




Joe Bugajski
Senior Analyst
Burton Group
jbugajski@burtongroup.com

Bob Grossman
Managing Director
Open Data Group
rlg@opendatagroup.com

Data Governance for Improved Information Quality

MIT IQ Industry Symposium
Cambridge Massachusetts USA
16-17 July 2008

All Contents © 2008 Burton Group and Open Data Group. All rights reserved.





Data Governance for Improved IQ

2

Thesis

- Good Information Quality (IQ) needed to comply with Sarbanes Oxley, EU Privacy, and Anti-Terrorism Acts
- Audits, business process controls and spot checks are necessary but insufficient to assure good IQ
- Data field reuse changes meaning of information
 - May result in sensitive data stored in unprotected data fields
 - Increases transaction failure risks that reduce profitability
- Data Authority Reference Model lowered liabilities and recovered USD \$2 billion annual sales
 - Measure and improve semantic reliability
 - Provides effective data governance for high IQ



Data Governance for Improved IQ

3

Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions





Data Governance for Improved IQ

4

Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions



Authors



5

Joseph M. Bugajski: Senior Analyst, Burton Group Inc.

- Research data and technology governance, architecture, information quality and standards. Formerly Chief Data Officer at Visa Inc. responsible worldwide for information architecture and quality. Previously, Visa chief enterprise architect and resident entrepreneur for information strategy, risk systems, and data warehousing. Before Visa, CEO of Triada, a business intelligence software firm.

Robert L. Grossman: Managing Partner, Open Data Group

- Deliver management consulting, outsourced analytic services, and analytic staffing for data mining; Director at the National Center for Data Mining and Prof. at University of Illinois at Chicago (UIC). Also, Chair of the Data Mining Group (DMG) and advisory board member of InfoBlox. Before Open Data Group, founder and CEO of Magnify a data mining software provider to financial services. Before Magnify, was co-founder and co-director of the National Scalable Cluster Project, and co-founder and co-director of the PASS project.

