EPA's National Emission Inventory Criteria Data System Functions for Streamlining Data Processing and QA

Steven Boone
Director of Information Technology
and
Donna McKenzie
Database Analyst
E. H. Pechan and Associates, Inc.
3622 Lyckan Parkway, Suite 2002; Durham, North Carolina 27707; 919-493-3144

Rhonda Thompson
United States Environmental Protection Agency
Office of Air Quality Planning and Standards
Emissions Monitoring and Analysis Division
Emission Factors and Inventory Group
D205-01
Research Triangle Park, North Carolina 27711; 919-541-5538

ABSTRACT

The Emission Factor and Inventory Group (EFIG) of the United States Environmental Protection Agency's (EPA) Emissions Monitoring and Analysis Division (EMAD) compiles the National Emission Inventory (NEI) on an annual basis. For inventory data processed up through the 1999 base year, the final production criteria pollutant NEI data reside in an Oracle database on an EMAD Unix-based server. Data are submitted to EPA in the NEI input format (NIF) by State, local, and tribal agencies. To improve data quality, data integrity, increase data security and facilitate data analysis, the data undergo a multi-step process that involves data merging, processing, validating and updating.

Each emission inventory data file submitted to EPA are validated with EPA's Quality Assurance tool to determine gross issues with the data file. The criteria pollutant data are imported from the submitting NIF file types (Microsoft Access, ASCII or XML) via a Microsoft Access/Java-based tool with an graphical user interface into an Oracle transfer database that resembles NIF version 2.0 format. Additional fields are contained in the Oracle database to provide for data audit and log tracking capabilities as the data moves through the multi-step process. Counts of records and emission summaries are automatically performed on the input and resulting files to ensure full transfer of the data. Once the data are in Oracle, the data begin a series of steps to where detail diagnoses, data scrubbing and data augmentation are performed. Finally, data are moved to the NEI Oracle database structure, which is a more normalized database than the NIF, through extract, transfer and load (ETL) algorithms.

Criteria pollutant emission inventory data for the Mobile, Area (including nonroad and paved and unpaved roads) and Point source sectors are processed via the NEI Staging Database Process. Data are rediagnosed with Oracle-based routines to determine the details of data issues. The issues are organized by agency and provided to the agencies for comment. Agency

comments are incorporated into the process either via new data submittals or ad-hoc scrubbing procedures. When data are scrubbed (updated) by means of a standard scrub, which is performed on all data, or an ad-hoc scrub, which is performed for specific conditions, all data are audited. Auditing consists of archiving a copy of an entire data record before any changes are made. If a record contains several data columns that are to be updated either through standard or ad-hoc scrubbing procedures, the auditing of the record takes place before each update. The result is that a single data record can appear in the audit tables multiple times before the entire data update process is complete. During scrub intensive processing time periods which generally span about three months twice a year, literally tens of millions of records are updated or created in each workday. Data from states submitting emission inventory data are merged with data from the previous inventory from states that did not submit an updated file to achieve a national inventory database for reporting purposes.

INTRODUCTION

The Emission Factor and Inventory Group (EFIG) of the United States Environmental Protection Agency's (EPA) Emissions Monitoring and Analysis Division (EMAD) compiles the National Emission Inventory (NEI) on an annual basis. The inventory data are needed to evaluate emission trends in each State and to compare emission trends between States. The NEI is used as the basis for modeling and regulatory analysis conducted by EPA, States and Regional Planning Organizations (RPO's). Finally, the NEI comprises the information behind the National Air Pollutant Trends Update, published annually by EPA.

