Data Roadmap

Scott Thorne
Mary Weisse
Agenda

• Current state and background
• Issues and problems
• Approach
• Next steps
  – issues to be addressed
  – issues deferred
MIT Data Warehouse - Overview

• Most administrative information is available
• Widely used and relied on
• Recognized by peer institutions as one of the best in higher education
• Improvements can make it more useful

Examples:
  – adding more data
  – cleaning up data
  – real time data
  – true 24x7 availability
Why the Warehouse was Created

• Primary reporting platform
  – provide a common access mechanism for all Institute wide
    administrative data
  – easier access provided by a simpler view of data

• Support departmental reporting needs
  – ability to combine central and local data
  – reduce rekeying of data

• Improve data quality through usage

• Self service
  – minimize support & operational costs
  – DLC's don’t need to rely on a central reporting group
What Data is Available

- Academic Chairs
- Alumni Biographic, Occupation & Degree
- Association of Student Activities Group, Requests and Space
- Benefits
- Budgets (Nimbus)
- CIP (Classification of Instructional Programs)
- Concur Travel
- Coop Textbook
- Course Catalog
- Cybersource MIT Storefront transactions
- Department Master Hierarchy (and other hierarchies)
- EHS Contact, Room set, Sarah, Training Data
- Facilities/Plant Maintenance Orders
- Fidelity – Benefits
- Financial
- Financial Aid
- Fund data (Resource Development)
- Graduate Admissions
- Graduate Awards
- HR (Employees, Positions, & Appointments)
- Cyborg & SAP
- IAP Credit & Non-Credit Activities
- Instructor
- Institutional Research - Honors & Awards
- Kerberos
- Keys
- Libraries (Financials, Orders, Circulations, Collection Management, Serials Control Processing)
- Metadata (Table, view, & field definitions, data models; report templates)
- MIT Card – Pictures
- MIT ID
- Moira List data
What Data is Available (cont)

- OSP Proposals and Award Cost Sharing
- Overhead Rates
- Payroll (Deductions, Distributions & Certifications)
- Pension
- Person
- Pre-registration & Registration
- Procurement (Purchasing, Requisition & Credit Card)
- Request Tracker
- Roles /Authorization
- Space
- Student Biographic
- Student Degree
- Subject Enrollment
- Student Grades
- Training & Events
- TLO

- Telephone Detail (ice 9)
- Student Term Enrollment
- Undergrad Admissions
- Vacation/Absences/Leaves
- Web Statistical Data
Issues and Problem Collection

• Who
  – Data Warehouse Community Survey
  – Task Force Reports
  – Reporting Tool Review Committee
  – DW Staff
  – Interview data users
    • Institutional, School, Department
    • Central, Research, Academic

• What
  – Document
    • specific issues
    • brief analysis
  – Identify next steps
Categories of Data Issues

• Warehouse infrastructure & operations
• Reporting environment
• Missing data
• Data access policy
• Strategic reporting (hard to answer questions)
Warehouse Infrastructure & Operations

Issues identified by the Warehouse Staff

• Job scheduling
  – start time manually set
  – extra time used to pad
  – hard to maintain

• 7x24 availability

• Congested nightly loads
  – scheduling
  – full/incremental
  – a few intensive jobs
Warehouse Infrastructure & Operations

- Performance of large detailed tables as history grows
- Managing the daily operation
  - reduce the number of non standard processes
- Incomplete Documentation
- Lack of Real time data feeds
- Lack a process to build and deliver higher level information
Reporting Environment Issues

• Data definitions need improvement
• Need intuitive table and field names
• Report descriptions aren't sufficient
  – hard to find the one you need
  – report names should be in the footer, so once you have paper you can find the report used to produce it
  – organize reports better
  – easy to identify reports that need to be run by DLC monthly
• Easier way to get the data from the Warehouse and format in Excel
• Easy way to join data from multiple domains
Data Missing from Warehouse

• OSP
  – Pre-Proposal
  – Award Budgets
• Property
• TLO – Basic
• Alumni
• ISO information
• Travel charges that haven't been submitted on a trip report
• Travel Guest Card charges not available for DLC's
• Historical Request Tracker Information
• Profit Center Group History
• Building Data – Date Built and Date Renovated
• Space Survey Data
• Building Depreciation
Data Access Policy Issues

• Room usage type
• Appointment supplements
• HR information for people affiliated with labs & centers
  – paid but not appointed
  – faculty not charged or appointed
Strategic Reporting  (Hard to answer questions)

• What does it cost MIT to deliver its programs?
  – Undergraduate Education, Graduate Education and Research
  – How do costs vary by school and program?
  – How do education costs relate to tuition charged (sticker price) and tuition collected (net tuition revenue)?
  – How do research costs relate to research revenues collected (direct and indirect)?