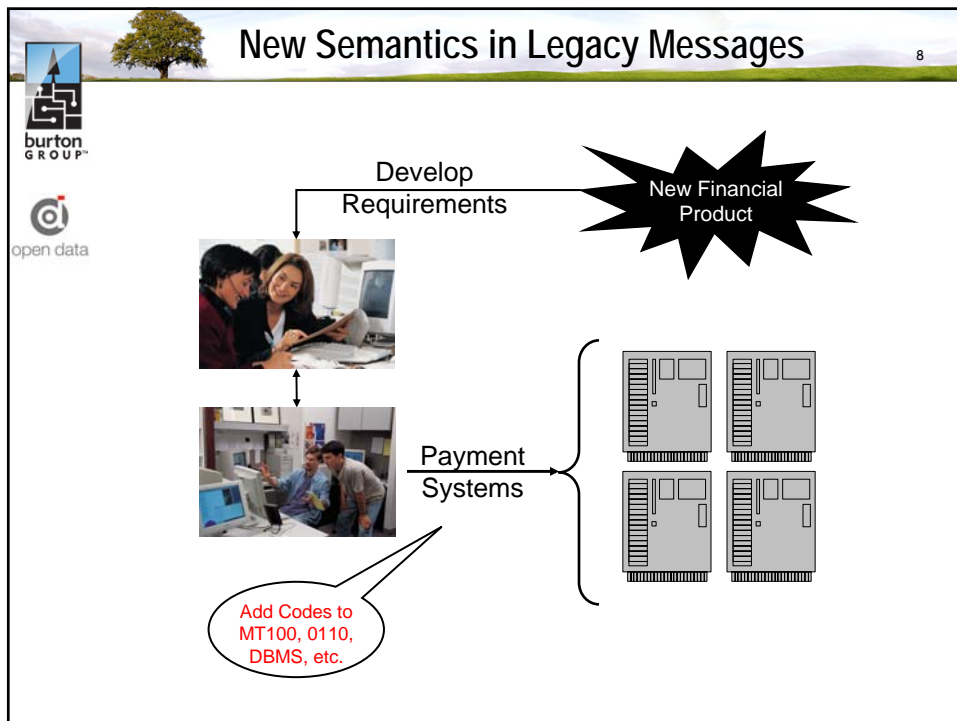
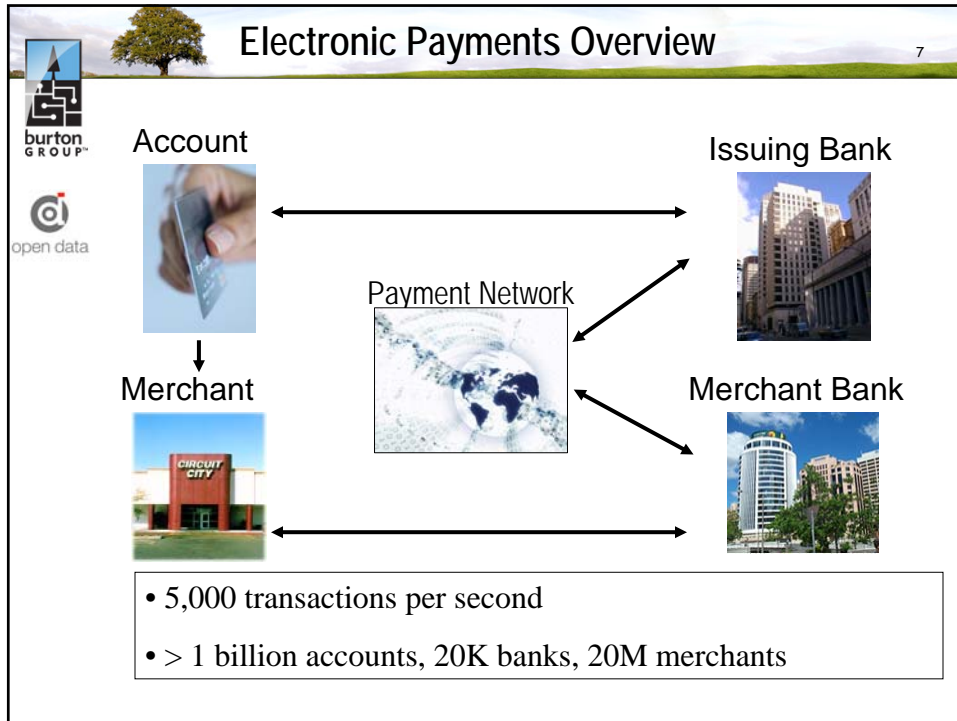





Data Governance for Improved IQ

6

Agenda

- Authors
- **Managing information in a global payment network**
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions





Information Quality Challenges

11

Problem Definition

- High costs and long lead times for adapting legacy systems for new products demands data field reuse
- Data field reuse creates confusing semantics and it may let sensitive data into unprotected data fields
- Multiple parties are involved in message networks, any of whom can modify data that heighten risks
- Faulty messages increase processing costs, result in lower sales revenue, and raise liability risks
- Data governance is required to lower risks

