Warehouse with Many Floors and Many Doors:

Data Warehousing and Data Mining at the Iowa Department of Revenue

Prepared By
Mike Lipsman
Tax Research and Program Analysis Section
Iowa Department of Revenue

FTA Revenue Estimating and Tax Research Conference
Portland, OR – September 17 – 20, 2006

Introduction

The Iowa Department of Revenue (IDR) maintains a number of information systems on different hardware platforms (floors) and provides a number of software solutions (doors) for accessing, querying, and analyzing the data (Figure 1). The most comprehensive system is the Department's Integrated Revenue Information System (IRIS), which consists of a large transactional database and associated applications that can display individual taxpayer information, check tax return computations, assess penalties and interest, issue refund warrants, and generate correspondence. Second in size is the Compliance Division's Enterprise Data Warehouse (EDW) which supports the work of the Department's Tax Gap Program. The third system, the Department's smallest and most recent addition, is the Tax Research and Program Analysis Section's SAS Business Intelligence Server, where we are in the process of building an analytical data warehouse.

This paper consists of four sections. The first section provides a short history of the evolution of information systems within the Iowa Department of Revenue. The second section sketches pictures of the three information systems previously identified and what functions they serve. The third section describes some applications that have been developed or that are being developed in the new Tax Research BI Server environment. The final section discusses the challenges faced in implementing the Tax Research data warehouse and speculates on future changes in the nature of tax research activities that may arise from new capabilities provided by the data warehouse.

Background: The Evolution of Information Systems at the IDR

In the beginning there were what are now referred to as the Legacy Systems – AWWD, INTX, INKO, RCTX, RCET, SLTR, WHNA … They were written during the 1970s using COBOL and CICS (Customer Information Control System) software. Design work on the current main transactional database (IRIS) began during 1991. IRIS was developed by IDR staff over a 14 year period. The system consists of two major components, Registration developed between 1992 and 1995 and Transactions. The Transactions component includes elements for each of

the major taxes administered by the Department. The tax transaction elements went live at different times between 1995 and 2004: corporate income tax (1995), motor vehicle fuel tax (1996), individual income tax (1997), sales and use taxes (2002), and withholding tax (2004).

In 1997 the Iowa General Assembly authorized the Department of Revenue to initiate a public-private partnership for the purposes of "identifying nonfilers of returns and nonpayers of taxes." In addition, the legislation authorized the Department to cover the cost of this partnership from funds generated by the enhanced compliance program rather than from annual appropriations.[1] This self-funding mechanism provided the resources for the development of the Department's Tax Gap Enterprise Data Warehouse (EDW).

In November 1999, the Department entered into a three-year partnership with NCR-Teradata to design, develop, and implement a data warehouse solution. Given that the program was self-funded the private sector partner placed a premium on generating revenues quickly. The Department recognized the first revenues from this program in April 2000. Since its inception the Tax Gap program has generated over $71 million.

This past year the Department's Tax Research and Program Analysis Section began building its own data warehouse. This was facilitated by the installation of a dedicated SAS Business Intelligence (BI) Server. Tax Research staff had been using PC-SAS on their desktops since 2002 and prior to that time they used SAS on the State's mainframe computer. Some of the reasons behind the move to the server environment include improved data security, improved data management, and enhanced application development capabilities.

In addition to the three information systems just described, the Department maintains a number of other databases on a variety of hardware and software platforms. Legacy systems still remain in use for some of the smaller taxes, such as the bank franchise tax, for the management of accounts receivable billings and collections, and for property tax declaration of value information. The Department's electronically filed return database for individual income tax is maintained on an MS SQL server as well as is a new statewide tax credits tracking database. Furthermore, various work units in the Department have developed databases and applications in

[1] House File 266, 77th General Assembly, 1st Session, codified as Chapter 421, Section 17, Subsection 23, Iowa Code (2005)

Access, Excel and Visual Basic for such purposes as tracking legislation, tracking return and payment processing activities, and documenting local option tax distributions to cities, counties, and school districts.

Information System Platforms and Software Applications

The Integrated Revenue Information System (IRIS)

IRIS is maintained on the State's IBM-z/OS mainframe computer system. The IRIS database consists of a hierarchical structure created using the Computer Associates IDMS (Integrated Data Management System) software with screen display applications created using ADS/O (Application Development System/Online). (Figure 2.)

