

## Master Data Management for Big Data

John R. Talburt, PhD, IQCP, CDMP  
 Black Oak Analytics, Inc, USA  
 MIT International Conference on Information Quality  
 Workshop June 21, 2016, Ciudad Real, Spain

---

---

---

---

---

---

---

---

### My Background

- Currently
  - Chief Scientist for Black Oak Analytics, Inc.
  - Professor of Information Science and Coordinator for the Information Quality Graduate Program at the University of Arkansas at Little Rock (UALR)
- Previously
  - Business Leader for Data Research and Development at Axiom Corporation

© Black Oak Analytics

2



---

---

---

---

---

---

---


---

### Talk Outline

- Business Case for MDM
- Technical Foundations of MDM
  - Entity Resolution
  - Entity Identity Information Management
  - Master Data Management
- The Need for Entity Resolution Analytics
- Investing in Clerical Review for Continuous Improvement
- Large-Scale MDM Using Distributed Processing

© Black Oak Analytics

3



---

---

---

---

---

---

---

---

## The Value Proposition for MDM

© Black Oak Analytics

4




---

---

---

---

---

---

---

---

## The Business Case for MDM

- ◆ Customer Satisfaction and Entity-Based Data Integration
- ◆ Better Service
- ◆ Reducing the Cost of Poor Data Quality
- ◆ MDM as Part of Data Governance

© Black Oak Analytics

5




---

---

---

---

---

---

---

---

## Customer Satisfaction

- ◆ MDM has its roots in the customer relationship management (CRM) industry.
- ◆ The primary goal of CRM is to improve the customer's experience and increase customer satisfaction
- ◆ The business motivation for CRM is to
  - Increase customer retention rates
  - Lower customer "churn rate"
  - Gain new customers gained through social networking and referrals from satisfied customers.
  - Costs less to keep a customer than to acquire a new customer

© Black Oak Analytics

6




---

---

---

---

---

---

---

---

## Better Service

- ◆ Healthcare
  - Improved clinical care, complete view patient encounters
  - Improved medical research, find related cases
  - The value proposition is "better quality of life"
- ◆ Law Enforcement
  - Many entity types- suspects, autos, airplanes, boats, phones, places, ...
  - Helps to bridge the many disparate and autonomous jurisdictions
  - The value is more efficient and more effective investigation – cases closed

© Black Oak Analytics

7




---

---

---

---

---

---

---

---

## Reducing the Cost of Poor Data Quality

- ◆ A major cause of data quality problems is "multiple source of the same information produce different values for this information."
  - Lee, et al, "Journey to Data Quality"
- ◆ A result of missing or ineffective MDM practices.
- ◆ Taguchi's Loss Function - the cost of poor data quality must be considered not only in the effort to correct the immediate problem but also include all of the costs from its downstream effects.
- ◆ MDM is considered fundamental to an enterprise data quality program

© Black Oak Analytics

8




---

---

---

---

---

---

---

---

## MDM as Part of Data Governance (DG)

- ◆ DG is a program for managing information as an enterprise asset
- ◆ DG provides a single-point of communication and control over information in the enterprise
- ◆ DG has created new management roles devoted to data and information
  - CDO, Chief Data Officer
  - Data Stewards
  - MDM and Reference Data Management (RDM) are regarded as essential components of mature DG programs

© Black Oak Analytics

9




---

---

---

---

---

---

---

---

## Technical Foundations of MDM

Entity Resolution, Entity Identity Information Management, and MDM

© Black Oak Analytics

10




---

---

---

---

---

---

---

---

## Three Related Concepts

- ◆ Entity Resolution (ER)
- ◆ Entity Identity Information Management (EIIM)
- ◆ Master Data Management (MDM)



© Black Oak Analytics

11




---

---

---

---

---

---

---

---

## Entity Resolution (ER)

- ◆ The process of determining whether two references in an information system are referring to the same real-world object or to different objects (Talbert, 2011)



Record-linking  
Record-deduplication  
Data matching  
Co-reference problem  
Semantic resolution

If they refer to same real-world object, they are said to be "Equivalent"

© Black Oak Analytics

12




---

---

---

---

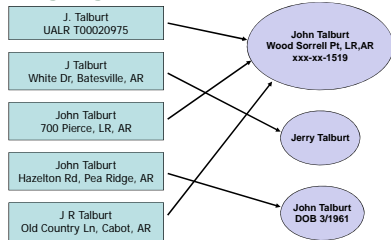
---

---

---

---

### Which belong together?



© Black Oak Analytics

13




---

---

---

---

---

---

---

---

### 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

© Black Oak Analytics

14




---

---

---

---

---

---

---

---

### 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*

© Black Oak Analytics

15




---

---

---

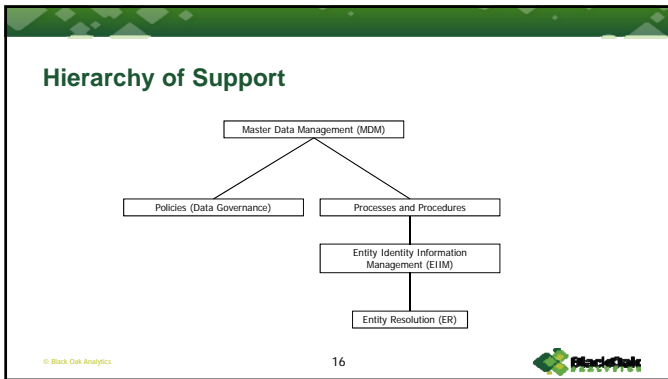
---

---

---

---

---




---

---

---

---

---

---

---

---

### Most Common MDM Mistakes Organizations Make

- ◆ Fail to quantitatively and systematically measure and improve Entity Identity Integrity achievement (Lack of QC and Continuous Improvement)
- ◆ Apply QA processes at the sourcing step, but not at the linking step (Partial QA – Lack of Review Indicators)
- ◆ Failure to address the life cycle of entity identity information
- ◆ The EIIM information architecture is inadequate
- ◆ The EIIM process is embedded in other ETL processes