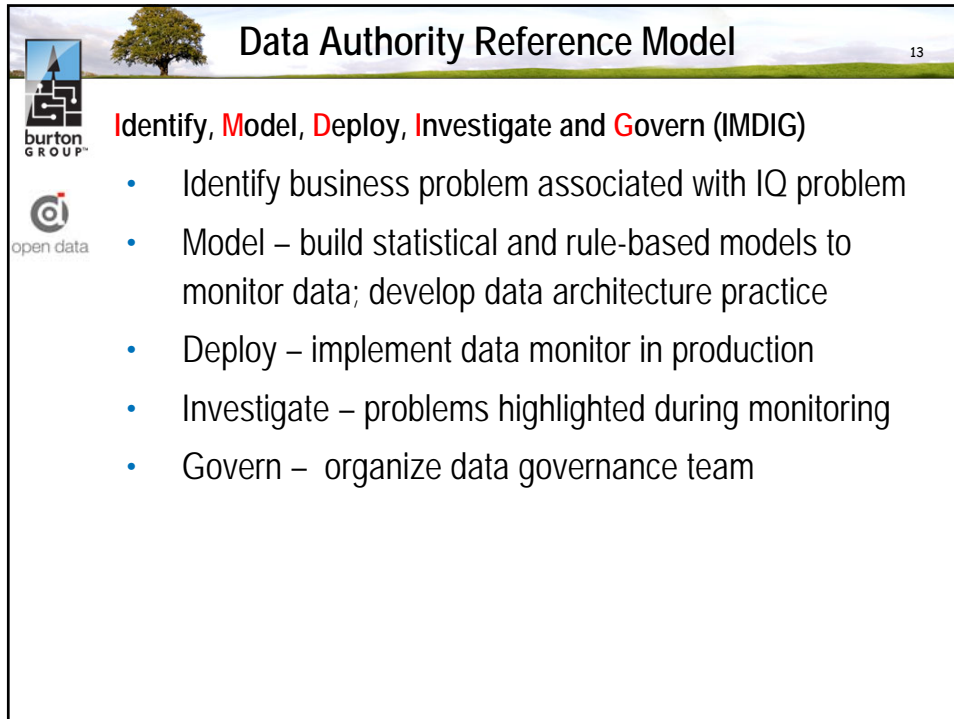


Data Governance for Improved IQ

12

Agenda

- Authors
- Managing information in a global payment network
- **Data Authority Reference Model**
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions

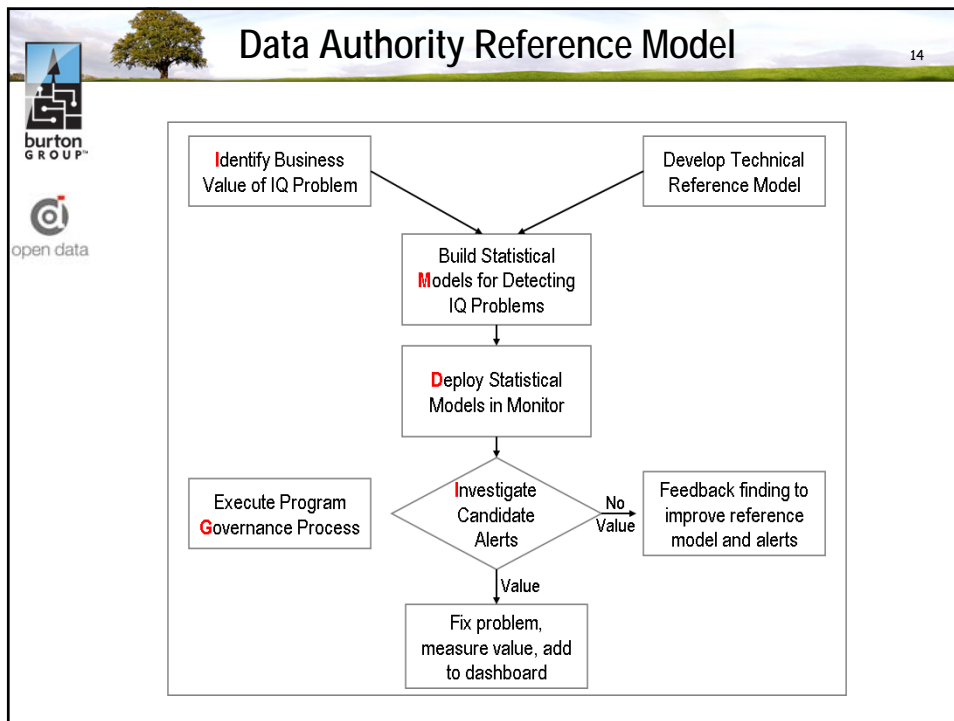


Data Authority Reference Model 13

Identify, Model, Deploy, Investigate and Govern (IMDIG)

