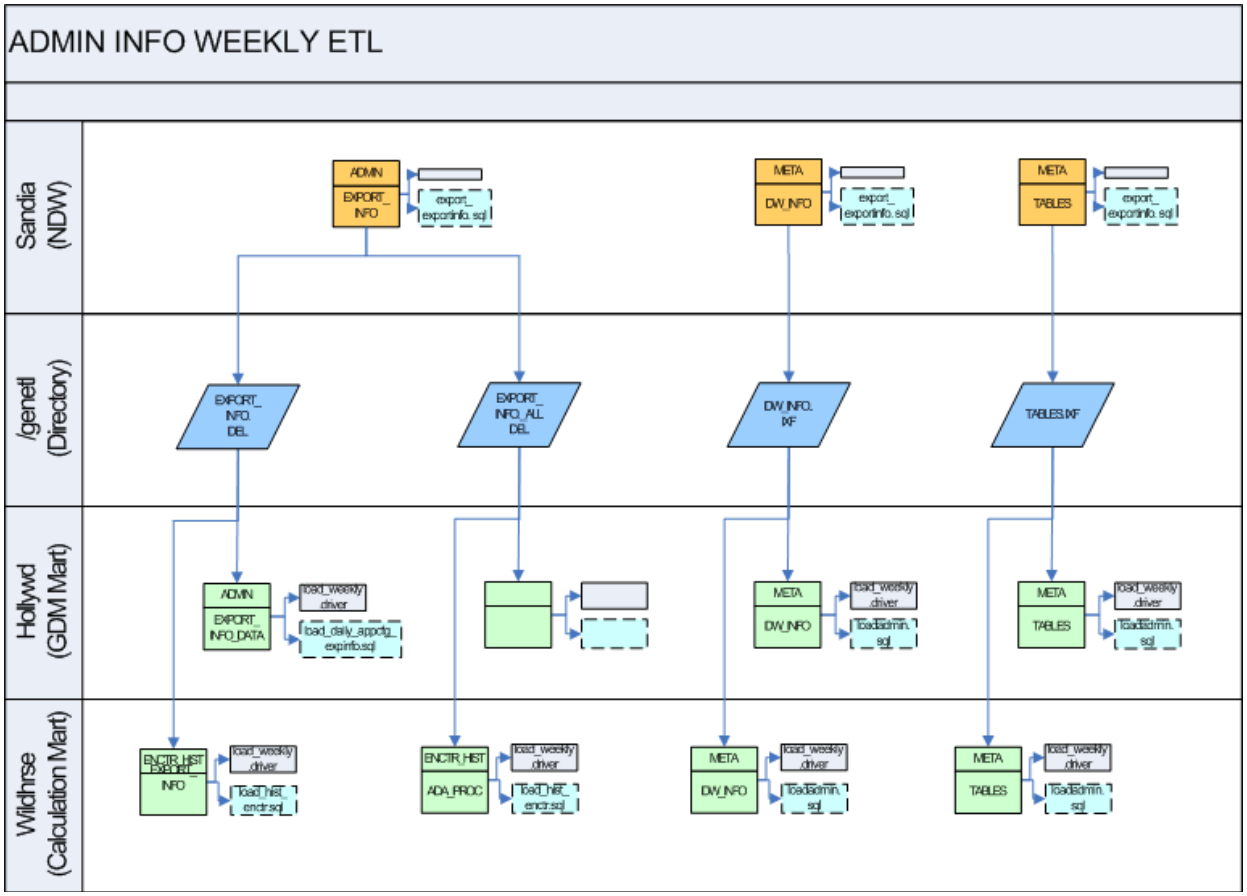


Figure 27 - ADMIN_INFO weekly ETL

Appendix C: NPIRS SAN Switch Port Summary

FABRIC A				FABRIC B			
SANSW1		SANSW3		SANSW2		SANSW4	
Port	Usage	Port	Usage	Port	Usage	Port	Usage
0	ISL to SANSW3	0	ISL to SANSW1	0	ISL to SANSW4	0	ISL to SANSW2
1	----------	1	gollum_A_fcs0	1	----------	1	gollum_A_fcs2
2	----------	2	smeagol_A_fcs0	2	----------	2	smeagol_A_fcs2
3	----------	3	ircsrv6_fcs2	3	----------	3	ircsrv6_fcs1
4	----------	4	SVC3p1	4	----------	4	SVC3p2
5	----------	5	SVC3p3	5	----------	5	SVC3p4
6	SVC1p1	6	SVC4p1	6	SVC1p2	6	SVC4p2
7	----------	7	SVC4p3	7	----------	7	SVC4p4
8	SVC2p1	8	Blade 1	8	SVC2p2	8	Blade 3
9	SVC1p3	9	Blade 2	9	SVC1p4	9	Blade 4
10	SVC console p1	10	ircsrv6_fcs0	10	SVC console p2	10	ircsrv6_fcs3
11	SVC2p3	11	LTO4_drive10	11	SVC2p4	11	LTO4_drive9
12	smeagol_B_fcs1	12	arwen_fcs1	12	gollum_B_fcs1	12	arwen_fcs0
13	gollum_B_fcs0	13	----------	13	smeagol_B_fcs0	13	----------
14	----------	14	----------	14	----------	14	----------
15	XIV_mod5_p3	15	XIV_mode4_p1	15	XIV_mod4_p3	15	XIV_mod5_p1
16	----------	16	----------	16	----------	16	----------
17	LTO4_drive1	17	----------	17	LTO4_drive4	17	----------
18	LTO4_drive2	18	----------	18	LTO4_drive5	18	----------
19	LTO4_drive3	19	----------	19	LTO4_drive6	19	----------
20	LTO4_drive7	20	----------	20	LTO4_drive8	20	----------
21	bilbo_A_fcs0	21	----------	21	bilbo_A_fcs1	21	----------
22	bilbo_B_fcs0	22	----------	22	bilbo_B_fcs1	22	----------
23	----------	23	----------	23	----------	23	----------

Figure 28 – SAN Switch Port Summary