IBM InfoSphere Discovery: The Power of Smarter Data Discovery

Gerald Johnson
IBM Client Technical Professional
gwjohnson@us.ibm.com
Objectives

- To obtain a basic understanding of the IBM InfoSphere™ Discovery solution through discussion and demonstration
Agenda

- Introduction
- InfoSphere Discovery Overview and Foundational Capabilities
- Summary
- InfoSphere Discovery Demo: Business Objects and Relationship Discovery
You Can’t Manage What You Don’t Understand

- **Distributed Data Situation:**
  - Grows exponentially
  - Increasingly distributed
  - Poorly understood

- **Problems:**
  - Poor understanding = Poor IT agility
  - Poor understanding = Poor data governance
  - Poor understanding = Bad data = Bad business decisions
Poor Understanding = Unpredictable Project Deployment

Discovery is the first phase of information-centric projects

Data Growth
Consolidate
Privacy

Discovery Phase
Automate Discovery and Accelerate Information Understanding

- **Significant Acceleration of Information Agenda projects**
  - Data Growth Management
  - Test Data Management
  - Sensitive Data De-identification
  - Application/Data Consolidation, Migration and Retirement
  - Master Data Management and Data Warehousing

- **Why is this different?**
  - **Data-based discovery**
  - **Automate** discovery of business entities, cross-source business rules and transformation logic
  - **Evaluate** multiple data sources simultaneously
  - Identify and **remediate** cross-system rules and inconsistencies
IBM InfoSphere Discovery Components

**Cross-Profiler**
- Basic profiling plus automated primary-foreign key, business entity and cross-source overlaps discovery

**Transformation Analyzer:**
- Discover complex business rules and transformation logic between two data sources

**Unified Schema Builder:**
- Prototype empty targets from the combination of many data sources
Value to Data Management?

Accelerate understanding your existing distributed data landscape for:

- Data Archiving
- Test Data management
  - Sensitive Data
- Application and Data Consolidation, Migration and Retirement
- Master Data Management and Data Warehousing

Dramatic reduction in risk, time and effort for the discovery phase projects

- Automated discovery of business entities, cross-source business rules and anomalies
- Increased repeatability
- Verifiable results
How Information Management Products Work Together

- **Automated relationship discovery**
  - Accelerate and improve data quality monitoring and management

- **Consolidation prototyping capabilities**
  - Build a better quality solution the first time

- **Transformation rules discovery**
  - Save time finding implied logic in legacy applications

- **Automated business object identification and confidential data location**
  - Achieve regulatory compliance faster

**Products:***
- **Information Analyzer**
  - Assess, Monitor, Manage Data Quality
- **DataStage / QualityStage**
  - ETL / Data Quality, Master Data Management
- **FastTrack**
  - Capture Design Specification
- **Optim**
  - Data Privacy, Test Data Management, Data Growth & Application Retirement
InfoSphere Foundation Tools

Govern your information assets

- Manage Business Terms
  Business Glossary
- Design Enterprise Models
  Data Architect
- Monitor Data Flows
  Metadata Workbench

- Business Glossary Term Pack
- Leverage Industry Practices
  IBM Industry Models

Target Model

Shared Metadata

- Understand Data Relationships
  Discovery
- Capture Design Specifications
  FastTrack
- Assess, Monitor, Manage Data Quality
  Information Analyzer
Extending the Foundation Tools

InfoSphere Foundation Tools

Manage Business Terms
Business Glossary

Discover Data Relationships
New – Discovery

Design Enterprise Models
Data Architect

Capture Design Specifications
FastTrack

Assess, Monitor, Manage Data Quality
Information Analyzer

Monitor Data Flows
Metadata Workbench

Enterprise Projects

Test Data Generation
Application Retirement & Consolidation
Data Archival
Data De-identification
Data Quality
Data Integration
Master Data Management
Data Warehousing

Common Metadata

Discover
Extends and Completes Discover Phase

Govern

Discover
Design
Govern
IBM InfoSphere Discovery: The Power of Smarter Data Discovery © 2011 IBM Corporation

- Core Profiling
- Discovery Information Analyzer
- 360º data relationship discovery
- Automated discovery methodologies
- Discover business objects
- Identify confidential data
- Rapidly prototype data migration and consolidation
- Assess, monitor, manage data quality
- Across heterogeneous global systems
- Business friendly interface
- Re-usable rules library
- Collaborative exception management