- Identify business problem associated with IQ problem
- Model – build statistical and rule-based models to monitor data; develop data architecture practice
- Deploy – implement data monitor in production
- Investigate – problems highlighted during monitoring
- Govern – organize data governance team

Logos: burton GROUP, open data






Data Authority Reference Model 14

```

    graph TD
      A[Identify Business Value of IQ Problem] --> B[Build Statistical Models for Detecting IQ Problems]
      C[Develop Technical Reference Model] --> B
      B --> D[Deploy Statistical Models in Monitor]
      D --> E{Investigate Candidate Alerts}
      E -- No Value --> F[Feedback finding to improve reference model and alerts]
      E -- Value --> G[Fix problem, measure value, add to dashboard]
      H[Execute Program Governance Process] --> E
  
```

Logos: burton GROUP, open data




Data Authority Reference Model

15

Organizing the Data Authority Reference Model

1. Review data architecture across all I.T. systems
2. Measure production data quality (and interoperability)
3. Establish priorities for correcting data problems
4. Win commitment of CIOs to correct problems
5. Form a data governance team of experts, business and technical – team reports to CIOs
6. Build consensus solutions to top priority data issues
7. Write and maintain a Technical Reference Model
8. Report measurable progress to CIOs at least quarterly





Data Governance for Improved IQ

16

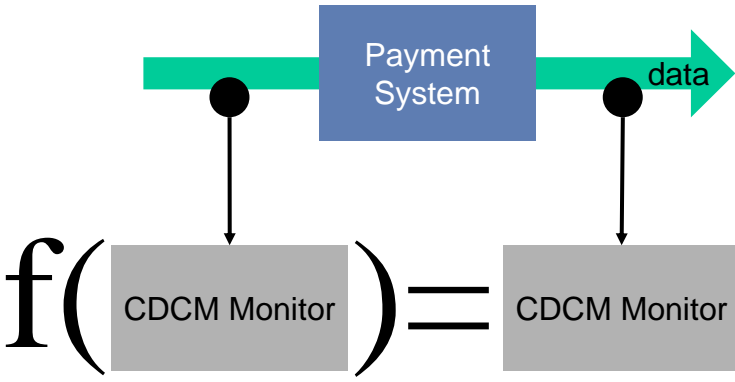
Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- **Measurements of Change Detection Using Cubes of Models (CDCM)**
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions

CDCM Data Measurements






- Apply CDCM at “tap points” across processing systems
- Tap at points to get “real” data – not cleansed
- Build baseline CDCM to compare with current sample

