Module L13
Data Integration Part 2

Presented by: John R. Talburt
Acxiom Chair of Information Quality, UALR
Chief Science Officer, Black Oak Analytics, Inc.
September 28, 2016
Overview

• Module Objective
  • Continue the overview of the role and challenges around data integration with special attention entity information life cycle management and Big Data

• Topics
  • What is entity identity information management (EIIM)
  • The CSRUD Model of the Entity Information Life Cycle
  • What is master data management (MDM)
  • What is the impact of Big Data on data integration?
  • Modernizing data integration
Entity Identity Information Management (EIIM)
We Want to Have Persistent Identifiers

Entity Identity Information Management (EIIM)
Provide persistent entity identifiers (tokens) that do not change over time

LinkID: 00008888

LinkID: 00008888

LinkID: 00008888
Entity Identity Information Management (EIIM)

• An extension of ER in two dimensions
  • Knowledge management
    • Creating, storing, and managing the information that represents the identity of an entity
    • Entity Identity Structure (EIS)
  • Temporal
    • Maintain persistent entity identifiers over time, i.e. process to process
• Essential for
  • Effective master data management (MDM)
  • Entity-based data integration
Identity Information has a Life Cycle
CSRUD Identity Information Life Cycle Model

• Capture of Entity Identity Information
• Store and Share Entity Identity Information
• Resolve and Retrieve Entity Identifiers
• Update Entity Identity Information
• Dispose (Retire) Entity Identity Information

Yinle Zhou, IBM
Capture Phase in an EIMS

Build the initial population of identities (EIS) to be managed.
Store & Share Phase

• The Identity Knowledgebase is the primary repository of identity information and provides a central point of management

• The knowledgebase comprises the set EIS that represent each identity under management

• EIS vary from system to system and use different formats, e.g. XML structures, relational database rows.
Update Phase (Automated)

- Entity References
- Link Index
- Clerical Review Indicators

Application System

EIM Service

Staging

ER: Identity Update

Current Identity KB

Rules

Updated Identity KB

Sept 26-29, 2016

(c) 2016 iCDO@UALR
Update Phase (Manual)

Clerical Review Indicators → Visualization Tool → Assertions

Application System
EIM Service

Current Identity KB → ER: Identity Update → Updated Identity KB

Rules

Sept 26-29, 2016
(c) 2016 iCDO@UALR
Rule Override: Asserted Linking

Known to be equivalent because Mary reported her change of name (marriage) and new address to the publisher of her favorite magazine.
Resolve and Retrieve (Batch)

Entity References → Link Index → Confidence Indicator

Application System → EIM Service → Staging

ER: Identity Resolution

Identity KB

Uses IKB Information, but does not change it

Rules

Sept 26-29, 2016 (c) 2016 iCDO@UALR
Resolve & Retrieve Phase (Interactive)

Application

Identity Information

Entity Identifier

API

Application System

EIM Service

ER: Identity Resolution

Current Identity KB

Does not alter the IKB
Dispose (Retire) Phase

- Eventually, some identities will no longer be relevant or active with respect to the application
- EIS can be moved from the IKB into an archive leaving only a placeholder in the IKB.
- Beware of schema change!
  - When the definition of EIS change, it can create a problem in the retrieval of archived information
Master Data Management (MDM)
Master Data Management (MDM)

• MDM is a collection of
  • Policies, Procedures, Services, and Infrastructure

• To support the
  • Capture, integration, and shared use

• Of
  • Accurate, timely, consistent, and complete

• Master data

David Loshin, *Master Data Management*
Enterprise Data Governance (DG)

- Business Glossary, RDM, DQ, ...
- Policies and Procedures (Master Data Governance)
- Master Data Management (MDM)
- Master Data System of Record
- Entity Identity Information Management (EIIM)
- Entity Resolution (ER)
MDM System of Record

• An authoritative source of information about each master data entity
• Sometime called the “golden record”
• Has a persistent identifier from the EIIM system
• In legacy systems, the EIIM system and System of Record were often the same
• In newer systems they are separate functions
Big Data and ER Performance
Increasing Performance

- Theoretical resolution is impractical for large datasets
- For performance reasons, most systems limit number of comparisons
  - Blocking
  - Sorted Match (Hash) Key (aka: Sliding Window)
  - Inverted Index
- Compromise: By limiting number of matches, always a chance some matches will be missed
Blocking Strategy
Match-Key Index

Example: Single-Key Index by Zip Code

Previously Clustered Records

Input Record
Zip = 72202

Zip Block 72201
(Only records with zip code 72201)

Zip Block 72202

Zip Block 72203

Input record only compared to records in Zip Block 72202

Sept 26-29, 2016
(c) 2016 iCDO@UALR
Compound Match-Key Index Example

- Attributes – FirstName, LastName
- Hash Algorithms
  - First Name: LeftSubstring(1)
  - Last Name: LeftSubstring(5)
- Example
  - FirstName: “PHILIP” → LeftSubstring(1) → ”P”
  - LastName: “DOE” → LeftSubstring(5) → ”DOE”
  - Match Key = “P” + ”DOE” = “PDOE”
New IT Paradigm of Big Data

• Move processes to data, not data to processes
• Ingest data first, then analyze and determine model, not design model first and force data to fit
• Parse and structure data on output, not on input
• De-Normalized key-value pair data stores, not normalized entity-relation schemas
• Implicit, middleware parallelism, not explicit coding
Entity Resolution is a (Noisy) Graph Problem

Simple Undirected Graph

Match Key Graph

Match Key Generator 1
Match Key Generator 2
Match Key Generator 3
Pre-Resolution Transitive Closure in Hadoop Map/Reduce
Post-Resolution Transitive Closure

- Hadoop M/R Entity Resolution on Match Key 1
- Hadoop M/R Entity Resolution on Match Key 2
- Hadoop M/R Entity Resolution on Match Key 3
- Final EIS
Incremental Transitive Closure

Sort by Match Key 1

Hadoop M/R Entity Resolution

EIS1

Transitive Closure of Entity Identifiers from EIS 1 and Match Key 2

Hadoop M/R Entity Resolution

EIS2

Transitive Closure of Entity Identifiers from EIS 2 and Match Key 3

Hadoop M/R Entity Resolution

Final EIS
Thank You!

- John R. Talburt
- jrtalburt@ualr.edu