Business Users
Technical Users
Automating the process of understanding how data is related across heterogeneous systems
Data quality assessment, analysis and continuous monitoring with exception management.
Complete Data Analysis & Quality Assessment Solution
IBM Solutions for your Information Agenda Projects

Analysis\Discovery

Operation\Run Time

MetaData Server

InfoSphere Discovery

Optim

InfoSphere Information Analyzer and Validator

InfoSphere Information Server

InfoSphere MDM Server and Data Warehouse
IBM InfoSphere Discovery Components

**Cross-Profiler**
- Basic profiling plus automated primary-foreign key, business entity and cross-source overlaps discovery

**Transformation Analyzer:**
- Discover complex business rules and transformation logic between two data sources

**Unified Schema Builder:**
- Prototype empty targets from the combination of many data sources
InfoSphere Discovery – Cross-Profiler

What is unique?
Only solution on the market that automatically discovers primary foreign keys, business entities, and performs cross-source analysis

Applicable to
- Data Archiving
- Test Data Management
- Sensitive Data Discovery

- Data profiling and cross-system overlap analysis of up to 20 systems simultaneously
- Automated PF Key and Business Object discovery
- Extremely easy to install and use

Distributed Enterprise Structured Data
Challenges in Building an Archiving Solution*

1. Identify the data source that needs growth management

2. "Subject Discovery": rough inventory of candidate data sources, tables and files.

3. Identify and Document Referential Integrity (RI).

4. Identify business objects (BOs) and document

5. In Optim, develop and configure archiving templates by reading RI and BO documents

*Same Challenges apply for Application Retirement

- Find information on "business subjects" such as "purchase orders", "portfolios". Figure out which area of the data need to be controlled and the central tables in these areas.

- Typical data source has large schema and implicit relationships that are not declared or documented. It is difficult to find these RIs without tooling support.

- Boundary of business objects is often not documented at all. It is very difficult to establish business objects without tooling support.
Building Archiving/Retirement Solutions Using Discovery and Optim

1. Identify the data source that needs growth management

2. Take rough inventory of candidate data sources, tables, and files.

3. Use PFkey Discovery to discover and establish RIs

4. Use Data Object Discovery to discover, review, and establish Business Objects.

5. Export discovered artifacts to Optim

6. In Optim, automatically configure archiving templates based on above discovered relationships and business objects.

7. Use InfoSphere Discovery to discover critical artifacts needed when archiving or retiring systems
Building a Test Data Solution Without Tooling Support

Compliance team develops sensitive data mandate

Identify existence of the sensitive data mandated. Document the location.

The types of sensitive data that must be masked in test data are known. Where these data reside in a large data set, is not known. Manually scanning for these sensitive data, is difficult without tooling support.

Identify the data source that will provide test data

Find information on “test subjects” such as “purchase orders”, “portfolios”. Identify subject data to be extracted to support testing

Identify the subject tables that support the test cases

Typical data source has large schema and relationships that are not declared or documented. It is difficult to find these RIs without tooling support.

Identify and Document Referential Integrity

Boundary of business objects is often not documented at all. It is very difficult to establish business objects without tooling support.

Identify business objects and document

Use SQL/custom masking functions to create test data extract based on above documentation

Audit and approve the extract with privatization.
Building TDM Solutions Using Discovery and Optim

Identify the data source that will provide test data

Identify the subject tables that support the test cases

- Primary-Foreign Key Discovery
- Business Object Discovery
- Sensitive Data Discovery

Export all discovered artifacts to Optim including sensitive data classification

In Optim, semi-automatically configure archiving templates based on above discovered RIs, business objects, and sensitive data elements.

Audit and approve the extract with privatization.