• What does it cost MIT to own and maintain its space?
  – capital (interest, depreciation)
  – operational (utilities, repair and maintenance costs)
  – How do costs vary by building and relate to market rates?
Strategic Reporting (Hard to answer questions)

• How much money has MIT received from all sources (gifts, research revenues, other) in support of major areas such as Cancer Research or Energy compared to total resources used and needed?

• What are MIT’s gross expenses by function before internal charges?
State of Data Practices

• Current software project development does not always include
  – reporting requirements
  – conceptual data model
  – data definitions

• Projects often focus on a transactional orientation
  – not what ways the information might be used in the future
Reporting Committee Findings

- IBM/COGNOS Tool Suite meets user requirements for functionality and look & feel.
  - Handling of local data
  - Web-based solution
- Data issues will not be resolved by a tool
- Rollout of the tool will be a multiyear effort
- Data issues should be addressed as the tool is rolled out
- Requires the involvement of the Business Owners and Community
Approach

• Demonstrate to the community that we are working on all aspects of the problem
• Create a repeatable process that can applied per subject
• Gain experience with the new tool
• Begin engaging Business Owners
• Partner with Development Projects
• Make demonstrable incremental progress
Areas of Focus

• Improving Reporting Infrastructure
• Improving Development Practices
• Improving Data Practices
Categories of Work

• Day to day operations
  – upgrading the environment
  – monitoring the loads
  – supporting minor changes

• Housekeeping
  – Metadata maintenance
  – Warehouse documentation

• New reporting tool administration

• Building reports

• Strategic questions

• Modeling

• New information in Warehouse
Resources

• Team
  – 2 Data Administrators
  – 2 Developers
  – 2 Reporting Consultants

• Project Teams

• Business Owners
Next Steps 2011

• Improving Reporting Infrastructure
  – migration to VM
  – develop plan for 24x7
  – reporting tool proof of concept

• Improving Development Practices
  – initial data modeling support

• Improving Data Practices
  – document problems & issues
  – Improve metadata documentation
Issues to be Addressed

• Tested in reporting tool proof of concept
  – better Excel integration
  – easier report delivery
    • automatic notification and delivery
  – better report descriptions
  – better report organization
  – simpler views of information
  – simpler joining of information

• Short term solution for Singapore access
Work Deferred (Post FY2011)

• Improving Reporting Infrastructure
  – real time data
  – 24x7 implementation
  – Job scheduling
• Improving Development Practices
  – to be determined
• Improving Data Practices
  – strategic (Hard to answer questions)
  – define and publish guidelines
  – define and publish roles & responsibilities
  – document data access policies
• New data in Warehouse
  – should be done with new tool
Improving Understanding of Data

• Simpler more focused views of information

• Better definitions and documentation

• Reporting tool proof of concept
  – including process for developing specialized views
Improving IS&T Practices

• Focus with the business on up front design
  – conceptual data model
  – business process models
  – reporting requirements

• Metadata documentation

• Document reconciliation processes
Improving Data Practices

• Publish principles & guidelines for data management
• Establish processes for data issue resolution and prioritization
• Gain consensus and document roles & responsibilities concerning data
• Develop a method to monitor progress
TIMELINE

July 2011

Improving Warehouse Environment
- Migrate to Linux/VM

Improving the Understanding of Data
- Reporting Tool Pilot
- Improving Meta Data
- Space Example

Improving IS&T Practices
- Establishing Modeling Conventions

Improving Data Practices
- Defining Roles & Responsibilities

Job Scheduling
- Partitioning

Rollout New Tool per Area
- Data Documentation per Area
- Define Focused Views per Area

Migrate to Linux/VM

7x24

Job Scheduling

Partitioning
Appendix

• Warehouse background
  – design principles
  – Warehouse processing
  – statistics

• Example of Addressing Issues in a single domain
  – Issues
  – short term plan
  – long term plan

• Issues found so far grouped by subject
Data Warehouse Guiding Principles

• Information in a single place
  – access any information in the same manner
  – integrated information
    • join information from several sources
• Structures should support the needs of all types of users
  – Central, DLC, Institutional
• Open access via SQL
  – not proprietary only access
  – uncommon in education and industry
  – can be accessed programmatically
• Warehouse is the data distribution hub
  – easy to verify when feeds and reports are from a single source
Data Warehouse Guiding Principles (CONT)

