Reporting Data Kwality Issues

Mike Glasser

Office of Institutional Research
University of Maryland – Baltimore County
Agenda

• Introduction
• Elements of Data Quality
• Correcting Issues
• Tables and Procedures
• Reports
• Data Quality Firewall
• Questions
UMBC

University of Maryland - Baltimore County

- Located in suburban Baltimore County, between Baltimore, MD and Washington, DC
- One of the three public research campuses in the University of Maryland System
- 10K undergraduate and 2K graduate students
- 2,200 Employees; 715 full-time faculty
- PeopleSoft, SQL Server 2008 R2
- 9 Pan-Am chess championships
# UMBC

## Data Warehouse & Reporting

<table>
<thead>
<tr>
<th>Headcount</th>
<th>IT</th>
<th>IR</th>
<th>Back Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Data Warehouse</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Reporting</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>FTE</td>
<td>2.5</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

- Student Administration module (incl Fin Aid)
- Finance module
- Learn Analytics module (beta)
Data Quality

• **Accurate**
  – Data entry
  – Outdated

• **Consistent**
  – Records in one table do not match other

• **Complete**
  – Missing data

• **Within Policy**
  – Violates a policy or practice

• **Reportable**
  – OK in system, but cannot be reported
Accurate Data

• Data Entry
  – Invalid Emplid
  – Invalid Institution
  – Assignment % for Class not 100%
  – Wrong County code for Maryland

• Outdated
  – Wrong contact information
  – Wrong birthdate
  – Student no longer in program
Consistent Data

• Inconsistent within record or tables
  – US Citizenship, F1 visa
  – Withdrawal code, date before classes
  – Class attribute not on catalog
  – StudentPlan without StudentTerm
  – Enroll Total in Class does not match registrations
  – New Plan not in DW setup table
Complete Data

• Missing data
  – Plan is missing CIP code
  – Class is missing Instructor

• Missing record
  – Class is missing Instructor
  – Course missing graded component
  – New Plan not in DW setup table
• Violation of Policy or Practice
  – Double majors for Grad students
  – Duplicate plans
  – 2\textsuperscript{nd} major is Undecided
  – Grad Assistant not enrolled
Reportable Data

- **Cannot report**
  - Unknown Gender (IPEDS)
  - Invalid major

- **Should not report**
  - Non-degree instead of degree seeking
  - Masters instead of Doctoral
  - Class Section changed after freeze
Remedies

• Meetings or Email
• Transaction system
  – Edit records manually
  – Change business process
  – Change PS data entry (if lucky)
• Data warehouse
  – “Unknown” values (key = -1)
  – Create fake records
    • Bad majors
    • “Two or More” Ethnicity
  – New fields
    • GenderIPEDS
  – Tweak data quality check
Philosophy

- Fix it in the transaction system
- Prevent / Fix it at data entry
- Fix it as soon as possible
- Fix it before IR census
- Fix it in data warehouse
Carrot / Stick

- Show them the errors
- Explain the impact
- Exposure to more users
- Data Management Cmte
- Provost
- OK, we will fix it in the DW
Practice

• **Data Quality Team**
  – Part of Data Management Cmte
  – Campus commitment to data quality
  – Identify responsible parties

• **Data Quality Reporting**
  – Identify errors
  – Report errors
Data Quality Team

- IR Data Administrator (convener)
- Data Managers/Stewards
  - Registrar
  - Scheduling
  - Financial Aid
  - Undergraduate Admissions
  - Graduate School
  - Finance
  - HR
Data Quality Team

- Kickoff meeting to explain philosophy and processes
- Meetings with back office(s) as needed
- Back office report developers
- Write SQL to identify issues
- Write PS Query in transaction system to identify issues
- Develop DW reports with more detail for specific errors
Data Quality Reporting

- Identify issue
  - Via meetings, IR, users, back office
- Write SQL
- Nightly procedures
  - Check everything, summarize
- Daily report
  - User subscriptions or on demand
- Fix errors
  - A little prompting
Tables

• Error Messages
• Daily Errors
• Copy of Yesterday’s Errors
• History of Errors
• Exceptions
Table for Error Messages

• 1 message per error check
• Error Message Number and Text
• Explanation and/or solution
• Module (Admissions, Registrar)
• Table name
• Field name
• Key fields
Table for Error Messages

