

Implementing Our Data Warehouse Roadmap:

It's more than Technology

Minnesota DOR

- **Serve population of about 5.2 million**
- **1287 employees**
 - Approximately 2/3 in direct compliance & collections functions
- **Filers**
 - 150,000 active Sales Tax filers
 - 164,000 active Withholding filers
 - 50,000 active Corporate filers
 - 3,600,000 Income Tax filers
- **\$17.5 billion in tax collections**
 - No motor vehicle, very small property tax
- **Collections arm brought in \$170 million in delinquent taxes & \$21.6 million in non-tax debt**
 - No Child Support

Minnesota DOR Strategic Plan

Vision

- **Everyone pays the right amount of tax**
- **Information is timely, accurate and convenient**
- **Employees have necessary skills, tools, and resources**
- **Revenue system works well, in policy and operation**

Mission

“Make the revenue system work well”

Strategies

- **Focus compliance efforts on those who deliberately evade the tax laws, not on those who make an effort to “get it right”**
- **Measure the effectiveness and cost of activities, and shift resources to those that demonstrate the greatest success in achieving our mission.**

MINNESOTA • REVENUE

MN DOR Warehousing in 2004

3 Warehouses

2 little used for compliance

1994 – Sales Tax Reengineering

1996 – Collections System Warehouse

2003 – Income Tax Reengineering

3 Database platforms

No Integration

No Data Warehouse Strategy

No formal organization inside DOR

Spotty expertise

MINNESOTA • REVENUE

A Plan

- **Initiatives**
- **Develop Data Warehouse plan**
 - Outcome had to be affordable and sustainable
- **Brought in Experts to analyze the situation and make recommendations**
 - Interviewed over 40 MN DOR staff
 - Documented current practices
 - Produced 96 page report with recommendations
- **Presentations to sell ideas**
 - *Who's going to read a 96 page report?*

MINNESOTA • REVENUE

From the Roadmap

A data warehouse environment is...

... driven by needs for business intelligence.

Reporting

Online Analytical Processing (OLAP)

Management Decision Support

Ad Hoc Reports

Data Mining/Predictive Modeling

... supported by technology and organizational components.

MINNESOTA • REVENUE

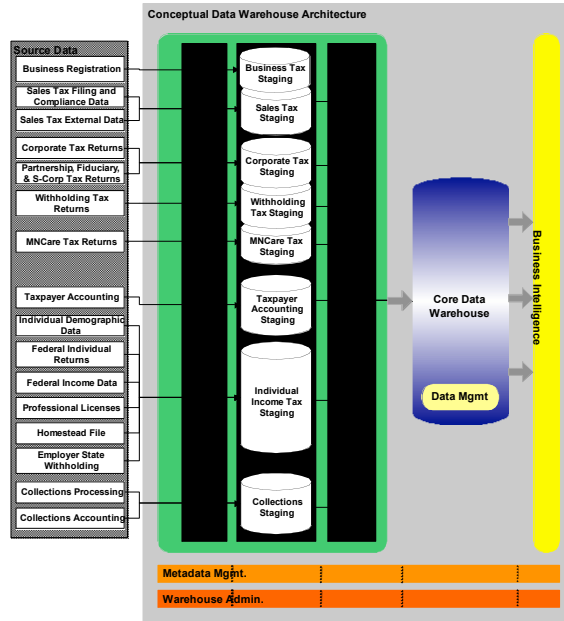
Technology requirements include...

Integration of data
future state is conceptually similar to what DOR is doing today

Storage of data
future state staging and warehouse databases use modeling techniques new to DOR

Access of data
future state recommendation is to implement a metadata layer with the BI tool to facilitate easier user access to data

Administration of warehouse
future state contains administration layers including ETL, metadata, data updates, BI tool(s)



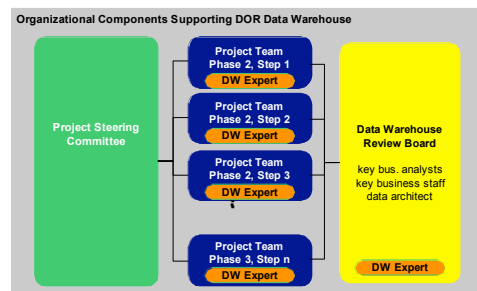
Organizational requirements include...