The NEI includes criteria pollutants, their precursors and hazardous air pollutants (HAPs). Upon completion, the NEI is expected to be used as the initial inventory for EPA regional- and local-scale modeling efforts to predict ambient concentrations, exposures, and the resultant risks to human health and the environment, for State/local/tribal (S/L/T) and RPO modeling efforts, for emission estimates for the "National Air Quality and Emissions Trends Report", for implementation of the 1990 Clean Air Act Amendments, for the Industrial SO_2 Report to Congress and to tracking progress toward program goals of the federal Government Performance and Results Act (GPRA). To support these uses, the NEI must be comprehensive, covering all areas of the United States. It covers all significant emission sources, including all stationary and mobile sources including large point/stationary sources and smaller sources wherever reported on an individual facility basis, area/non-point sources, onroad mobile sources licensed for use on highways and/or roadways and nonroad sources (e.g., construction, lawn/garden, boats, trains, airplanes).

For inventory data processed up through the 1999 base year, the final production criteria pollutant NEI data reside in an Oracle database on an EMAD Unix-based server. Data are submitted to EPA in various NEI input formats (NIF) by State, local, and tribal agencies. To improve data quality, data integrity, increase data security and facilitate data analysis, the data undergo a multi-step process that involves data merging, processing, validating and updating.

TOOLS

Since there are millions of records of information to be processed, the tools used for the NEI processing of criteria emissions consist of a relational database management system with robust transaction handling capabilities and various front end software tools to push the data into and through the process. The NEI database is a typical data warehouse designed to collect data from multiple sources within an organization or from many organizations. This data is then placed into a single (or virtually single) data storage area. The NEI is denormalized and relatively static. The data volume processing difference between a data warehouse and an on-line transaction processing (OLTP) system is the difference between a single load of a million rows of data (data warehouse) vs. a million loads of a single row of data (OLTP). The primary purpose of the NEI is to provide a platform to query, report, extract and analyze data.

In the NEI data warehouse, data is intended to represent a "snapshot" of a particular set of data. It is not updated by users - rather it is a collection of user data that usually has already been processed by an OLTP or other front-end system. The NEI maintains sets of data from many time periods for the purposes of developing trends analyses. Periodically (typically twice a year for the NEI) the NEI database is updated. This information (after cleaning and enhancement) is then made available through on-line analytical processing (OLAP) tools, web reports, and extraction to other data marts.

The NEI's back end relational database management system consists of an Oracle 9*i* database. Oracle is well-known for its robust transaction handling capabilities. In addition to its capability to import and process millions of rows of data efficiently, it has auditing and logging features that allow the NEI processing team to preserve all changes to the data and to document and preserve copies of the database tables during processing runs. By contrast, these same data were processed in Microsoft Visual FoxPro and Microsoft Access databases just three years ago. The new Oracle-based process is much more efficient, secure and self-documenting.