- **Create table** *DW.Data_Quality_Error_Messages* (  
  - **Error_Msg_Nbr** int IDENTITY(1, 1),  
  - **Error_Msg_Text** varchar(100),  
  - **Error_Msg** AS (((Error_Msg_Text + ' [') + CONVERT(varchar,Error_Msg_Nbr)) + ']'),  
  - **Error_Msg_Severity** varchar(10),  
  - **Error_Explanation** varchar(500),  
  - **DQ_Module** varchar(3),  
  - **Addl_Recipients** varchar(200),  
  - **Table_Name** varchar(128),  
  - **Field_Name** varchar(128),  
  - **Process_Name** varchar(128),  
  - **DW_Load_Dttm** datetime,  
  - **Key_Fields** varchar(500),  
  - **DQ_Sub_Module** varchar(5))
Table for Daily Errors

- Error Message Number
- Value of the field with error
- Value of fields to identify record
- Date/time
- New error indicator
- Current error indicator
Table for Daily Errors

- create table DW.Data_Quality_Daily ( 
  - Error_Msg_Nbr     int,
  - Fieldvalue       varchar(100),
  - Key1             varchar(100),
  - Key2             varchar(100),
  - Key3             varchar(100),
  - Key4             varchar(100),
  - Key5             varchar(60),
  - Key6             varchar(60),
  - Key7             varchar(60),
  - Key8             varchar(60),
  - Key9             varchar(60),
  - DW_Load_Dttm     datetime,
  - Key_Values       varchar(1000),
  - New_Error_Yn     varchar(1),
  - Exception_YN     varchar(1),
  - Current_Yn       varchar(1))


SQL

• SELECT statement to identify the errors
• Has to return the key fields and the field with error (if applicable)
• Convert to INSERT statement for table Data_Quality_Daily

