



F E D E R A L  
S T U D E N T A I D

*We Help Put America Through School*

**FSA Modernization Partner**

**NSLDS II Reengineering  
Data Architecture Detailed Design**

DRAFT

Version 1.0

**September 30, 2002**

## Table of Contents

<b>1</b>	<b>INTRODUCTION .....</b>	<b>3</b>
<b>2</b>	<b>APPROACH .....</b>	<b>4</b>
2.1	ENTERPRISE DATA WAREHOUSE (EDW) .....	4
2.2	EXTRACT, TRANSFORMATION, AND LOAD (ETL) PROCESS.....	5
2.3	DATA MART.....	6
2.4	METADATA REPOSITORIES .....	6
2.5	CLIENT ACCESS.....	6
<b>3</b>	<b>SCOPE.....</b>	<b>6</b>
<b>4</b>	<b>NSLDS DATA DEFINITION LANGUAGE (DDL).....</b>	<b>8</b>
<b>5</b>	<b>LOGICAL DATA MODEL.....</b>	<b>9</b>
5.1	PURPOSE OF LOGICAL DATA MODELING .....	9
5.2	EDW DATA MODEL COMPONENTS.....	9
5.2.1	<i>EDW Subject Area Diagrams.....</i>	<i>10</i>
5.2.2	<i>EDW Enhancements .....</i>	<i>10</i>
<b>6</b>	<b>MAPPINGS TO NSLDS II EDW DATABASE FROM NSLDS MAINFRAME.....</b>	<b>13</b>
6.1	DATA CONVERSION APPROACH.....	13
6.2	DATA EXTRACTION.....	13
6.3	DATA TRANSFER.....	14
6.4	DATA LOAD .....	14
6.5	DATA VALIDATION .....	14
<b>7</b>	<b>APPENDIX A - DATA DEFINITION LANGUAGE .....</b>	<b>16</b>
<b>8</b>	<b>APPENDIX B - EDW DATA MODELS.....</b>	<b>17</b>
<b>9</b>	<b>APPENDIX C - DATA MAPPINGS FROM NSLDS TO NSLDS II.....</b>	<b>18</b>

## Document Control

Version Number	Description	Release Date	Author
1.0	Initial Design of DDL, Data Model, and NSLDS to NSLDS II EDW database	09/30/2002	Terry Helwig

## 1 Introduction

The National Student Loan Data System (NSLDS) was established as part of the Higher Education Act of 1965, as amended, to provide a comprehensive repository of information about Title IV recipients and their loans, grants, lenders, guaranty agencies (GAs), servicers and schools. As NSLDS has evolved since its implementation in 1994, it has become hampered by a number of challenges related to discrepancies between the quality and timeliness of NSLDS data and the system of record, its analytical capabilities, and operating costs. Given these challenges, a project to modernize the system has been undertaken with the following objectives:

- **Data Warehousing:** Improve usability of the NSLDS data repository through new tools.
- **Internal Federal Student Aid (FSA) Direct Access:** Improve customer satisfaction through better quality and usability of NSLDS information.
- **Outsourced Enrollment Tracking:** Balance FSA data needs with burdens placed on the financial aid community.
- **Financial Partner Data Feed Reengineering:** Take greater advantage of data resources available within FSA and from the financial aid community
- **Common Record Extension:** Improved financial integrity, reduce FSA costs associated with NSLDS and related operations.

The first phase of the NSLDS Reengineering effort is called NSLDS II. The first release of NSLDS II will focus on the Data Warehousing and Internal FSA Direct Access Opportunities, as well as assessing ways to support existing requirements through the NSLDS II or other modernized systems. Later releases of work will address the other objectives and additional requirements that FSA may have.