Data Warehouse Expert whose primary responsibilities are:

- Drive the organization along the course set by the roadmap
- Mentor business and technical users of the data warehousing environment
- Facilitate communication of data warehouse efforts at all levels of the organization

Data Warehouse Coordination Team whose primary responsibilities are:

- Review system and data models
- Establish a process and schedule for reviewing and approving changes submitted by project teams.
- Provide a means of communicating scheduled and/or approved changes to the data warehouse environment.
- Assist with resolution of design and implementation conflicts between data warehouse project teams.



MINNESOTA • REVENUE

Roadmap

Increase Revenue

- Capability to produce more profitable audits
- Ability to identify a wider spectrum of audits
- Ability to more efficiently tap the information currently available
- Increase ability to contact non-filers and under reporters



**Overall
Roadmap
Benefits**

Increase Efficiency

- Enable far more in-depth reporting, including drill down, historical trend analysis, and anomaly identification
- More efficient audit process
- Reduce time spent on non-productive audits
- Reduce erroneous contacts by having more accurate information...rely less on a wide-sweep process
- Reduce redundant processes

Reduce Cost

- Improve accuracy in reporting
- Ability to capture and measure performance metrics
- Provide more visibility and capability in reducing the tax gap
- Make "better use of data" increasing productivity and reducing workload
- Standardize DOR data warehouses on one platform and DBMS

MINNESOTA • REVENUE

Using the Roadmap

- **Created Data Warehouse Steering Team**
- **Temporary/Permanent Data Warehouse Staff**
- **Created Data Warehouse Coordination Team**
- **Purchased Extract, Transform and Load Tool**
- **Began using enhanced Data Quality & Match techniques and technologies**

MINNESOTA • REVENUE

Using the Roadmap

- **Explored options for a primary Business Intelligence tool**
- **Migrated to One Integrated Warehouse**
- **Added multiple new data sources**
- **Pilot Data Mining projects using University of Minnesota graduate students**
- **Started to measure value of Data Warehouse sources**

MINNESOTA • REVENUE

Created Data Warehouse Steering Team

- **Membership**
 - Director of Sales & Corporate Taxes
 - Director of Individual Income Tax
 - Director of Tax Research
 - Director of Information Systems
 - Data Warehouse Coordinator
 - CIO Office
- **A Key to success**
 - Provided visibility, resources, and direction

MINNESOTA • REVENUE

Temporary/Permanent Data Warehouse Staff

- **Started with Contractors**
 - Data Warehouse Coordinator
 - Data Architect
 - ETL Programmer
- **Provide proactive leadership and management of the warehouse environment**
 - 100% of time on warehouse
 - Co-located away from IS division
- **Mentor and collaborate with Business and IT staff**
- **Relationship with IS and business stakeholders is the key**
- **Difficult to fill permanent positions**
 - Classification

MINNESOTA • REVENUE

Created Data Warehouse Coordination Team

- **Primary users/suppliers of the warehouse**
 - 3 warehouse, 3 IT, 1 Security, 16 Divisional
- **Prioritize data warehouse efforts and develop recommendations for the Steering Team**
- **Communicate throughout the agency**
- **Share how the warehouse is used**
 - Improved, more consistent processes
- **Meet bi-weekly**
- **Very successful**
 - Collaboration, better decisions on where to focus resources, elevated view from unit to department

MINNESOTA • REVENUE

Purchased Extract, Transform and Load Tool

- **Primary tool for building the warehouse**
 - Gives the ability to get external data, clean it, match to other data and transform into a form that is more appropriate for queries and reporting
 - Replaces using custom code and stored procedures
- **Long evaluation process**
- **Expensive**
 - Initial cost
 - On-going
- **Great tool – Wacky, changing industry**
- **Tool expertise can be hard to hire**

MINNESOTA • REVENUE

Began using enhanced Data Quality & Match techniques and technologies

- **To find new audit cases, new sources of data, usually external, needs to be used**
 - We usually have little control over the quality of the data we receive from these external sources
 - Have to ensure that one taxpayer's information in existing data sources gets linked accurately to that taxpayer's information in the external data source
 - **Fewer and fewer external sources have SSN**
 - **Have to allow for missing or incorrect SSN/FEIN**
 - **Given the volume of data, human intervention and manual data correction is not possible**
- **How have we done this?**
 - By using data quality & match software
- **What does this software do?**
 - Standardizes address so that all addresses are conformed to a single format (Street vs. St., 1st Ave vs First Avenue, etc.)
 - Parses names into individual components (John Doe is broken out into John as the first name and Doe as the last name)
 - **Uses dictionaries present in software to accomplish this**
 - **Uses advanced algorithms for matching**
 - *The software will indicate "John Doe Enterprise" in one source and "J. D. Enterprise" in another source are very similar, though not exactly the same*