• Written by me or back office report developer
SQL

**Academic Plan has invalid CIP code**

```sql
INSERT (DW_Load_Dttm, Fieldvalue, Error_Msg_Nbr , Current_YN, Key1, Key2)
SELECT getdate(), a.CIP_Code, 180, ‘Y’
    a.ACAD_PLAN, a.EFFDT
FROM iPSSA.Source.PS_ACAD_PLAN_TBL A
LEFT JOIN iPSSA.Source.PS_CIP_CODE_TBL B
    WHERE b.CIP_CODE IS NULL

* Removed effective dating logic for simplicity*
Table for Error History

- Summarize daily error messages
- One record per message per day
  - Errors
  - Unique values
  - Exceptions
  - New errors
  - Current errors
create table DW.Data_Quality_History (  
  Error_Msg_Nbr int,  
  ETL_Load_Dttm datetime,  
  ETL_Date varchar(10),  
  Error_Count int,  
  Unique_Values_Count int,  
  Exception_Count int,  
  DW_Load_Dttm datetime,  
  New_Error_Count int,  
  Current_Count int)
Table for Exceptions

- Some errors can be warnings
- Exceptions are on individual case
- Still identified during the checks
- Exceptions are deleted from daily errors after summary, so exceptions can be counted
Table for Exceptions

- create table DW.Data_Quality_Exceptions ( 
  - Error_Msg_Nbr int, 
  - Fieldvalue varchar(100), 
  - Key1 varchar(100), 
  - Key2 varchar(100), 
  - Key3 varchar(100),  
  - .....  
  - Key8 varchar(100), 
  - Key9 varchar(100) 
)

<table>
<thead>
<tr>
<th>Error_Msg_Nbr</th>
<th>Fieldvalue</th>
<th>Key1</th>
<th>Key2</th>
<th>Key3</th>
</tr>
</thead>
<tbody>
<tr>
<td>179</td>
<td>%</td>
<td>SPSE PBC</td>
<td>%</td>
<td>%</td>
</tr>
</tbody>
</table>

Error Message 179: Academic Plan is missing Degree code
Procedures

- **Nightly Data Quality**
  - Run Data Quality Checks
  - Cleanup error table
  - Identify errors as new
  - Summarize errors
  - Delete exceptions

- **Data Quality Checks SA**
  - SQL for individual DQ checks
Reports

• Microsoft Reporting Services
• Summary report for any date
  – Count of errors, current and new
  – Broken down by module
  – Links to details of error message
• Detail report for last night
  – Which records had the errors
• List of possible error messages
  – With explanation and relevant data
## Summary Report

### Error Messages from Oct 9 2012 5:46AM

<table>
<thead>
<tr>
<th>Data Quality Module</th>
<th>Error Message</th>
<th>Errors</th>
<th>Unique Values</th>
<th>Except</th>
<th>New Errors</th>
<th>Current</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>1 messages</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>AD</td>
<td>Student has invalid Plan in Admissions stack</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>CC</td>
<td>1 messages</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>CC</td>
<td>Invalid EMPLID</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>IR</td>
<td>2 messages</td>
<td>28</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>IR</td>
<td>Academic Plan not found in OIR,DW,UW_Acad_Plan_Tbl</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>IR</td>
<td>Class Section changed after Day 10</td>
<td>26</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SR</td>
<td>18 messages</td>
<td>1290</td>
<td>154</td>
<td>13</td>
<td>0</td>
<td>506</td>
</tr>
<tr>
<td>CORE</td>
<td>Academic Plan has an invalid CIP code</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>DEG</td>
<td>Degrees in different Plan, same HEGIS</td>
<td>24</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DEG</td>
<td>More than one degree with same Academic Plan</td>
<td>17</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Report of Daily Errors

Data Quality Details for Last Night

Academic Plan has an invalid CIP code [180]

DQ Module: Student Records (CORE)
Table Name: PS_ACAD_PLAN_TBL
Field with Error: CIP_Code
Keys in Table: Acad_Plan ~ Effdt

Total Errors: 1

<table>
<thead>
<tr>
<th>Error Value</th>
<th>Keys</th>
<th>New</th>
<th>Current</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCT UDC~Jan 2 1901 12:00AM</td>
<td>N</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>

- Error Message with Number
- Field with error
- Keys in table
- Value of field with error
- Value of keys to identify record
- Error explanation on hover
Subscriptions

• Microsoft Reporting Services
• Setup email recipients
  – Can be anyone
  – Don’t need access to report
• Any schedule
  – Recommend daily new, weekly all
• Set report parameters
• Choice of formats
• Link to report online
## Error Messages from Oct 9 2012 5:46AM

<table>
<thead>
<tr>
<th>Data Quality Module</th>
<th>Error Message</th>
<th>Errors</th>
<th>Unique Values</th>
<th>Except</th>
<th>New Errors</th>
<th>Current</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>AD: Student has invalid Plan in Admissions stack</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>CC</td>
<td>CC: Invalid EMPLID</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>IR</td>
<td>IR: Academic Plan not found in OIR.DW.UW_Acad_Plan_Tbl</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>IR</td>
<td>IR: Class Section changed after Day 10</td>
<td>26</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SR</td>
<td>SR: 18 messages</td>
<td>1290</td>
<td>154</td>
<td>13</td>
<td>0</td>
<td>506</td>
</tr>
<tr>
<td>CORE</td>
<td>CORE: Academic Plan has an invalid CIP code</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>DEG</td>
<td>DEG: Degrees in different Plan, same HEGIS</td>
<td>24</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DEG</td>
<td>DEG: More than one degree with same Academic Plan</td>
<td>17</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
HR Data Quality

- Process built prior to BbA
- Similar to SA process, but not same tables or procedures
- Email summary sent daily with NEW errors, weekly with ALL
- Email created with SQL, not RS
- Details reported with Crystal
- 180 error checks
Comparing yesterday's errors with today, the following NEW errors were found ...

<table>
<thead>
<tr>
<th>Count</th>
<th>Error Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Eligible retirement code not found</td>
</tr>
<tr>
<td>1</td>
<td>EMPL_CLASS inconsistent with EEO6CODE</td>
</tr>
<tr>
<td>1</td>
<td>Unable to find Benefit_Plan for employee</td>
</tr>
<tr>
<td>6</td>
<td>EMPLID not found in UM_Person_Info table</td>
</tr>
</tbody>
</table>

=====================  
9  NEW errors  
1085  Total errors

Email produced by Email_New_ETL_Errors_Sp on Oct 9 2012 12:55AM
Data Quality Firewall

- **Procedures**
  - missing data
  - foreign key discrepancies
  - duplicate source keys

- **Warnings**
  - Usually missing data
  - Loaded as “Unknown” (Key = -1)

- **Critical Errors**
  - Usually duplicate source keys
  - Only the first is loaded
Why don’t we use it?

– I already had a system in place for HR data
– I did not know much, if anything, about it
– I did not know XML
– Maybe we could, if time were invested
Recap

- Identify people responsible for data quality in each area
- Back office commitment
- Identify issues and resolutions
- Use DW to capture and report issues
- Report issues to appropriate people
- Impact of data quality
- Accept that some things are wrong only for reporting
Wrap Up

Any Questions?

Mike Glasser
University of Maryland - Baltimore County

mglasser@umbc.edu  (410) 455-3577

Source code is available upon request