A major advantage of IRIS relative to the systems it replaced is its "entity-based" structure. Rather than managing data for each revenue source separately taxpayer entities (individuals and businesses) can be linked across revenue sources in IRIS. The registration component of the system provides the common taxpayer identification information that supports this linkage.

As stated previously, IRIS is a transactional system. Its primary purpose is to serve those Department personnel involved in the processing of tax payments, the examination of tax returns, corresponding with taxpayers, and responding to taxpayer inquiries. The registration component of the system contains information on 4 million individuals and businesses. The transactions component contains over 16 million records. Multiple years of transaction data are maintained in IRIS with the number of years varying by revenue source. Currently, IRIS contains about 62 gigabytes of data.

IRIS serves as the primary source of State taxpayer, tax payment, and tax return data for both the Tax Gap Enterprise Data Warehouse and for the Tax Research Data Warehouse. The data warehouses are refreshed and updated on different frequencies. To the horror of many this means not only does the Department have three different sets of much of the transactional data, but in addition the datasets are not coordinated in terms of their update frequencies. However,

given the differences among how the databases are used this is not as big a problem as one may suspect.

The Tax Gap Enterprise Data Warehouse (EDW)

The Tax Gap Enterprise Data Warehouse began as a joint effort involving several departments of State government. Reflecting this broader purpose its official designation is just the Enterprise Data Warehouse. Enterprise in this sense applies to all of State government. However, because of budgetary problems experienced by other departments, the Department of Revenue's unique self-funding mechanism, and the large volume of data the Department desired to load to the warehouse to support enhanced compliance activities the Tax Gap Program has become the dominant user of the EDW. Also, initially within the Department of Revenue the EDW was intended to serve a broader customer base than just the Tax Gap Program. However, the demands placed on the staff charged with managing the EDW by the Tax Gap Program and the fact that the Department's share of EDW costs comes solely from Tax Gap Program generated revenues has limited usage outside the Compliance Division.

The EDW, which holds about 100 gigabytes of data, is a relational database housed on a Teradata Server. The Teradata Server and associated operating system software were replaced during FY 2006 at a cost of about $1 million. The Department's share of server related hardware upgrades and software license fees are covered by Tax Gap Program generated revenues. Sixteen examiners, two consultants, one support person, and a manager are funded from Tax Gap Program revenues. In addition, over 100 other Compliance Division staff access reports generated from the Tax Gap EDW. Also, all audit programs are now processed through the EDW. Program costs totaled $3.2 million during FY 2006. In comparison $16.3 million in additional revenue was generated by the program during FY 2006.

At the present time data from eight major sources have been loaded into the EDW. These sources are:

- The Department's Integrated Revenue Information System (IRIS)
- Human Resources payroll data for the Iowa Department of Revenue
- The State of Iowa Financial Accounting System (I/3)

- Iowa Department of Workforce Development unemployment tax files
- The Department's Accounts Receivable System
- The Department's Collections System
- Internal Revenue Service business tax files (BMF/BRTF)
- Internal Revenue Services individual tax files (IMF/IRTF/IRMF)

The EDW supports three major applications:

- Lead Generation -- By applying specific business rules and criteria, the EDW provides a list of audit leads on taxpayers who have not complied with Iowa tax laws. These leads are loaded from the data warehouse into a Web-based Audit Component application for further investigation.
- Audit Component Application – This is a case management application on a separate hardware platform from the EDW. However, this application uses the Teradata database. Potential non-compliance cases are assigned, worked and tracked by examiners and auditors using this system. Data from the EDW is used to calculate taxes due.
- Value Added Application – Through query development, the value added application brings together data from the various sources stored in the EDW to serve as a single view which supports the identification and working of audit leads. These queries are written in either Teradata SQL or Business Objects.

The two consultants coordinate the loading of data extracts from IRIS and other sources; they design the Business Object universes; and they develop queries and reports used by Tax Gap Program examiners to identify audit leads.