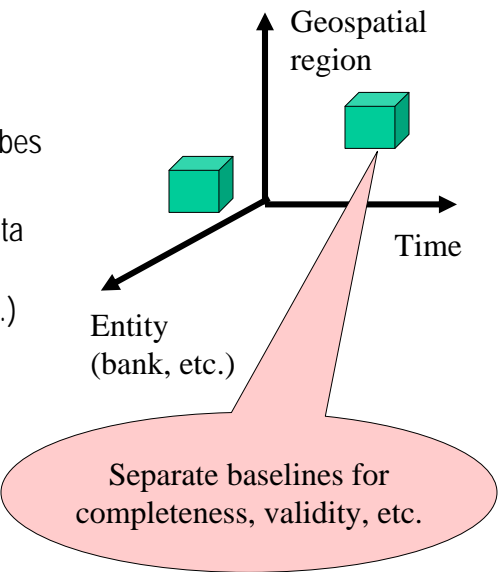


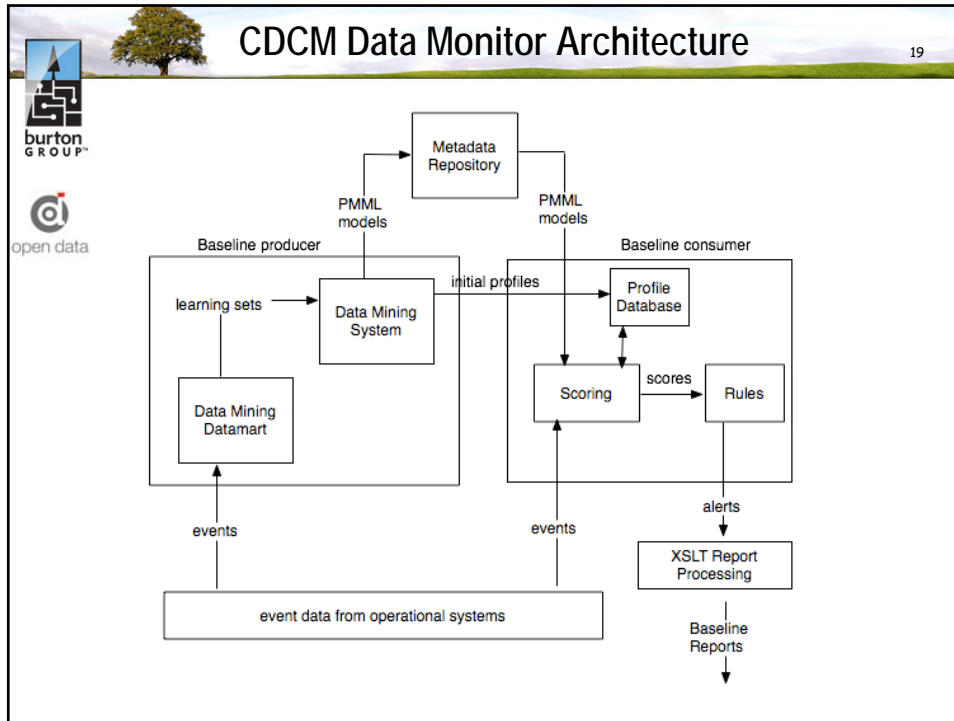
f

CDCM Logical Design

- Focus on Baselines & Changes
- Divide & conquer data (segment) using multidimensional data cubes
- For each cube, establish separate baselines by data quality dimension (e.g. completeness, validity, ...)
- Detect changes from baselines







Example: Bivariate Distribution Change

Baseline Model	
Value	Percentage
90, -	0.13
90, blank	0.21
05, -	0.01
05, blank	0.01
etc.	etc.
Total	100.00

Observed Model	
Value	Percentage
90, -	0.13
90, blank	0.63
05, -	0.01
00, blank	0.11
etc.	etc.
Total	100.00

Logos for 'burton GROUP' and 'open data' are present in the top left corner.



CDCM Launch Summary

21

Change Detection using Cubes of Models implementation

1. Select "best" data tap point(s) into production systems
2. Build baseline development and production monitor
3. Develop scoring process by quality measures: completeness, validity, etc.
4. Build baseline CDCM models and add to monitor
5. Score data from tap points – highlight serious problems
6. Generate trouble alerts then start improvement process
7. Track improvements and report results in CIO dashboard
8. Update Data Authority Reference Model with findings






Data Governance for Improved IQ

22

Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- **Building a Global Data Governance Program**
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions



Create Data Governance Program

23

Data governance program designed to deliver business value

- Use data architecture study and baseline data quality measurements to focus program on fixable problems
- Define measurement and correction program and build support for it among all data experts
- Ask finance for estimate of monetary value of fixes
 - Build a financial model; e.g., a P&L for program; and live with it!
 - Set financial objectives for top problems list
- Earn support for program at “C” level; start with CIOs
 - Set realistic monetary value recovery objectives
 - Make business case using simple and real examples of problems
 - Assure proper level of staff support to actually fix problems