## 2 Approach

For the current design of the NSLDS II data architecture, there are five main components:

- **Enterprise Data Warehouse (EDW):** This database is similar to the current NSLDS database on the mainframe.
- **Extract, Transformation, and Load (ETL) Process:** This process details the loading of data from the trading partners along with the replication of data from the Enterprise Data Warehouse database to the Data Mart database.
- **Data Mart:** This database contains a subset of the Enterprise Data Warehouse designed for aggregation of data for reporting needs.
- **Metadata Repository:** The Metadata repositories are “data about data” that define the business and technical needs of the users through natural language rather than relational database terms.
- **Client Access:** There are multiple clients to the NSLDS II system that access data through web sites, data feeds, and direct SQL access. The data architecture has to be flexible enough to handle all of the client user requirements.

### 2.1 Enterprise Data Warehouse (EDW)

Figure A represents a graphical depiction of how the NSLDS II data will be architected. While this diagram is subject to change based on technical limitations found during the build cycle, the current design is the EDW to be populated through the conversion of the NSLDS mainframe database. This EDW database will approximate the current NSLDS design and be the transactional database for the NSLDS II system. All updates and inserts will be done to the EDW and replicated to a Data Mart database that will provide the structure for all reporting requirements. The logical model for the Data Mart will be built in a dimensional model designed in a snowflake schema.