• Data should be stable
  – consistency between reports generated on the same day
  – ability to generate the same report at any point in time

• Data has to be accurate
  – stale data is better than inaccurate data

• Structures should make it easy to report

• Data is not corrected or modified in the Warehouse

• Access rules are shared by the Warehouse and transactional systems
  – access managed by Business Owners
Data Warehouse Guiding Principles (CONT)

- Data always comes from a system of record
- Users should be able generate and share reports
- Metadata is available for all information
- Access control done at the database level
  - enables direct access via any means
- Metadata driven transformations and loads
  - to know how data was arrived at
  - same code called thousands of times every night
Data Warehouse Process Stages
Metadata Driven

• Extract
  – minimal logic
  – full or Incremental files generated

• Transfer
  – usually pushed automatically
  – encrypted/decrypted
  – simple integrity checking done

• Convert
  – data into reporting format and structures
    • star schemas

• Load
Metadata

Warehouse Meta Data

Input Files

Mapping & Conversion Information

Conversion Control

Conversion_Code
Conversion_Desc
Provider_Code
Pre_Batch_Script
Post_Batch_Script
Master_control_indicator
Password_file_name
Last_Run_Date
Last_Run_Status

Conversion

Conversion_Code
Mapping_Id
Input_Id
Output_Id
Load_Sequence_Number
Acceptable_Error_Percentage
Loader_Control_File
Remove_Data_Script
Remove_Data_Args
Create_Index_Script
Drop_Indexes
Revert_Indexes
Pre_Conversion_Script
Pre_Conversion_Args
Post_Conversion_Script
Post_Conversion_Args
Post_Load_Script
Post_Load_Args

Output Files

Output Field Definition

File_Id
File_Name
Extension
Disk_Name

Field Mapping

Conversion_Code
Mapping_Id
Input_Field
Output_Field

Field

Conversion_Code
Mapping_Id
Output_Field_Sequence
Conversion_Function
Conversion_Parameter_1
Conversion_Parameter_2
Conversion_Parameter_3
Conversion_Parameter_4

Underlying Table Information

View Rules

Table_Id
View_Id

Base Table

Table_Name
Field_Id
Field_Name

Subject Area (Star)

Provider Code
Provider_Desc
Business Contact MIT ID
Business Contact Email
Technical Contact MIT ID
Technical Contact Email
Source System
Warehouse Directory Path

Subject Area Views

Subject Area Id
View_Id

User View Field

View_Id
Field_Id
Name
Definition
Length
Data Type

Field Dictionary Definition

Field_Id
Field_Name
Definition
Length
Data Type

User Visible Meta Data

Note: Users see a view of field info combining the definitions with the other field info. Table/view descriptions are maintained separately from base table descriptions.

Tables derived from other tables -- how do we see the representation (mapping? Dates of last structure modifications? -- Variations?)
Integrity Checking in all Processing

• Correct files on hand before job runs
• Record & byte counts
• Comparisons of control file to data file
• Conversion type checking, number and types of fields
• DBMS constraints
• Error checking and validation routines
• Mail sent to Warehouse & Business Owners
System Availability

• Approaches 24 x 7
  – except Saturdays from 3:00 pm – 3:10 pm
    • snapshot
  – some tables are temporarily unavailable in the early morning as new day’s data is loaded
## Statistics

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobs per night</td>
<td>~300</td>
</tr>
<tr>
<td>Input files per night</td>
<td>875</td>
</tr>
<tr>
<td>Field transformations</td>
<td>12,847</td>
</tr>
<tr>
<td>Distinct functions</td>
<td>78</td>
</tr>
<tr>
<td>Tables</td>
<td>1,522</td>
</tr>
<tr>
<td>Views</td>
<td>1,150</td>
</tr>
<tr>
<td>Records</td>
<td>1,904,948,496</td>
</tr>
<tr>
<td>Total database size</td>
<td>575g</td>
</tr>
</tbody>
</table>

1/26/11
Statistics

Fiscal Year 2011 Quarter 2  (October 1, 2010 – December 31, 2010)

