Create a SQL-Based Data Warehouse

End Product

Build a SQL-based data warehouse or data depository that integrates data from multiple sources into a single database for data mining.

Downloadable Documents: (Click on items below to download)

- NC Data Warehouse Description & Visualization
- NC Data Warehouse: Final Project Proposal
- NC Data Warehouse: Work Plan and Milestones
- NC Data Warehouse: Data Dictionary
- NC Data Warehouse: Look Up Tables
- NC Data Schematic and Linking Instructions
- NC ACIS Schematic and Linking Instructions
- NC ACIS Data Relational Schematic
- NC Data Warehouse: User Documentation
- NC Data Warehouse: Training Videos
- Travis County, TX Data Dictionary
- Travis County, TX Visual Schematic and Linking Instructions
- Travis County, TX Data Architecture Description

Operational Definitions

- **Data Warehouse**: A data warehouse or data depository is the technological infrastructure used to house large amounts of data. A data warehouse typically integrates data from multiple sources into a single database for data mining.
- **Big Data**: “Big Data” is an industry term in information technology referring to large, complex data sets. It references the programming skills and experience required to build platforms that can process and handle large data sets efficiently with fast response times.
Data Warehouse Project Overview

The purpose of a data warehouse is to house, clean, and integrate data from multiple sources if necessary. A data warehouse will provide access to multiple users simultaneously and automate many data production steps, thus saving time and labor. There is no one way to create a data warehouse. There are a variety of programming tools available with which to build a data warehouse, from SQL-server to open source programs, and which you select will depend on your existing data infrastructure and technical expertise of your staff. NCSEP elected to build a Microsoft SQL Server data warehouse because North Carolina’s court system uses SQL Server, but your IT contractor can help you make an informed selection.

Data Warehouse Elements

To be over simplistic, there are two parts to a data warehouse: 1) the server program itself, which is a collection of program routines that integrates the data into a single database, and 2) the user interface that provides access to the data. Again, there are many options for how to do this and you will work with your IT programmer to identify what will work best for you. NCSEP utilizes IBM SPSS statistical software to analyze data. Consequently, NCSEP designed our data warehouse to interface with SPSS. If we want to change the underlying structure of the data warehouse, i.e., add a new variable, NCSEP will use Microsoft SQL Server Management Studio. Management Studio is free software designed specifically to work with SQL-Server databases.
**Key Project Steps**

Listed below are the key project steps to creating a data warehouse:

1. Identify project staff and resources
2. Assemble data warehouse background information
3. Identify potential IT contractors or in-house IT programmer
4. Interview potential IT Contractors
5. Develop contract with identifiable deliverables and project milestones
6. Work with IT contractor to create data warehouse
7. Test data warehouse

The work plan outlined in this toolkit is a generalized work plan based on our experience. The project work plan in your jurisdiction may differ in the details depending on individual circumstances.

**Identify Project Staff and Resources**

*Project Time:* The total time duration for the project is approximately 10-12 months.

*Staff*
Managing the project is estimated to require 20% of a FTE position for 10 months based on NCSEP’s experience. The project requires varying levels of intensity. There will be stretches where staff will devote 40 hours per week and stretches where little or no labor is needed. Select staff with the flexibility in their work schedule to accommodate these peaks and valleys of intensity. Staff should have both research and database design experience if at all possible. If staff lack this experience, you can expect both the length and cost of the project to increase.

*Costs*
Below are some direct project costs to anticipate. We assume you have access to a server that will host the data.

- **SQL Server Software:** A single license costs around $5,000. However, Microsoft makes a free version of SQL, SQL Server Express, which is a more limited version of the software (including limits as to number of records it will house). ([http://www.microsoft.com/web/platform/database.aspx](http://www.microsoft.com/web/platform/database.aspx))

- **Computer and Monitor:** Powerful computers with high levels of RAM and fast multiprocessing speed, such as a Dell T3500 Work Station, cost approximately $2,500 to $3,500 each.