MINNESOTA • REVENUE

Began using enhanced Data Quality & Match techniques and technologies

- **How does this help?**
 - The software can identify one taxpayer's information in multiple sources, even when information is slightly different or has minor mistakes
 - Matches can be made with much more accuracy than basic programming techniques can provide
 - The process can be tuned to require little to no human intervention
 - Now able to match and use data that previously had to be discarded
- **What can the software not do?**
 - It cannot eliminate all mistakes in matching. Some incorrect matches will still occur.
 - Marginal matches may need to be discarded if human intervention is not possible or not cost effective.
 - It needs to be configured and set up, which may require individuals with specific skills

MINNESOTA • REVENUE

Explored options for a primary Business Intelligence tool

- **Goal was to find a tool or toolset that would allow less technical folks to use the warehouse**
- **Extensive evaluation**
 - Many different tools and toolsets
 - Industry keeps changing
 - So easy a monkey can use it – but you better understand the work involved before you give it to the monkey
- **We have a winner – Maybe not**
 - First choice turned out to be too expensive
 - Second choice bought our ETL tools data quality tool, then they were bought by a large ERP
 - Put purchase on hold
- **At the recommendation of the Warehouse Coordination Team we created an in-house SQL class (ad hoc queries)**
 - Over 100 employees took the class
 - Knowledge of SQL helps in most tools
- **Still pursuing our original goal**

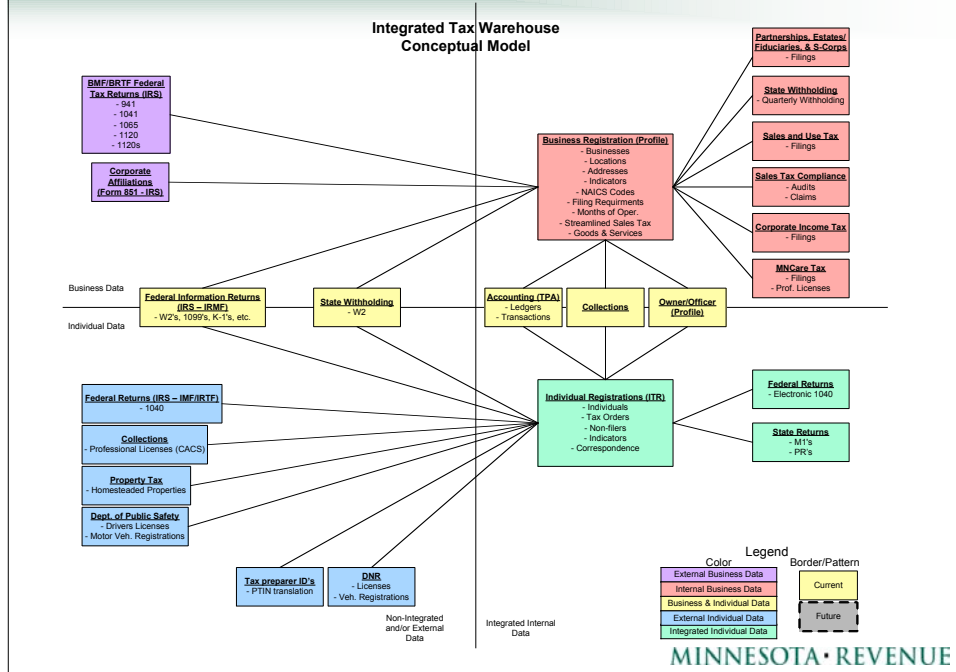
MINNESOTA • REVENUE

Migrated to One Integrated Warehouse

- **Using ETL tool we migrated our Business Tax and Collection data warehouses to DB2 database**
 - One platform
 - Similar database design
 - Enhanced naming conventions
- **Greater value**
 - Easier to use
 - **Develop answers faster**
 - More likely to try something new