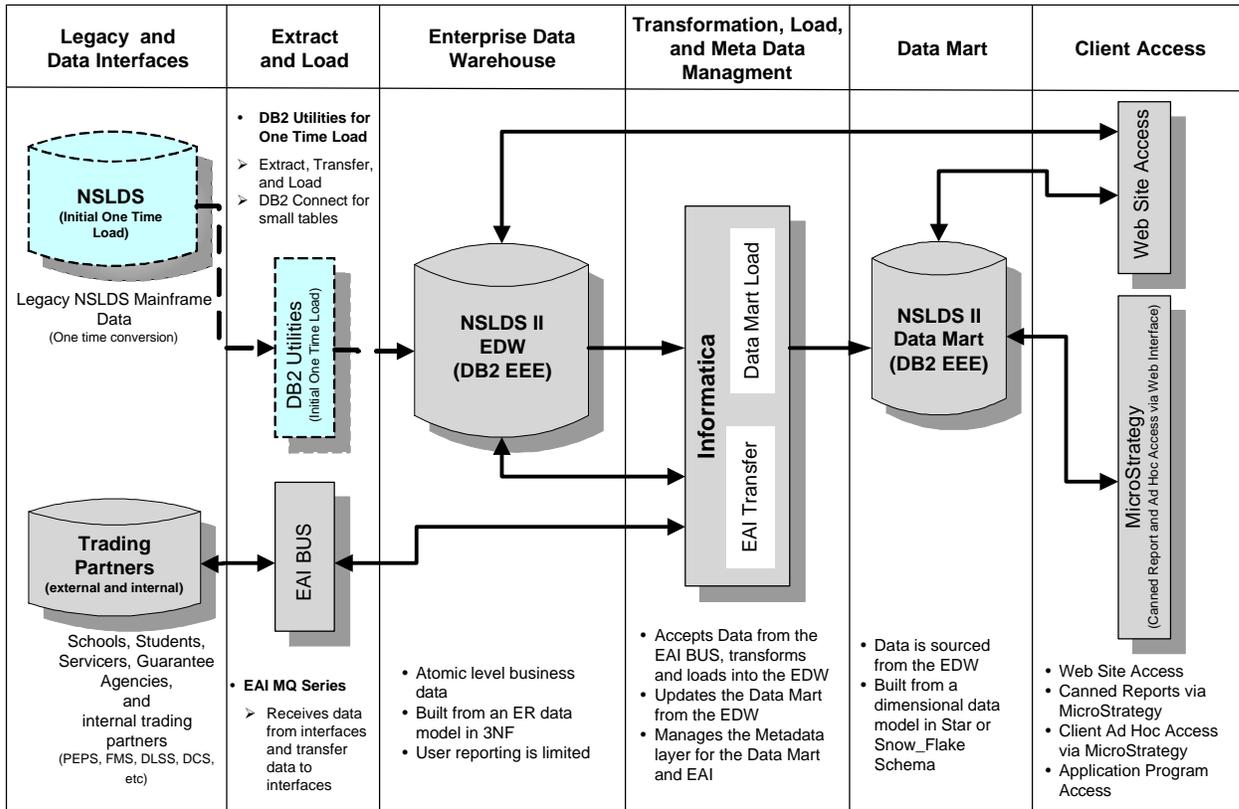


Figure A. NSLDS II Data Architecture

## 2.2 Extract, Transformation, and Load (ETL) Process

To load data into the system and handle all transformation and replication chores, two COTS products have been chosen, IBM's MQSeries software and Informatica. By using COTS tools, the following benefits are realized:

- Less code to maintain.
- Less development time by leveraging existing code or GUI interfaces for data loading, transformations and replications.
- Less maintenance and operations cost as vendors continually update their products with new features and fixes.

The Enterprise Application Integration (EAI) team is a group of individuals whose mission is to provide a set of common technology services that enables the sharing of processes and data of disparate systems to support end-to-end business processes. The EAI architecture enables many "stovepipe" applications to exchange information via common, reusable methods and infrastructure. By leveraging IBM's MQSeries software suite, the EAI team provides the capability to integrate web-based applications, Data Warehouse environments, COTS packages, and existing legacy systems within the FSA technical environment. The cost of the software suite and hardware to run it is shared among the projects within FSA that utilize the bus.

For the NSLDS II project, the EAI bus will dynamically load data via the MQSeries interface for trading partners who are on the bus and the student internet aid gateway (SAIG). All data will be delivered via

EAI to directories on the Informatica server, which will perform the load and any transformations into the DB2 EEE EDW database. Informatica will also be utilized for replication between the EDW database and the Data Mart database.

### **2.3 Data Mart**

In order to provide reporting capabilities a second database, the Data Mart, is designed as a subset of the EDW for specific ad hoc queries and standard reports. This database is a multi-dimensional database with the ability to drill down to a specific level of detailed defined by the user community. This database uses the COTS tool, MicroStrategy, for its reporting presentation layer. The MicroStrategy tool is the standard reporting application for on-line analytical processing (OLAP). It is a web based tool that analyses data from a number of different perspectives or business dimensions by drilling up or down through the data provided in the Data Mart.

### **2.4 Metadata Repositories**

The Metadata repositories are “data about data” that define the business and technical needs of the users through natural language rather than relational database terms. Objectives of the metadata are the following:

- To provide a means to improve the productivity of the administrators/developers and the reliability of the business intelligence solution.
- To provide a means to assist business analysts in locating and understanding the data in the data warehouse environment.

There are two main areas that contain metadata in the NSLDS II data architecture. The MicroStrategy reporting application creates a metadata layer with pointers to the actual data, attributes of the data, and parameters of the reporting project. Additionally, the Informatica application has a metadata layer that stores configuration information and data about the different transformations performed and data loaded into the EDW database. The Informatica server’s metadata management tool will be used to maintain all metadata layers in the NSLDS II data architecture.

### **2.5 Client Access**

Client access to the application will be primarily through the NSLDS II Financial Aid Professional (FAP) web site and the Student Access Financial Aid Review web site. These web sites are for schools, lenders, guarantee agencies, and students to examine the history of various loans. Additionally, the FAP web site will have a link to the MicroStrategy reporting server where users will be able to run ad hoc and pre-built reports. The web sites will access a combination of the EDW database for updating information and the Data Mart for read only information on the web site. A secondary path for data access will be for a small super user community to have SQL access to the EDW and Data Mart databases for running ad hoc queries.