• 664 distinct users logged into the Data Warehouse

• Representing 117 org units

• Generating 782,177 sessions
Statistics

DW User Community: Sessions by Department (400+ Sessions) FY 2011 Q2

- Environment, Health & Safety Office
- HQ, Information Services & Technology
- Others (research affiliates and scientists, postdoctoral fellows, visiting faculty/students/scholars, non-current employees, etc.)
- Center for International Studies
- Human Resources
- Research Laboratory of Electronics
- Office of Sponsored Programs
- Sloan School of Management
- Information Services & Technology
- Microsystems Technology Laboratories
- GL Ops & Reporting & Journal Vouchers
- Budget, Finance and Treasury
- Sponsored & Genl A/R & Cashier
Data Issues Example
(Space)

• Buildings need to be mapped to their cost collectors manually
• Building costs are collected via secondary postings
  – many of these are done manually requiring significant effort
• Utility usage metadata not available in the Warehouse
• Flaws in the accuracy due to the fact that some utility meters cover more than one building
• The older organization unit identifiers (6 digit) make it hard to relate space information to other domains
• All systems do not use a standard format when referencing location information making integration difficult
• Building built and last renovated dates are not shown in the Warehouse
• Access to space information in not consistent or documented
• Building depreciation information not in the Warehouse
Space Data Short Term Term Plan

• Use new HR org unit numbers in the feeds and views of the Warehouse
• Add the cost collector that relates to the building in the building feed
  – include the cost collector key in the Warehouse view
• Add new fields to the building feeds and views for date built and date renovated
• Add geocode fields to the room feed and views
• Update the Warehouse metadata
  – add information about utility usage
Space Long Term Plan

• Work with Facilities and others to agree on guidelines for integrating location information
  – building and room identifiers
  – geocoding
    • guidelines for tracking geocode for things like equipment
• Work to gain consensus and document access policies concerning space information
• Ensure that space survey data it in the Warehouse
• Encourage reporting of space information from the Warehouse where appropriate
  – possible pilot web reporting
• Bring building depreciation information into the Warehouse
  – create an integrate view of all space costs in the Warehouse
HR Data Issues

- Faculty promotion and Tenure dates have to be cleaned up by hand
  - SAP & business policy does not support correction of some information
- CIP (Academic Discipline) attached to All Academic Staff
- Race & Ethnicity not consistent over time
  - self identified ethnicity needs to be managed and synchronized with Institutionally identified
- Historical data is inaccurate because old records are changes without the modified date and effective dates properly maintained
- Many places where more than one affiliation occurs but always need one marked "Primary" to support non-duplicated counts
- HR data not accurate.
  - How many researchers have PHD?
- Many things could be calculated in a standard way to make it easier and reduce discrepancies.
  - For example: years at MIT, could be provided in addition to the appointment dates.
- There seem to be different tallies for the Annual Head Count Report (October 31st) which cause confusion.
- Inaccuracies in Faculty appointment dates
- Inaccuracies in Faculty Leave information
- Ethnicity & Race not maintained accurately
- How many FTE by FY for a DLC for the last 10 years
Finance Data Issues

- Hard to maintain CEMIT Groups
  - can't tell if they're used
    - so hesitant to delete them
  - Need to verify that they are all correct each month
    - maybe an exclusion list would help (to see if the right gl_accounts are excluded and only those)
    - mistakes lead to reporting errors
- Difficult to separate out internal charging and transfers
  - some specific GL_Accounts for internal that can be excluded for this view
  - could use SI transactions instead of SA, but no way to ensure people do that correctly
  - no way to enforce integrity rule in SAP
- Many Document Types not used, but not filtered out of list
- Sponsor Billing done at Parent WBS level, so children look in deficit
  - no common definition of what "deficit" accounts are
- GL_Account_Report is confusing.
- Travel charges that haven't been submitted on a trip report can't be reported on
- Guest Card charges not available for DLCs
- Sponsor Billing done at Parent WBS level, so children look in deficit No common definition of what "deficit" accounts are
  - billing lag, means DLCs and Central might have different views and definitions
- Hard to see if changes in a Person's status from ON/OFF Campus have been made before the postings
- Need Profit Center Group History so that we can recreate old reports and understand the differences in numbers
Student System Issues

• What is the Cost of Education?
  – By Department and Program
    • per Student
  – How much Tuition is Charged?
    • What is waived?
  – How much of Tuition is paid?
    • by the Student
    • by a Sponsor

• Don't have accurate info on Grads.
  – Where did they get their undergrad degree?
  – When did a student enter in which degree program?
  – They have Grad entry term, but not cohort.
  – Are they fully supported (external sources hard)?