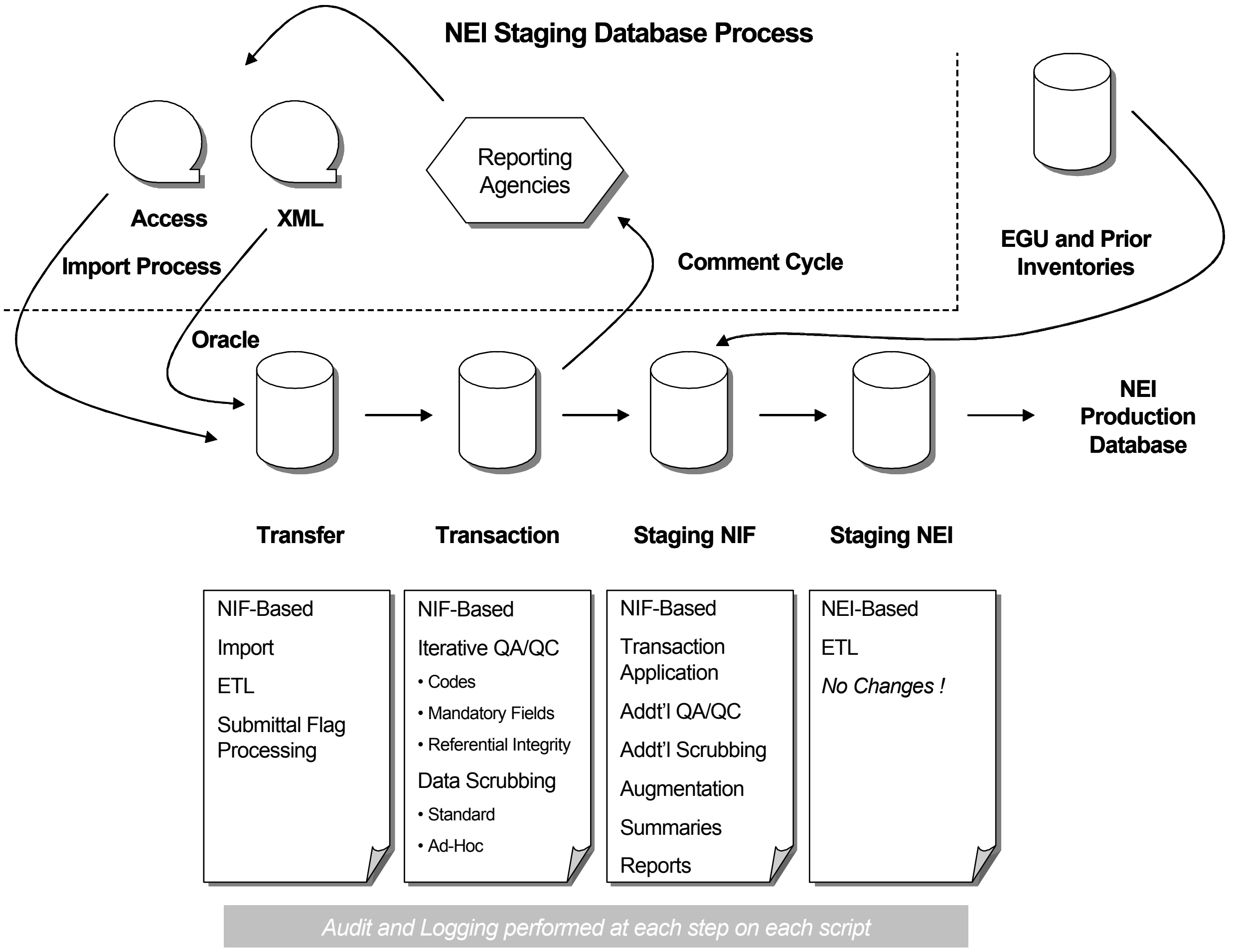
The front end tool set consists of several different tools that perform specific functions with the data. A Microsoft Access tool provides data loading and data extraction capabilities. The data loading tool reads NIF transactions in Access format and loads the data to Oracle using SQL*Loader, an Oracle-based data loading tool. The Access tool provides an interface for importing data in XML format and interfaces with a Java-based tool that was built to process the XML data and load it directly into the appropriate Oracle tables and fields. The Access tool also provides an interface with export capabilities of the Oracle data by State into Access NIF files for distribution.

The remaining front end tools consist of structured query language (SQL) scripts to process the data that are run via Oracle tools such as SQL*Plus and SQL*Reports and the use of Oracle tools such as Enterprise Manager to perform database administration tasks. The general script categories consist of Extract, Transfer and Load (ETL); quality assurance and quality control (QA/QC); data scrub (updates); data augmentation; summaries and reports.

PROCESSES

The NEI Staging Database Process consists of multiple, discrete processes. An overview of the process is shown in Figure 1.

Figure 1
The NEI Staging Database Process Overview



<u>Import</u>

The import processes are handled by a Microsoft Access tool, described earlier. Acceptable imports formats are Microsoft Access NIF format or XML encoded ASCII files. The import application loads the data into the Oracle-based system.