## **3 Scope**

This document will detail the data architecture design for the NSLDS II Release 1 phase of work. For this draft version, this document will focus on the following tasks:

- **Data Definition Language (DDL):** The DDL is used to create and delete databases and database objects. Database administrators use the DDL to create new and empty tables as well as delete and modify existing tables.
- **Logical Data Model:** A logical data model is a precise and unambiguous expression of business facts. The data model shows entities, attributes, keys and relationships.
- **Data Mappings from Legacy NSLDS Mainframe to Mid-Range NSLDS II:** This section details the conversion mappings for porting data from the legacy NSLDS mainframe to the new NSLDS II enterprise data warehouse. These mappings detail at the field level the transition of data from the mainframe to the field level on the NSLDS II EDW database.

This document is currently in draft format and will be expanded to include the configuration of the NSLDS II development, test1, test2, and production databases. Additionally, this document will include the archive strategy, AIX operating considerations, development Data Mart design, and setup of each of the databases.

## 4 NSLDS Data Definition Language (DDL)

The NSLDS II DDL will be used to create the tables inside the enterprise data warehouse and Data Marts, which includes table names, column names, data types as well as partitioning keys and indexes. Using prepackaged modeling software, Computer Associates AllFusion ERwin Data Modeling suite, a DBA will create the DDL. Additionally, the ERwin software will be used to help create and maintain all of the databases, data warehouses and enterprise data models. By using this software databases can be developed more quickly with more efficiency and maintainability.

Once ERwin creates the initial copy of the DDL by re-engineering the existing NSLDS database, the DBA will modify the file to reflect the new requirements and architecture of the NSLDS II data warehouse environment. Some examples that a DBA would need to modify the DDL would be the addition of an indexing strategy or choice of singular or composite partitioning keys for multiple-partitions tables.

There will be two separate DDL files defined for NSLDS II. The first will be used to construct the EDW portion of the system. The second will construct the tables that encompass the NSLDS II Data Mart.

The following examples of syntax will help users to get a better understanding of what the DDL does:

This command will create an empty database called 'NSLDS':

```
CREATE SCHEMA NSLDS
```

The following CREATE TABLE statement adds a table called STUDENT in the database with the field's *first name*, *last name* and *student ID*. Each of these three columns must not have NULL values. The final clause contains the partitioning key. The partitioning key acts as the unique identifier for each row. There will be no duplicate values of the student\_id. This key also helps in partitioning the table across multiple nodes, which will aid in processing speed.

```
CREATE TABLE student  
(first_name char(20) NOT NULL,  
last_name char(20) NOT NULL,  
student_id INT NOT NULL  
PARTITIONING KEY  
(student_id))
```

Please refer to Appendix A for a complete data definition language for the enterprise data warehouse. Future iterations of this document will include the DDL for the NSLDS II Data Mart.

## 5 Logical Data Model

The new NSLDS II data architecture will consist of two logical data models, the EDW and the Data Mart. Data Models contain entities, attributes, and the relationships between them. Entities are items of relevance about which information can be kept. They refer to persons, places, things or concepts. Examples of entities are Loans, Students, Schools, and Lenders. Attributes are quantitative or descriptive characteristics of the entity. For instance, student name is a descriptor detail that users associate with the Student entity. Keys identify properties of entities, and relationships show association between two entities. This information, the entities and relationships between them, is represented in the logical data model in Entity / Relationship (E/R) relationship diagrams. The NSLDS II logical data model is recorded in the ERwin data modeling software. For this draft version, only the EDW data model will be presented in Appendix B.

### 5.1 Purpose of Logical Data Modeling

The two primary reasons for creating a logical data model are:

- **Improve communication:** Communication between the business users of the data and the developers is essential to a successful development project. Eventually, all of the work that a business does comes to rest in its data. Understanding the business nuances that are inherent in the data is one of the quickest and surest ways for a developer to obtain an understanding of the business. A logical model helps depict those nuances as it denotes entities and their relationships to each other. In developing a data warehouse, the developer needs this understanding to ensure that the data warehouse can answer the questions posed by the business.
- **Provide input into physical design:** The logical data model is one of the inputs into the physical data design. The other main inputs are access patterns, the performance criteria, and the technical infrastructure. These inputs will be documented in future releases of this document.