- **IT Programmer:** The NCSEP data warehouse took an IT programmer 318 hours to build at $150 per hours for a total cost of just under $60,000. This included documentation, training manuals, and a small number of instruction videos. The IT contractor estimates that having the documentation, schematics, list of deliverables, and sample code will significantly reduce the cost of replicating the data warehouse in other jurisdictions, between 20% and 50% depending on the extent to which your project differs from the NCSEP project. The NCSEP data warehouse was on the more complex side, requiring the integration of more than 15 separate data tables into a single database. The NCSEP data warehouse was built by Take Note Technologies, [www.takenote.com](http://www.takenote.com), and we could not have been happier with their services. The IT contractor was paid on an hourly basis.
• **SPSS Statistical Software**: A single user license of SPSS Base plus SPSS Custom Tables costs approximately $3,000 per license for government agencies. (http://www-03.ibm.com/software/products/en/category/business-analytics)

**Assemble Data Warehouse Background Information**

Before you meet with any IT contractor, you should have the following information assembled:

1. **Data**: Identify all data sources to be integrated into the data warehouse, both current and future data. Identify the data you currently have access to, as well as data that you would like to add to the data warehouse in the future. It is important that the IT contractor be aware of future data sources, so that the data warehouse will be built with ability to grow and expand. NCSEP only had access to two of the desired data sources. To accommodate future expansion, the IT contractor built into the NCSEP data warehouse the capacity to add tables as well as variables and developed training videos to walk NCSEP how to modify the data warehouse program once these data sources become available.

2. **Data Storage & Growth**: The size of your data warehouse, including data warehouse backups, will depend on the size of your data sets and growth projections over time. Identify the number of records to be uploaded to the data warehouse initially as well as an estimate of the number of records that will be added to the warehouse annually. You also will need to find out how much “space” is available on the server that will host your data and provide this information to the IT contractor.

3. **Variable Spreadsheets**: For each data source, create a spreadsheet(s) of every variable in the data set, including the variable format, width, type, codes, etc. (see NCSEP downloads for an example).

4. **Data Relational Schematic**: Develop a schematic that shows how the data is to be linked if you have more than one data source or a data source with multiple tables (see NCSEP downloads for an example).

5. **Data Mining Software**: Identify how you will analyze your data, i.e., what program will be used (SPSS, SQL, SAS, etc.). You will need to verify with the IT contractor how the program will work with the data warehouse.

6. **Number of Users**: Identify which users will have access to the data warehouse and whether you have any security issues that will need to be addressed, such as limiting some users to “Read Only” access, etc.

7. **IT Contractor Deliverables and Data Warehouse Visualization**: Create a list in layman’s terms of what you want the data warehouse to do for you and how it will fit into existing workflow processes (see NCSEP downloads for an example). This list will be the blue-print for defining IT contractor “deliverables.” In addition, create a flow chart or graphic that visualizes the deliverables. The key to acquiring the data warehouse that you need is a thorough understanding on your part of what the data warehouse will do.

**TIP**: One of the major challenges in IT projects is developing a shared understanding of what you want with the IT programmer. The IT industry has a language and terminology of its own that is difficult for lay persons to understand. All deliverables should be described in terms that you understand. It is a mistake to rely on IT terms when drafting the contract and assume that the IT programmer understands what you intend. Although it can be taxing to all parties, keep the conversation going until all deliverables are described in a language that is clear to both you and your IT programmer.
8. **Identify Specialized Needs**: You should build a data warehouse only after you have a thorough understanding of your data and your research needs. You want to build a data warehouse that will support your research needs and this may involve some specialized items not normally seen in the typical data warehouse. For example, NCSEP court system data is collected at the charge level. Most NCSEP research, however, involves analyzing “cases” not “charges.” To allow NCSEP to analyze case information, the IT contractor built a feature into the NCSEP data warehouse that generates a second “case” table containing only case-level variables, specifically to support NCSEP research.

### Identify Potential IT Contractors or In-House IT Programmer

The optimal IT programmer or contractor should have experience with “Big Data” and building databases containing millions of records. “Big Data” is an industry term in information technology referring to large, complex data sets. It references the programming skills and experience required to build platforms that can process and handle large data sets efficiently with fast response times.

### Tips and Lessons Learned in Developing a Data Warehouse

Below are some tips and lessons that NCSEP learned undertaking this project.