MINNESOTA • REVENUE

Added multiple new data sources



Pilot Data Mining projects using University of Minnesota students

- **Worked with Professor Jaideep Srivastava of the University of Minnesota**
 - A well known expert in the field of data mining
- **Over the course of 14 months we had the opportunity to do data mining pilot projects with the help of 5 University of Minnesota PhD candidates working under the professors tutelage.**
- **Used the data mining techniques for :**
 - Individual Income tax,
 - Sales and Use tax,
 - Corporate tax and the
 - Partnership/Estate/Fiduciary/S-Corp area taxes.
- **Requires very specialized tools, knowledge, skills and experience requiring a significant commitment of resources for many years.**
- **We often lacked the needed data or quantity of needed data to create good models**
 - A couple hundred audits is not a particularly good data set.

MINNESOTA • REVENUE

Sales & Use Pilot

- **Goal**
 - Find productive smaller cases for entry-level auditors and to reduce our number of no change audits.
 - We defined a productive audit as an audit resulting in an assessment of at least \$500 per year, or \$1500 per case.
- **Categories**
 - Large Use & Large Sales
 - Small Use & Small Sales
 - Large Use
 - Small Use
- **Removed the usual suspects**
 - Fortune 500
 - Top 200 per region

MINNESOTA • REVENUE

Sales & Use Pilot Results

Results in % for All Categories	Pre Data Mining Avg Success Rate	Data Mining Predicted Success Rate	Actual Success Rate
Sales	29%	38%	37%
Use	39%	56%	51%

MINNESOTA • REVENUE

Sales & Use Pilot Results

Results in \$ for All Categories	Pre Data Mining Avg Dollars	Data Mining Predicted Dollars	Actual Dollars
Sales	\$6,497	\$11,976	\$8,186
Use	\$5,019	\$8,623	\$10,829

MINNESOTA • REVENUE

Sales & Use Pilot Results

Results of 414 Audits	Overall Total Assessed	Overall Average Assessed
Large Sales & Use	\$1,399,436	\$19,437
Small Sales & Use	\$72,605	\$2,504
Large Sales	\$6,229,248	\$23,776
Small Sales	\$101,895	\$1,998
Combined Totals	\$7,803,184	\$18,848

MINNESOTA • REVENUE

Sales & Use Comments

- Data mining can confirm what you think you know about your audit population: first hits came back identifying bars and large businesses as good candidates.
- Don't count anything out: we used every element from other business returns that we could, including Corp, S-Corp, Partnership,
- Don't be discouraged at preliminary results; more refinement will eventually get you where you want to be.
- Our goal was to find productive smaller cases for entry-level auditors and to reduce our number of no change audits. We defined a productive audit as an audit resulting in an assessment of at least \$500 per year, or \$1500 per case.
 - **Data mining identified a large number of small businesses we would otherwise have chosen only by chance.**

MINNESOTA • REVENUE

Income Tax Schedule C Loss Project

- Also known as the “hobby-loss” project.
- Attempt to identify taxpayers who are incorrectly using a Schedule C (i.e. a business) to reduce their taxable income by participating in activities that are not consistent with a business that is trying to produce revenue.
 - One goal of this project was to determine if data mining could reduce the number of unsuccessful audits
- Defined success as an audit that generated at least a \$1,000 assessment.
- Created model and then ran model against audit list created under old process

MINNESOTA • REVENUE

Schedule C Loss (Hobby) Pilot

255 Schedule C Loss Audit Results	Actual W/O Data Mining	Actual with Data Mining
Success Rate	76%	83%
Dollars Assessed	\$3,606	\$3,917

MINNESOTA • REVENUE

C Corp and S Corp Income and Expense Audit Pilot

- Data mining in our experience did not work out as we had planned.
- Income and Expense audits for C Corps have only been done for about a year. S Corps have been doing them longer, but they often have other issues that may cause problems with the coding of the audit. Because there are often more than one issue, the amount of the assessment would not be a true indicator of the amount assessed for the income and expense audit. Without this historical data it was difficult to get useful results.
- Another problem we ran into was the number of audits done are so small in comparison to the number of returns, it is difficult to find patterns.
- Another issue we had for C Corps, not all the fields are captured, so it limits what information we can mine.
- Currently we are working to keep better records in one central location so in the future we would have some historical data to use.