Initially, the Tax Gap Program focused primarily on identifying corporate nonfilers. This effort used federal tax information, state tax information, employment data, and other private sector data to identify corporations that had economic or physical nexus in the State but that were not filing Iowa corporate income tax returns. Similar data mining efforts have resulted in the identification of sales and use tax and individual income tax nonfilers.

More recently attention has turned to identifying individuals and businesses that appear to be underpaying their taxes. These efforts have involved segmenting taxpayer populations using NAICS codes obtained from Department of Workforce Development unemployment tax data. Then, employing various sources of information on different types of businesses, expected tax liabilities are computed and compared with tax liability amounts reported on filed returns.

Over the six years the Tax Gap Program has been in place $23.4 million in additional corporate tax revenue has been generated, as well as $9.0 million in additional sales and use tax revenue and $32.1 million in additional individual income tax revenue.

The Tax Research Data Warehouse

During July 2006 the Department installed SAS BI Server software on a separate server dedicated to databases and applications developed and maintained by the Department's Tax Research and Program Analysis Section. The hardware and operating system software platform consists of a Hewlett-Packard DL380 server with dual Pentium 4 processors, 3 gigabytes of RAM, and 200 gigabytes of disk storage running the Windows 2003 Standard Server operating system. The hardware and operating system software cost approximately $10,000. The SAS BI Server software installation was customized to our specifications. The SAS software components that were installed include:

- SAS Business Intelligence Server Tier
  - Base SAS
  - SAS Graph
  - SAS Stat
  - SAS ETS
  - Metadata Server
  - Integrated Technologies – Workspace Server
  - Stored Process Server
  - SAS Connect
  - SAS Access for PC File Formats
  - SAS Access for ODBC

- SAS Client Tier
  - Enterprise Guide
  - Add-in for Microsoft Office
  - Information Map Studio
  - Management Console

The initial SAS software costs totaled $70,000, plus $10,000 for installation consulting services and $9,000 for training. The annual license renewal cost is expected to equal about $14,000. However, this is not all new costs because we have been able to eliminate about $20,000 in annual PC-SAS license renewal costs.

A case can be made that the EDW should be able to support both compliance and tax research types of data analysis, but this proved to not be the case for various institutional and technical reasons. The reasons the Department decided to pursue the SAS BI Server solution for tax research activities are as follows.

First, the large amount of data stored in the EDW requires separate universes, which represent virtual subsets of the data, be developed to facilitate the efficient querying of the data.

The data elements defined for each universe are customized to satisfy different Tax Gap Program needs. Due to the heavy demands the Tax Gap Program places on the universe and application developers little time remains to serve the needs of the Tax Research Section. We found the universes designed to support the Tax Gap Program did not work well for tax system analysis and fiscal and economic modeling purposes.

Second, the Business Objects client software solution adopted for the EDW works well in an environment where a few power users develop all the queries and the remainder of the EDW users run standard reports generated from the queries. The types of issues confronted in analyzing tax policy, developing economic and revenue forecasts, and estimating fiscal impacts for proposed legislation require that each analyst has the ability to perform ad hoc queries and develop their own applications.

Third, the work done by Tax Research Section staff requires economic and fiscal modeling and statistical analysis capabilities that exceed those of the Business Objects software selected for the Enterprise Data Warehouse. Also, familiarity with SAS products by Tax Research Section staff influenced our selection of the SAS BI Server solution for the analytical data warehouse.

Fourth, the demands placed on tax research staff, particularly during the legislative sessions, require staff to multi-task. Also, expectations for the rapid turnaround of requests for information and analysis are ever increasing. Four or five years ago legislators and the Governor would allow a day or longer to turnaround requests for the analysis of proposed legislation. Now we are often only allowed a few hours to respond to requests. Moving the execution of our individual income tax micro-simulation model and other SAS applications from our desktops to the BI Server has reduced run times to around two minutes for even our most complex programs. As an added benefit we are able to continue working on other tasks on our desktops while the models are executed. Previously, when the models were running on our desktops other software applications could not run at the same time.

Fifth, the desktop client software that is packaged with the BI Server (Enterprise Guide) provides a variety of productivity features that allow Tax Research staff to work more efficiently

and to develop applications that may be shared with legislative staff and staff in other departments of state government.