Transfer

The Transfer tables accept the data from outside sources into the system. In the Transfer tables, data are subjected to summaries and counts to ensure the imports worked to completion. Preliminary submittal flag process is performed and data are moved into the Transaction tables where more intense data processing begins.

Transaction

The Transaction tables are where the bulk of the data processing on the NEI data takes place. There are a set of tables contained in the database which provide full audit capabilities of every change at each stage, automatically. Oracle triggers have been established on each table such that when a record is inserted, updated or deleted the old record is written to the audit tables. All insertions, updates or deletes are also logged so that changes can be undone in a particular sequence, if necessary. At the Transaction stage, all data are run through an iterative process of QA/QC and standard data scrubbing and ad-hoc data scrubbing until the QA/QC results are either clean or contain data problems that can be repaired by the data submitters, such as invalid codes or missing data. Once all of the data anomalies have been identified, the data are transferred to the Staging tables, and the submitting agencies can submit comments in the form of NIF transactions to add to, change or delete previously submitted data based on the findings of their original data.

Once the agency comment period ends, all update submittals are introduced into the system and exposed to the same QA/QC rounds as the original submittal. Once data are deemed acceptable, the are applied to the Staging tables through the processing of the submittal flags.

Staging

In the Staging area, the data are augmented with electric generating utilities data, ammonia emissions, $PM_{2.5}$ emissions, daily emissions are calculated and selected data from past inventories are used. Many different emission summaries and reports are generated and verified against previous inventories. Examples of some of the summaries and reports are emissions by pollutant by state, emissions by pollutant by county, change in emissions by pollutant by state, change in emissions by pollutant by county, tier reports by pollutant by year and tier reports by year.

Export of the data is performed from the Staging tables through the Access import/export tool to generate data files for distribution to agencies.

STREAMLINING

The process of handling and validating all of the NEI data has been refined each time the process is performed. Selected scripts are rewritten with greater performance efficiency, and additional quality checks and data scrub routines are built into the system. A system for versioning the data between releases has been developed so that a distinguishing internal version number denotes the difference between two releases of the data. The documentation contains the details of each change, the entity requesting the change, the data the change was requested, the date the change was applied, who applied the change, what script was run to apply the changes and the name of the person(s) who performed quality assurance on the changes. The documentation step is contained in a separate database system that handles the tracking of all changes and updates to the NEI data. Additionally, the use of automated audit trails and logging has increased the efficiency of undoing changes and recovering from and hardware failures.

Data for all emission sectors (point, area, onroad and nonroad) and all submitting agencies are processed in the general same time frame. A base year plus two additional grown years of data are usually compiled to make up a data version of the NEI. Since there are millions of records for each sector, there is a potential for many bottlenecks in the process. The NEI Staging Database Process was designed and developed to facilitate the efficient handling of large amounts of data, reduce bottlenecks and increase performance. Since introducing the process into Oracle, the time to process a submittal has decreased over 50 percent while at the same time the number of diagnostics has increased dramatically.

Even with all of the software and database enhancements to the process, there are issues that remain that reduce the efficiency of the overall system. Consistent use of the submittal flag, use of a consistent submittal format, adhering to specified lookup tables and adhering to mandatory field designations in submittals would greatly increase the efficiency of the process. Even though the NIF is somewhat dynamic, the use of the most current NIF available in MS Access shell is greatly encouraged to reduce data reformatting and compatibility issues.

PLANNING FOR THE FUTURE

The EPA plans to release NIF Version 3.0 in April of 2003, scheduled before the release of this paper. This new release encompasses changes from comments received on NIF Version 2.0 as well as incorporates the EPA Federal Data Standards for specific fields and data types. EPA is developing XML capabilities and central data submission repositories for all submittals. The XML formatted NIF files will help to increase the efficiency and compatibility of the transfer of data from submitting agencies to the Oracle data processing software.

Plans for additional data checks and cross-checks are now being implemented so that the inventories of the future will be the most reliable for modeling, reporting and analyses.