Organize Data Governance Team

24

Team members are data experts with “C” level credentials

- Data experts are from business units and I.T. groups
- Choose members with knowledge and collaboration skills
- Have members approved by their “C” level executives
- Host a “kick-off meeting of data governance team”
 - Agree upon financial objectives and dashboard reporting
 - Agree upon data quality measurement rules
 - Approve initial version of Data Authority Reference Model
- Meet telephonically at least bi-weekly
- Meet in person at least quarterly






Lead the Data Governance Team

25

Team members are have real work to do!

- Write problem “alerts” as individual business cases
- Assign each alert to a governance team member
- “C” level granted authority to team members to deliver business value by fixing data problems
- Team members report progress to financial objectives
- Leader solves common problems and meets with CIOs
- Feedback from team members regarding alerts and quality improvements provides foundation which drive
 - Updates to Data Authority Reference Model
 - Improvements to CDCM and Alerting process



Data Governance Program Review

26

Support data quality improvement with appropriate governance

- Align corporate objectives, program oversight, and alert investigation and improvement processes
- Report to CIOs on dashboard
 - Top priority concerns as agreed by the governance team
 - Strategic objectives met through program operations
 - Total value recovered versus agreed financial objectives
- Adjust alert workflow by refining quality rules, number of CDCM statistical models, and alert value thresholds
- Support the governance team members at all times
- Maintain the Data Authority Reference Model






Data Governance for Improved IQ

27

Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- **Recapturing lost revenue through improved IQ**
- Additional work required
- Conclusions

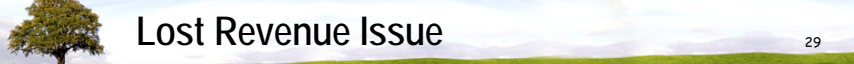




Customer Satisfaction Problem

28

Chip Card Terminal Coding Error

- Cards with “contact chips” common outside USA
- Terminals in USA read “contactless chips”
- Data field codes not well defined in technical manuals
- CDCM monitor notes contact chip activity in USA
- VIP customer shortly thereafter receives payment decline
- Business case for solution is clear
 - Update technical manuals with clearer definitions of chip codes
 - Send technical letter terminal suppliers and banks with correct codes
- Problem resolved within weeks






Lost Revenue Issue

29

Incorrect Country Code Error

- Payments messages record location of merchant
- CDCM finds mismatch between country and city
- Problem results in higher than normal rate of declines
- Inappropriately declined payments reduces revenue
- Governance team member worked with merchant's bank to define data problem and a business case for fixing it
- Problem resolved within a few weeks with corresponding increase in payment revenue





Overpayment Problem

30

Incorrectly Coded Sales Channel

- Payments messages record sales channel and terminal type; e.g., face-to-face and card present or e-commerce
- Channel and terminal data helps with risk assessment and also can be used to set payment processing fees
- CDCM finds mismatch between channel and terminal
- Problem leads to higher than expected payment fees
- Governance team member worked with bank to define data problem and the business case for fixing it
- Problem resolved within a few weeks with corresponding reduction in payments followed by additional business




Sensitive Data Values in Wrong Field

31

Data field reuse commonly applied for new products

- Payment messages designed to 1987 ISO standard
- These legacy messages used in thousands of systems
- Infrastructure costs for changing systems is HUGE!
- Reuse (overloading) of data fields is a best practice for recording data needed for new products
- Certain types of payment products required more information about cardholders than message held
- Decision to reuse data field for personally identifiable, non-public information reversed by governance team





Data Governance for Improved IQ

32

Agenda



- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- **Additional work required**
- Conclusions