Tax Research Data Warehouse Applications

Currently, the Tax Research BI Server houses about 70 gigabytes of data. Much of the data replicates data contained in the Tax Gap EDW. However, the structure of the data is different. Most of the data is stored as SAS files. In most cases the structure of the files differs from the EDW files in terms of the variables they contain as well as in terms of the metadata stored with the files, such as the variable names, types, formats, labels, etc.

The data stored on the BI Server support a variety of tax research activities. These include:

- Estimating the fiscal impacts of proposed law changes.
- Making economic and revenue forecasts.
- Developing statistical reports.
- Responding to ad hoc requests for information.

In addition, to warehousing data the BI Server software provides the capability for accessing data stored on other hardware platforms and in different formats. Mainframe files can be accessed using SAS Connect. Access, Excel, text, and other PC and LAN files can be accessed using the SAS Access for PC file formats utility. Data stored on SQL servers can be access using the SAS Access for ODBC utility.

The Department's individual income tax micro-simulation model is the most complex application run on the BI Server (Figure 3). The database used in this application consists of variables derived from the State's individual income tax master file and two years each of Internal Revenue Service IMF (Individual Master File) and IRTF (Individual Return Transaction File) files and their associated non-resident "tickler" files. Standard parameters, such as tax rates, bracket amounts, credit values, filing thresholds, etc., are passed to the model from an Excel table. Output reports that summarize impacts by level of adjusted gross income get passed back to Excel. A version of this model that will run against a synthetic database rather than actual return

data is being developed. Our intent is to compile this version of the model as a stored process and provide it to legislative staff for their use.

Other advantages of the BI Server/ Enterprise Guide solution include the software's point-and-click programming, project management, and project documentation features. Recently Tax Research Section staff developed an Iowa Leading Indicators Index (ILII) and implemented it using Enterprise Guide (Figure 4). This application illustrates the self-documenting feature of the Enterprise Guide software. Each icon in the project diagram represents a dataset, a query, a program module, or output. Data that is imported into the project can either be maintained in its native form or converted to a SAS dataset. Each icon can be opened to reveal the underlying data table, query structure, program code, or report. The final output can be exported to Word, Access, or Excel, or converted into a pdf (Acrobat Portable Document Format) file for publication. In addition, the BI Server software supports Web publishing, but we do not currently use this feature of the software due to some security concerns.

Working effectively in this environment does require staff to have basic SAS programming knowledge, even though many standard data manipulation and statistical analysis procedures are available as pre-programmed routines. These applications can be combined into projects with various data sets to generate common types of statistical output, such as one-way frequency distributions, two-way cross tabulations, regressions, and forecasts. Furthermore, wizards exist for many of these applications that allow a certain degree of customization. When the application executes the underlying SAS code is generated, which can then be modified and customized further. But such additional customization requires staff to possess knowledge of the SAS programming language.

Another advantage of having a data warehouse dedicated to tax research is that this arrangement allows activities that previously required considerable work on the part of the Department's information technology staff to be assumed by Tax Research Section staff. For example, the existing programs used to produce quarterly and annual sales and use tax statistical reports and annual individual income tax statistical reports need updating. These programs are

written in COBOL and they have not been updated to any great extent since the 1970s. The effort required to update these programs would be substantial. Also, many of the information technology staff that wrote these programs have retired or are nearing retirement. So, rewriting the programs the traditional way would require considerable reeducation for information technology staff. The Tax Research Data Warehouse and associated software now make it possible for Tax Research staff to redesign these reports on their own with minimal support from information technology staff.

Furthermore, given the ease with which the SAS programs can be modified it is more likely the statistical reports will be updated more frequently to better satisfy the changing information needs of their audiences. For example, the highest income range for which information is provided in the existing individual income tax statistical report is $100,000 and over. Even in Iowa this is a fairly low top income range for statistical reporting today. Redefining income ranges in COBOL programs is a major undertaking. In SAS this takes a half hour or less to complete.

One of the greatest benefits of having a data warehouse designed to meet the needs of tax research is the increased ability to respond quickly to ad hoc requests for information. Common types of requests we receive include:

- How many taxpayers claimed a given tax credit each of the past five years and how much was claimed?
- How many taxpayers earned over $1 million last tax year?
- So far this processing season how do wage, interest, dividend, and farm income compare to last year for taxpayers that filed returns both years?
- What has been the amount of taxable sales for a given type of business in a five county area each of the past four quarters?
- How much have consumers spent on motor fuel purchases each month for the past three years?