### 5.2 EDW Data Model Components

The EDW data model will have the similar logical and physical structures as the NSLDS data structures. The underlying EDW data model is in the third normal form (every non-key attribute should be functionally dependent on the full key and nothing but the full key), which translates to the EDW data model being normalized to ensure consistency, reduce redundancy, and maximize stability in the model. Normalization drives the following:

- Ensures all entities or attributes occur only once in the model.
- Assists in associating data attributes with entities based on properties of the data rather than on application requirements.
- Puts the keys and data together so that it can be maintained without jeopardizing the information's integrity.

The EDW data model is created on the ERwin data-modeling tool by reverse engineering the existing NSLDS DB2 database on the mainframe. All NSLDS tables were reverse engineered into ERwin and only entities required for NSLDS II are selected into the EDW data model. The EDW data model represents all the entities, attributes and relationships in a single E/R relationship diagram for all subject areas.

### **5.2.1 EDW Subject Area Diagrams**

A total data model diagram that lists all inter-relationships is too large for viewing in Microsoft Word on normal 8.5 x 11 printed formats. An effort is underway to print out the entire data model on poster board format for viewing. This will be part of the NSLDS II detailed design deliverable on Nov. 8<sup>th</sup> 2002. For this draft deliverable, the E/R relationship diagrams were created by subject area for easier viewing. A subject area centers on a particular subject matter as opposed to being generic and open to many different types of information. The following are major subject areas identified at FSA, which are displayed in Appendix B.

LOANS  
STUDENTSSCHOOLS  
LENDERS  
GUARANTY AGENCIES  
DEFAULT RATES  
FDSLPL  
TRANSFER MONITORING

For each subject area listed above, an E/R relationship diagram has been created in ERwin.

### **5.2.2 EDW Enhancements**

While the EDW data model has similar data structures as the legacy NSLDS mainframe database, there are a few changes to it. The first is that each EDW table will be expanded to include a date and time stamp to signify the occurrence of an update event in each row in the table. A new **TIMESTAMP** column is added to each table to reflect the date and time an update is applied to each row in the table. The new column will be as follows:

**DT\_TM\_STAMP** **TIMESTAMP** **NOT NULL** **WITH DEFAULT**

The data type of this column is **TIMESTAMP** and the default will be

0001-01-01 00:00:00.000000 where

Year = 0001  
Month = 01  
Day = 01  
Hour = 00  
Min = 00  
Sec = 00.000000

When updates are made to the EDW on a daily, weekly and monthly basis, each update will provide a current timestamp in the **DT\_TM\_STAMP** column. The ETL tool will examine this column on each update to the EDW. Other changes made to the EDW tables and columns are as follows:

NSLDS Table Name	NSLDS Column Name	Proposed Change	New Column Name	New Table Name
FS_SBMTL_RUN_ERR	FFEL_DUP_ID	Change the column name to some abbreviated form of "indicator of Separate Loan"	IND_SEP_LOAN	n/a
GA_SBMTL_RUN_ERR	FFEL_DUP_ID	Change the attribute name to some abbreviated form of "indicator of Separate Loan"	IND_SEP_LOAN	n/a
LOAN	FFEL_DUP_ID	Change the attribute name to some abbreviated form of "indicator of Separate Loan"	IND_SEP_LOAN	n/a
REINSUR_CL_RFD	n/a	Change the table name to some form of « Brankruptcy Claim Payment “. Only refunds on claims related to bankruptcy are tracked	n/a	BNK_CLM_AMT
LOAN_RPMT_PLAN	DT_ENTR_RPMT	Rename column to indicate this is the date the student entered a repayment PLAN	DT_ENTR_RPMT_PLAN	n/a
STU_DEM	CALNDR_YR	Rename column to award year	AWARD_YR	n/a
SCH_BR_SVR	n/a	Rename table to indicate that it stores "PERKINS"	n/a	PERK_SVR
LEN_BR	all	Not populated anymore. This table can be removed from EDW and Data Mart references.	n/a	n/a
LEN_BR_HOL	LEN_BR_CODE	Remove this column and rename table to indicate "LEN_HOL" rather than "LEN_BR_HOL"	n/a	LEN_HOL
LEN_BR_SVR	LEN_BR_CODE	Remove this column and rename table to indicate LEN_SVR rather than LEN_BR_SVR	n/a	LEN_SVR
STU_BR_ID	n/a	Rename table to	n/a	STU_DESIG