Use InfoSphere Discovery to discover critical artifacts needed for creating test data with privatization.
Schema with no declared/document keys
Schema with discovered key relationships
Data Object Integration with IBM Optim

The same data object imported into IBM Optim™ Designer and ready for:

- Archiving
- Test data
- Masking
Basic Discovery Summary

- Analyze up to 20 data sources simultaneously
- Discover column level statistics
- Automatically discover primary-foreign keys
- Discover business objects
- Generate referentially consistent sample sets
- Identify critical data elements and overlaps across data sources
IBM InfoSphere Discovery Components

**Cross-Profiler**
- Basic profiling plus automated primary-foreign key, business entity and cross-source overlaps discovery

**Transformation Analyzer:**
- Discover complex business rules and transformation logic between two data sources

**Unified Schema Builder:**
- Prototype empty targets from the combination of many data sources
InfoSphere Discovery Transformation Analyzer

- Automates discovery of:
  - Cross-system business rules and transformations
  - Data inconsistencies
- Detailed data mapping between two data sources
- Discrepancy discovery
- Cross source trouble-shooting workbench

**Applicability**
- Discover cross-source rules for data consolidation
- Add new sources to existing MDM hub or EDW
- Map an MDM hub to consuming applications

**What is unique?**
- Discovers cross-system business rules, transformations and data exceptions by examining data values

Distributed Enterprise Structured Data

If age<18 and Sex=M then 0
If age<18 and Sex=F then 1
If age>=18 and Sex=M then 2
If age>=18 and Sex=F then 3

= Example
IBM InfoSphere Discovery Components

**Cross-Profiler**
- Basic profiling plus automated primary-foreign key, business entity and cross-source overlaps discovery

**Transformation Analyzer:**
- Discover complex business rules and transformation logic between two data sources

**Unified Schema Builder:**
- Prototype empty targets from the combination of many data sources
Master Data Management Project

- Community
- CRM
- Region

Goal: build a new master from multiple sources:

- What data is in each source?
- How do you combine columns together?
- What are the matching keys used to align the rows?
- What is the trust precedence for each attribute?
Deploying a Master Data Management System Without Tooling Support*

- Identify data for MDM: What are the data to be moved, where are they?
- Common schema: Known models: how do I know they will fit? No models, how do I build one?
- Map sources: How to map? How do I make sure all maps are “compatible“?
- Cleansing, enrichment:
- Match: Match keys are hard to come by
- Merge: Not sure which source is more trustworthy or why

*IMPORTANT Similar use cases when:
- Migrating multiple sources to new operational system
- Deploying a data warehouse
Deploying a Master Data Management System Using InfoSphere Discovery

**Identify Data for MDM**

*Use Discovery to discover and establish Data Inventory: Column Statistics, Relationship Analysis, Data Overlaps*

*Use Profiling results to inventory Critical Data Elements*

**InfoSphere Discovery Foundational Steps**

- Unified modeling
  - Build new model using CDEs
  - Assess completeness of a unified model

- Unified profiling of maps and domains
  - Prototype maps using data views and stats
  - Unified profiling to align maps cross-source

**InfoSphere Discovery Unified Schema Builder**

- Match key discovery and analysis
  - Discover match keys
  - Analyze the strength of a match key

- Conflict detection and resolution analysis
  - Trusted source analysis
  - Conflict resolution rules analysis prototyping
InfoSphere Discovery - Unified Schema Builder

- **Unified Schema Builder:**
  - Workbench for prototyping Master Data Management (MDM)
    - Identify critical data elements
    - Draw asset overlap topology
    - Build unified model
    - Map sources, validate across
    - Prototype match & merge
  - Complete methodology for MDM data discovery
  - “Test it out before you roll it out”

- New Application
- Data Warehouse
- MDM Hub

What is unique?
Prototypes empty targets from existing source data (MDM, EDW, data migration) and populates prototype with consolidated source data
Why is it such an “easy button”? 

- It is not.

- Discovery provides a rich environment for user to guide the discovery process
  - Confirm discovered relationships
  - Assess/establish known/heard relationships
  - The more review you do, the more accurate it will get
  - …but, it is kind of an easy button because
    - Discovery guides you through a well-defined workflow to get to the end results.
Reminder: Value of InfoSphere Discovery

- Accelerate understanding and validation of your existing distributed data landscape for:
  - Data Archiving
  - Test Data management
  - Application/Data Consolidation, Migration and Retirement
  - Sensitive Data
  - Master Data Management and Data Warehousing

- Dramatic reduction in risk, time and effort for the discovery phase of your project
  - Automated discovery of business entities, cross-source business rules & anomalies
  - Increased repeatability
  - Verifiable results
Questions