Data Semantics Wrong from the Start 33

70+ case studies show problems start with data definitions

- Confusion about business definitions for data values
- Confusion about ways to correctly code data values
- Confusion about meaning of messages expressed by combinations of data values
- Reuse of legacy data fields for new semantic value adds to confusion and admits potential for sensitive data entries into fields where controls might be less stringent
- Complicated, unrecorded, metrics for data analysis result in inconsistent risk analysis and reporting





Financial Industry Tries to Fix Semantics 34

Work begun at ISO, UNCEFACT, OMG, other standards groups

- Solution to semantic variability problem is development of accurate, computer readable data models (e.g., UML)
- ISO 20022 (UNIFI*) standard developed to create XML messages from computer data models
- UNCEFACT Core Components applies technology similar to UNIFI to update EDI specification
- OMG Conversion Models for Payment Messages (CM4PM) applies UNIFI to generate legacy messages
- XBRL, FIX, IFX and others intend to follow suit

UNiversal Financial Industry (UNIFI) message scheme



Solutions Still Needed

35

High Value for solution to semantic variability problem

- Problem with defining elemental business terms
- Problem reusing elemental business terms in messages
- Modeling tools emit inconsistent XML form of models
- Software development tools today do not share models
- Metadata repositories do not support data governance
- Data models repositories incapable of global scale
- Full scale tests of technology stack remains incomplete
- Lingering belief that “yet another message format” will solve all these data semantics problems





Data Governance for Improved IQ

36

Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- **Conclusions**




Conclusions and Summary

37

Data Authority Reference Model Program Delivers Value

- Program very effective for recovering business value and improving data quality
 - Authors recovered USD \$2billion for payment company over three years of program operation
- Identify the problem, Model the data, Deploy monitors, Investigate problems, and Govern program (IMDIG)
 - Build consensus about data problems and business value
 - Develop Change Detection using Cubes of Models (CDCM)
 - Monitor production data
 - Create alerts and track solutions
 - Governance team of "C" level delegates own dashboard reports about top problems, strategic accomplishments and value recovered



Recommendations

38

Build governance that returns value to your business

- Develop a Data Authority Reference Model program in your business or agency
- Report your results to this forum, ICIQ, DAMA
- Help design standards that address global problem with semantic variability
- Develop tools and technologies to solve open issues
- Continue research into root causes of data problems





Data Governance for Improved IQ

39

References

- Joseph Bugajski, Robert L. Grossman, An Alert Management Approach to Data Quality: Lessons Learned from the Visa Data Authority Program, 12th International Conference on Information Quality (ICIQ), 2007
- Joseph Bugajski and Philippe De Smedt, Assuring Data Interoperability Through the Use of Formal Models of Visa Payment Messages, 12th International Conference on Information Quality (ICIQ), 2007
- Joseph Bugajski, Robert L. Grossman, Chris Curry, David Locke, and Steve Vejck, Data Quality Models for High Volume Transaction Streams: A Case Study, 13th International Conference on Knowledge Discovery and Data Mining, 2007
- Joseph Bugajski, Robert L. Grossman, Eric Sumner and Steve Vejck, Monitoring Data Quality for Very High Volume Transaction Systems, Proceedings of the 11th International Conference on Information Quality, 2006
- Leo L. Pipino, Yang W. Lee and Richard Y. Wang, Data Quality Assessment, Communications of the ACM, Volume 45, 2002, pages 211-218
- The Predictive Model Markup Language (PMML), www.dmg.org
- The Augustus open source data mining system can be downloaded from www.sourceforge.net/projects/augustus
- ISO 20022 (UNiversal Financial Industry message scheme), www.iso20022.org
- Conversion Models for Payment Messages (CM4PM), approved and awaiting publication at http://www.omg.org/technology/documents/spec_summary.htm



Data Governance for Improved IQ

40

Thank You!

&

Questions?

Joe Bugajski <i>Senior Analyst</i> <i>Burton Group</i> jbugajski@burtongroup.com	Bob Grossman <i>Managing Director</i> <i>Open Data Group</i> rlg@opendatagroup.com
---	---