MINNESOTA • REVENUE

Pilot Data Mining projects using University of Minnesota students

- **We are still testing the models we developed to see how they perform compared to our old methods of audit selection.**
 - Not actively working with the U of M at this time.
- **We still need to complete a number of audits based on the data mining findings.**
- **Some models turned out to be better at telling us who not to audit rather point out good audit candidates.**
- **There appears to be promise in the use of data mining for tax compliance, but there are still many unanswered questions on where and when this tool can be cost effective and sustainable.**

MINNESOTA • REVENUE

Data Mining Skills

Methodology Step	Difficulty Acquiring Skill	Skill Required
Business Understanding	Medium	DOR needs to be able to better discern what tax compliance problems can be addressed by data mining and which ones cannot.
Data Understanding	Low	Basic statistical analysis (i.e. correlations, data profiling, sampling, etc.) of data to identify problems or basic relationships that will affect the later steps.
Data Preparation	Medium	Certain data mining techniques require data normalization techniques that use statistical procedures to modify data prior to the modeling phase.
Modeling	High	For a single data mining software package the tool provides 10 or more different data mining modeling algorithms, each algorithm requires the tuning of 10 or more parameters. Data mining techniques are evolving constantly and keeping up with changes will require additional time investment. Some modeling packages require programming skills and do not have graphical user interfaces to simplify the modeling process.
Evaluation	High	The ability to evaluate the statistical results produced by any of the modeling tool. Even more important and critical is having the knowledge and experience to know what to do next when any given model does not generate useful results.
Deployment	Low	If a given model is going to be run on a regular basis (every second to once a week), then the models need to be tied into an operational process.

MINNESOTA • REVENUE

Pilot Data Mining projects using University of Minnesota students

- **The Good**
 - Professor Jaideep Srivastava
 - Very bright, inquisitive people
 - Exposure to cutting edge tools and techniques
 - The Cost – but its wasn't cheap
- **The Not So Good**
 - Some students better suited than others
 - They are students, not professionals
 - Difficult to retain students for more than 6 months
 - It takes time to perform the audits suggested by data mining – students gone by the time results are in

MINNESOTA • REVENUE

Started to measure value of Data Warehouse sources

- **Huge reluctance to give credit to the warehouse**
 - You don't need a warehouse to know you should audit your top 50 businesses
 - Agencies have been bringing in audits dollars long before anyone heard of data warehousing
- **It is usually very difficult to measure the value of individual data sources**
 - It is by combining multiple data sources that you derive the most value
- **The days of auditing the guy who cut you off on the highway are long gone**
 - 91% of Sales Tax Audit dollars collected involved warehouse data (2006)
 - 94% of Individual Income Tax Audit and Non-filer dollars collected involved warehouse data (2006)
- **We have begun to measure the efficiencies gained for those performing the audit selection function.**
 - In many cases we have cut the time to do their jobs by 1/3 to 1/2

MINNESOTA • REVENUE

Lessons learned

- **There are many "best" ways of using your data warehouse**
 - Which one is best for you?
- **Need buy in and participation from agency leadership**
- **Need individuals with depth in tax knowledge to partner with technology staff to be successful**
 - Some of our best auditors are retiring we need to tap into this knowledge now before it leaves
- **The skills needed to pursue these goals are many and varied, acquiring and maintaining these skills while managing costs was and continues to be a challenge**

MINNESOTA • REVENUE

Lessons Learned

- **You need to be in it for the long haul**
 - May need to change your data collection processes
 - Time to perform the audits to validate/refine your queries and models
 - Train staff and build relationships

- **It is not just new data sources**
 - Determine the cost of data sources and check to see if they are providing adequate value
 - Look at new ways to use data
 - Enhance collaboration
 - **Not just Business and IT, but cross divisions and units**

MINNESOTA • REVENUE

Lessons Learned

- **Plan for growth**
 - The more data you have, the more time you will spend maintaining it
 - The more folks using the data, the greater the need to mentor and support their activities
 - All of your staff's time can quickly be absorbed just keeping the existing warehouse operational

- **Manage expectations**
 - Demands for additional data sources
 - Demands for enhanced security and tracking
 - Demands for more complete data
 - Demands for "cleaner" data
 - Demands for more frequent data
 - Demands for even greater matching
 - Demands for easier tools

MINNESOTA • REVENUE