The Enterprise Guide software through its pre-programmed procedures and wizards makes responding to these types of ad hoc requests relatively easy. Also, the SAS Information Map Studio, which is packaged with the BI Server / Enterprise Guide software, can be used to simplify responding to information requests that are of a similar nature. This utility is a Java-based

application for creating and managing metadata. This metadata layer enables users to query data without extensive programming or database structure knowledge.

A final capability we hoped to obtain in establishing the Tax Research Data Warehouse was the improved delivery of information to our end customers – other state government departments, local governments, legislators, media, and the public. The SAS OLAP (Online Analytical Processing) Server will provide this capability. However, due to Departmental security concerns we are not yet making use of this feature of the software.

Challenges and Outlook

Now that the Tax Research Section possesses its own data warehouse and the new server and client software are installed, our primary challenge is learning how to use all of the software's features effectively and efficiently. To accomplish this we have undertaken a multifaceted training program. First, in order to get all staff up to a base level of competency in the use of SAS a self-paced tutorial package was licensed for a year. Second, after completing the tutorial training, staff members that felt they would benefit from additional instructor led training were sent to intermediate training courses at SAS training centers. Third, all Tax Research staff attended an on-sight two-day training course in the use of the Enterprise Guide client software. Fourth, the end of October all Tax Research staff will be attending four days of on-site training in the use of the more advanced features of the Business Intelligence Server/ Enterprise Guide software (i.e., BI Server Overview, Microsoft Office Add-In, Information Map Studio, and Stored Processes). In addition, we expect to continue sending individual staff members to SAS training centers for advanced courses that meet needs associated with their specific job responsibilities. The cost of all of the training through the end of this calendar year will be about $30,000. One third of this amount was included in the installation and first year licensing cost for the BI Server/ Enterprise Guide software. The other two-thirds of the cost has been covered through the purchase of a package of 60 training units (days) through the SAS Enterprise Professional Training Program.

Another major challenge is getting organized. Our seven person section already has over 70 gigabytes of data residing in over 6,000 files loaded into the data warehouse. Keeping track of all the files, documenting their content, and weeding out the "junk" is a gigantic task. Accomplishing this has required us to designate one staff member to oversee the management of the data files, projects, and programs.

A third challenge involves identifying how we can better meet the information needs of our customers. Overcoming concerns about confidentiality presents a major impediment in this area. For example, one of types of information for which we receive frequent requests is retail sales information. Those making such requests often want such information for a particular type of business in a specific geographic area. Due to the frequency of this type of request it would be desirable to create an application that would allow customers to drill into our database and generate their own custom reports. However, if the query gets too specific in terms of business type or if it focuses on too small of an area it likely would yield information that violates the confidentiality of individual businesses.

The demands for more and better information are constantly increasing. Some of the reasons for this are:

- The realization that information is a strategic asset for State and local efforts to promote economic growth.
- The desire to better anticipate the fiscal impacts arising from fluctuations in the national and State economies.
- The need to develop more accurate and productive ways to identify all tax revenues that are owed to the State and to local governments.

The Tax Research data warehouse and the associated BI Server/ Enterprise Guide software provide us with the means to better satisfy these demands. These resources allow us to streamline work processes so that we can respond to information requests more quickly and in many cases more completely because we have access to better and more complete data. Furthermore, staff members are able to work more efficiently and undertake more sophisticated types of analysis than in the past.

So, where do we go from here? First, there remains the desire to add further analytical capabilities. One capability we desire to add is greater flexibility compiling information by location

using geographic information system (GIS) software. Second, as alluded to previously, we desire to provide the means for customers to query certain data themselves provided confidentiality concerns can be overcome. Third, the Tax Research Section has proposed assuming a more active role within the Department in providing measures of the State “tax gap” and in helping identify audit candidates.