- **Milestones, Tasks, and Estimated Hours**: Before you finalize your contract, ask the contractor to provide a list of milestones, tasks, and estimated hours to complete each task, such as the one depicted below. (For the complete document download the NCSEP Project Work Plan and Milestones document above.) NCSEP kept track of the project’s progress and costs with this document. At each milestone, NCSEP was provided a live demo of the data warehouse to review and amend if necessary. We strongly urge you to view the data warehouse as it progresses and not wait for a final product to have your first viewing, because by that point, it will be very labor intensive to make programmatic changes.

<table>
<thead>
<tr>
<th>Milestone 1</th>
<th>Milestone 2</th>
<th>Milestone 3</th>
<th>Milestone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Produce end user requirements</td>
<td>1. Schedule and deliver SQL Server Management Studio training</td>
<td>1. Develop and code schema to maintain link to database</td>
<td>1. Planning, phone calls, email and meetings with IT staff about Milestone 4 items</td>
</tr>
<tr>
<td>2. Produce end user requirements</td>
<td>2. Schedule and deliver SQL Server Management Studio training</td>
<td>2. Develop and code schema to maintain link to database</td>
<td>2. Planning, phone calls, email and meetings with IT staff about Milestone 4 items</td>
</tr>
<tr>
<td>3. Produce end user requirements</td>
<td>3. Schedule and deliver SQL Server Management Studio training</td>
<td>3. Develop and code schema to maintain link to database</td>
<td>3. Planning, phone calls, email and meetings with IT staff about Milestone 4 items</td>
</tr>
</tbody>
</table>

---

This project was made possible by a grant from the Open Society Foundations.
• **Response Time**: Discuss and monitor the response time of the data warehouse with the IT contractor. Response time is how long the data warehouse will take to perform specific tasks, such as upload data into the data warehouse or download data from the data warehouse to a user’s computer. Obviously, response time will have a significant impact on productivity and with very large data sets tasks may take hours to complete. You should know approximate expected response times for various tasks up front and develop processes to deal with lengthy tasks as necessary. For example, NCSEP court system data extracts contain over 15 million records to upload to the data warehouse at one time. This volume of data takes several hours to complete. Because work processes often encounter the unexpected, NCSEP’s IT contractor built in “pause,” “cancel”, “resume”, and “partial” upload functions into the data warehouse for lengthy tasks. In addition, he built data counters for these tasks to allow us to monitor a task’s progress and check for accuracy.

• **Data Warehouse Data Backup Plan**: Be sure the IT contract includes a plan for backing up warehouse data. Data is an investment in both time and money and you do not want to risk losing your data if your server crashes or some other crisis occurs. NCSEP has a dual backup plan in place. NCSEP’s server is regularly backed up by the NC Administrative Office of the Courts which handles this task for all court system actors. In addition, NCSEP purchased a 1TB portable USB flash drive to host manual backups at NCSEP’s discretion.

• **Maintenance and Documentation**: NCSEP did not have the funds or access to technical staff to “maintain” the data warehouse. We recognized that the research staff would be providing this function. Consequently, we built into the contract provisions for the contractor to provide us with detailed documentation and training for maintaining the data warehouse in the future. The IT contractor was asked to provide “descriptions” into the programming code itself so that anyone could read the program code and know the function each section performed. This documentation was to assist NCSEP as well as external jurisdictions who might want to download NCSEP’s data warehouse to tailor it to their purposes.

• **Reserve Funds**: Even with the best of intentions and hard work by all parties, miscommunication or the unexpected can occur. One way to address these potential problems is to have an additional 20% to 30% above the cost of the project held in reserve to allow you to address the unexpected. For example, as the project progressed, NCSEP realized that we needed the capacity to identify “changed” data between different point-in-time data uploads, both in terms of being able to identify and select records that had been changed and being able to view the changes themselves. Having access to “reserved” funds allowed us to build this critical, previously unidentified, feature into the data warehouse.

• **Request a Copy of NCSEP Data Warehouse**: Part of the IT data warehouse contract was the provision that NCSEP owned all programming code and would be provided with a complete copy of the data warehouse program that we could distribute for non-commercial use. If you would like a copy of the NCSEP data warehouse please contact the IDS Research Director at 919-354-7200. The NCSEP data warehouse was built with SQL Server 2008 R2 and is compatible with all later versions of SQL.