NSLDS Table Name	NSLDS Column Name	Proposed Change	New Column Name	New Table Name
		Student Designator – STU_DESIG		
SCH_ORIG_HIS	n/a	Not populated anymore. This table can be removed from EDW and Data Mart references	n/a	n/a
SCH_SBMTL_HIS	n/a	Rename table to Perkins Submittal History	n/a	PERK_SBMTL_HIS
SSCR_CYCLE	n/a	Not populated anymore. This table can be removed from EDW and Data Mart references	n/a	n/a

## 6 Mappings to NSLDS II EDW database from NSLDS Mainframe

### 6.1 Data Conversion Approach

NSLDS is updated daily through both internal and external interfaces as well as through the NSLDS Financial Aid Professional Website, CICS, and by the Raytheon Quality Assurance Team.

The NSLDS II data conversion will be broken into three separate conversions as shown in the figure below.

1. The initial conversion will be comprised of NSLDS data prior to a specific date.
2. The second conversion will load the interface files received post-extract into the NSLDS II EDW.
3. The third conversion will populate the NSLDS Data Mart with information from the NSLDS EDW.

The detailed data conversion procedures as well as the data mapping from the NSLDS II EDW to the Data Mart will be detailed in the final detailed design deliverable. For this deliverable, mappings from the NSLDS mainframe to the NSLDS II EDW database are detailed in Appendix C.