In conclusion, the types of analysis tax researchers are asked to undertake requires access to a great variety of data. In many states, like Iowa, the data is probably maintained on different information systems. The types of analysis and the short amount of time often allowed to complete studies or to respond to requests for information require ready access to the data. Having to access the data in its native environment can be a major impediment to meeting the expectations of those making study and information requests. Thus, having a data warehouse dedicated to tax research can be a significant asset. Furthermore, combined with the appropriate software a separate tax research data warehouse can enhance staff productivity, increase the value of the information content derived through data analysis, broaden the scope of research activities, and increase the availability of information to policy-makers and the public.

Figure 1: IDR Information Platforms

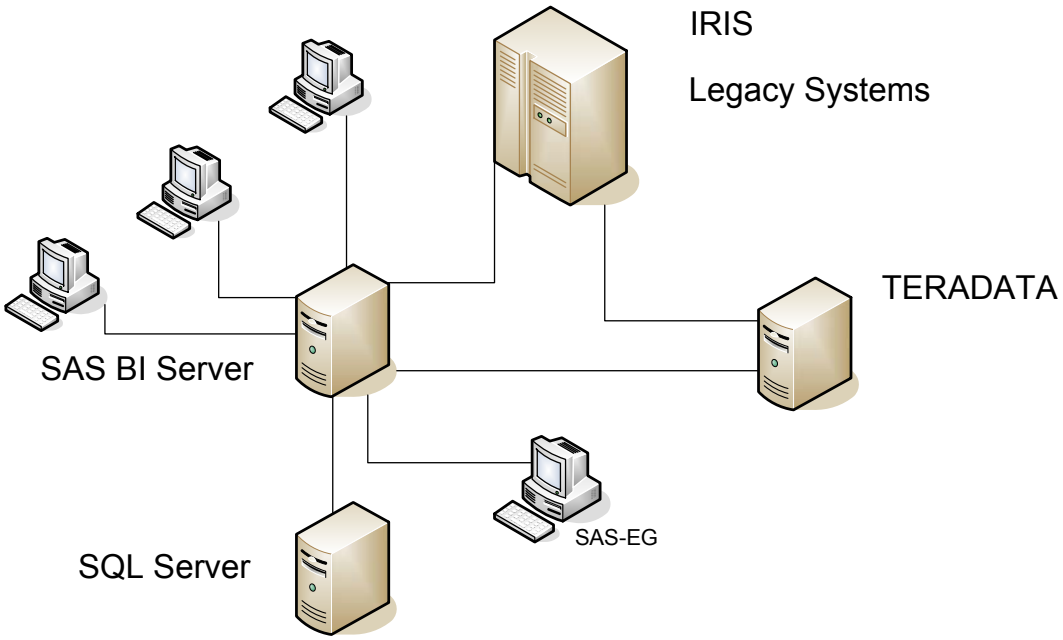

Figure 2: IRIS Hierarchical Data Structure

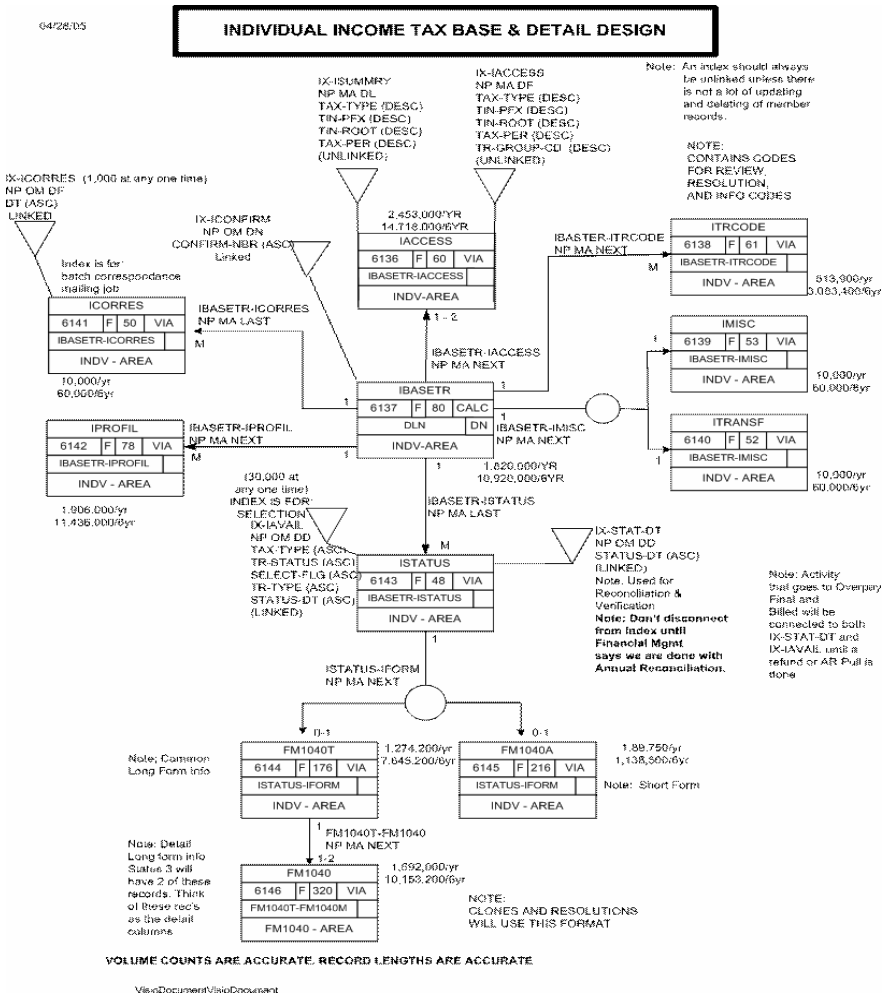

FIGURE 3: INDIVIDUAL INCOME TAX MICRO-SIMULATION MODEL

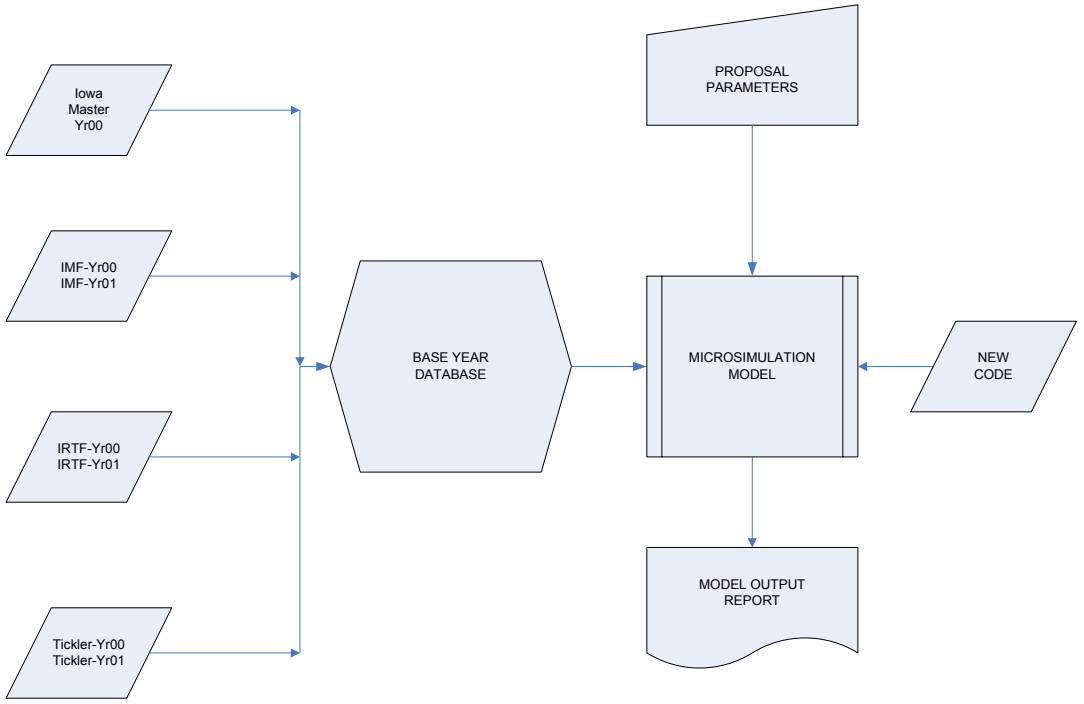

Figure 4: ILII Project Schematic Diagram