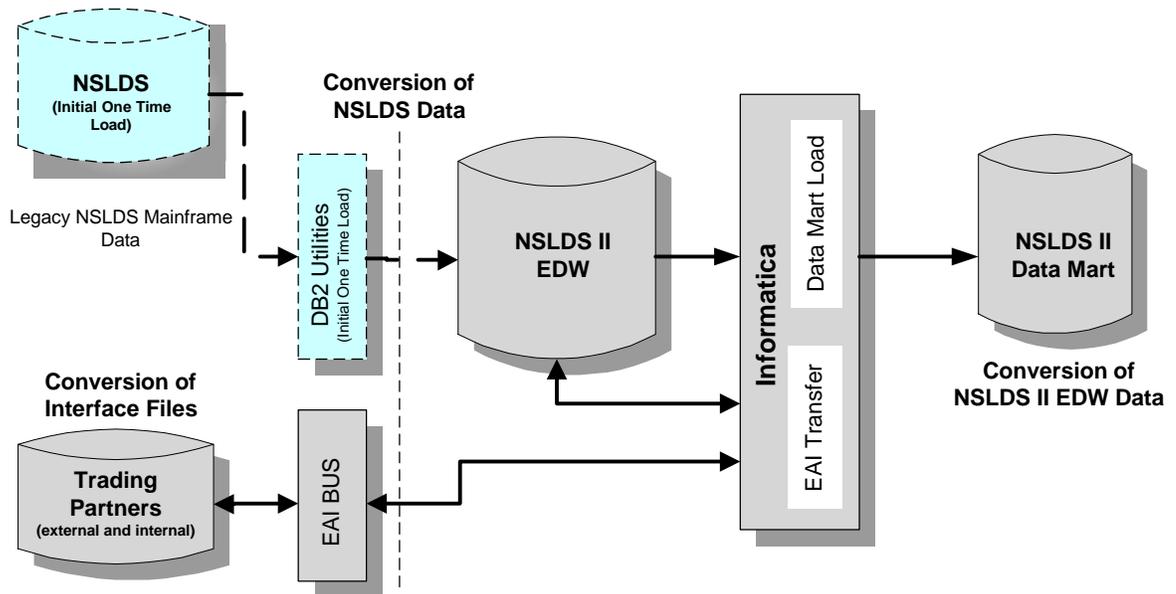


Figure B. NSLDS II Data Diagram

### 6.2 Data Extraction

NSLDS data will be exported from the mainframe using the DB2 Export Utility. The legacy NSLDS System Support entities will not be converted to the NSLDS II EDW, since these entities are used in the general operation of the legacy NSLDS database and not for storing data. Additionally, NSLDS entities that do not contain any data will not be converted.

### **6.3 Data Transfer**

The process for transferring NSLDS data will be developed as part of deliverable 94.3.3 Detail Design.

### **6.4 Data Load**

During the data load, the NSLDS data files extracted using the DB2 Extract utility will be loaded into NSLDS II using the DB2 Load Utility.

The NSLDS II EDW will be partitioned so that large amounts of data are stored across several disk arrays. A database partition is a part of a database that consists of its own data, indexes, configuration files, and transaction logs. A database partition is sometimes called a node or database node. In a partitioned database, entities may be located in one or more database partitions. When an entity is in a node group, which is a subset of one or more database partitions, some of its rows are stored in one partition and others are stored in other partitions.

Partitioning keys are used to determine the particular database partition where each portion of the data resides. The data imported from NSLDS must pass through a splitting phase where the data is divided and loaded into the correct partition. The DB2 AutoLoader utility will complete the split and load process using a hashing algorithm to partition the data. The AutoLoader utility then loads the data concurrently across the set of database partitions in the node group.

### **6.5 Data Validation**

Data will be validated after it is exported from NSLDS at two separate points during the conversion. First, data will be validated through the use of exception reports that will be reviewed after the data has been exported. Second, the data will be verified after it has been loaded into NSLDS II EDW by reviewing the following:

- The total record counts of the target entities and the source entities (e.g. NSLDS II EDW entities will be the target and the NSLDS entities will be the source entities).
- Sum totals for fields will be calculated in the target entities and compared to the totals in the source entities.
- Reports will be executed on the target system and the source system and reviewed for discrepancies.

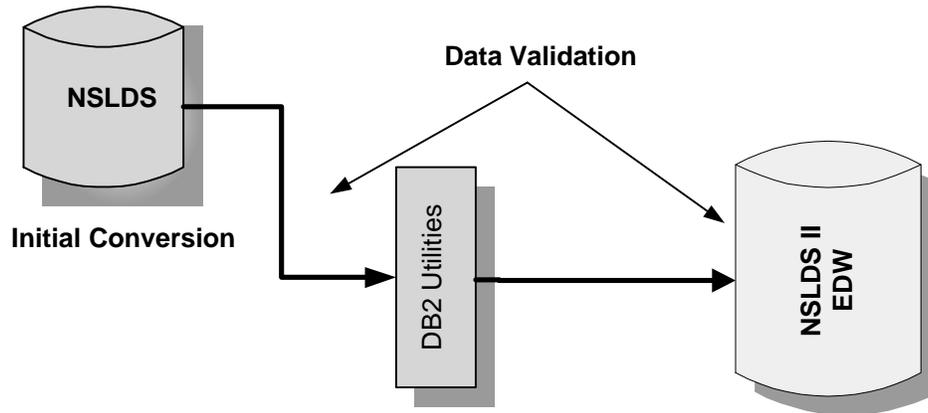


Figure C. Data Validation Process

## **7 Appendix A - Data Definition Language**

Please see NSLDS II Data Arch Appendix A & B Draft.doc

## **8 Appendix B - EDW Data Models**

Please see NSLDS II Data Arch Appendix A & B Draft.doc

## **9 Appendix C - Data Mappings from NSLDS to NSLDS II**

Please see NSLDS II Data Arch Appendix C Draft.doc