© Black Oak Analytics 17

---

---

---

---

---

---

---

---

### Measuring Entity Identity Integrity

- ◆ Linking Accuracy =  $(TP+TN)/(TP+FP+TN+FN)$
- ◆ False Negative Rate =  $FN/(TP+FN)$
- ◆ False Positive Rate =  $FP/(TN+FP)$

R = set of References  $|R|=N$   
D = All pairs in R,  $|D|= N*(N-1)/2$   
E = Equivalent Pairs  
L = Pairs Linked by Process

© Black Oak Analytics 18

---

---

---

---

---

---

---

---

## Measurement Techniques

- ◆ Truth set development
  - Small, but precise and time consuming
- ◆ Benchmarking over the same dataset
  - Large and fast, but less precise
- ◆ Stratified sampling of clusters by attribute entropy
  - In between, gives reliable accuracy statistics

© Black Oak Analytics

19




---

---

---

---

---

---

---

---

## Quality Assurance at the Linking Step

- ◆ Good MDM systems should produce "clerical review indicators"
- ◆ Clerical review indicators are signals from the system that false positive or false negative errors might have been made for certain linking decisions
- ◆ Clerical review indicators are implemented as "exception reports" that should be reviewed by true domain experts who can decide if the error was made or not
- ◆ If errors were made, the experts should be able to override the system and make corrections – "continuous improvement"

© Black Oak Analytics

20




---

---

---

---

---

---

---

---

## MDM Life Cycle Management

The CRUD Model

© Black Oak Analytics

21




---

---

---

---

---

---

---

---

### CSRU 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




---

---

---

---

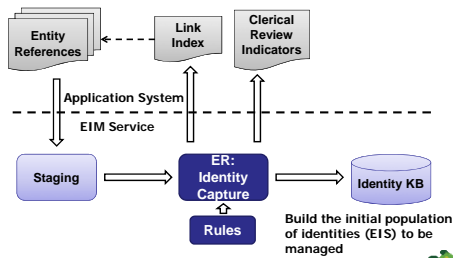
---

---

---

---

### Capture Phase in an EIMS




---

---

---

---

---

---

---

---

### 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.




---

---

---

---

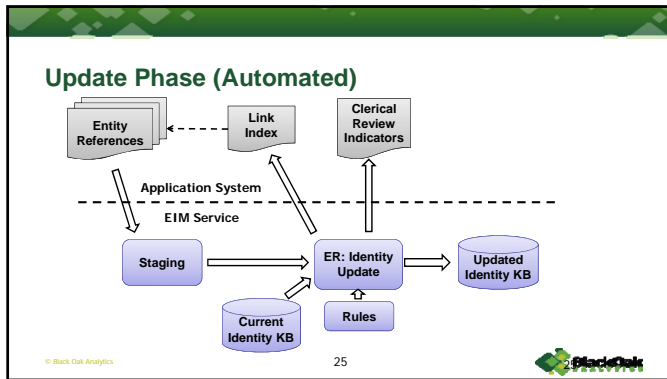
---

---

---

---






---

---

---

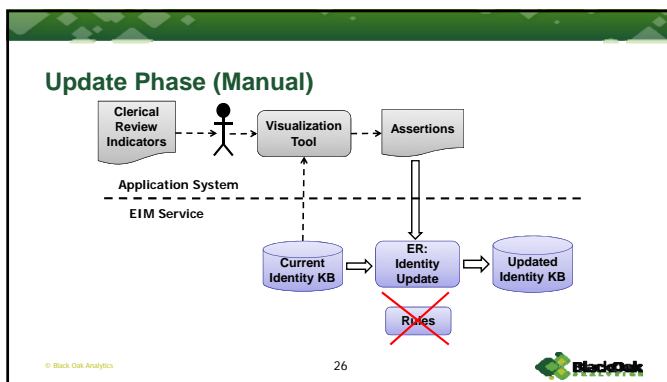
---

---

---

---

---




---

---

---

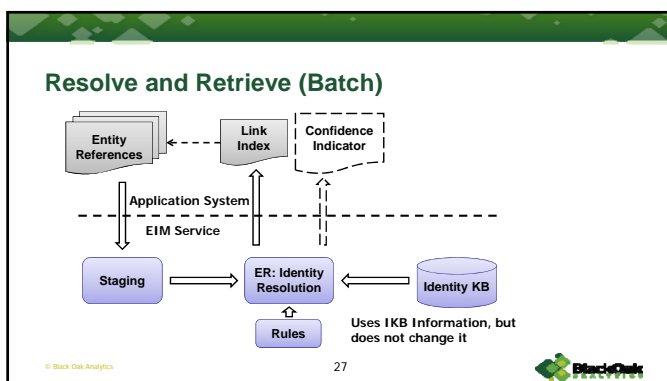
---

---

---

---

---




---

---

---

---

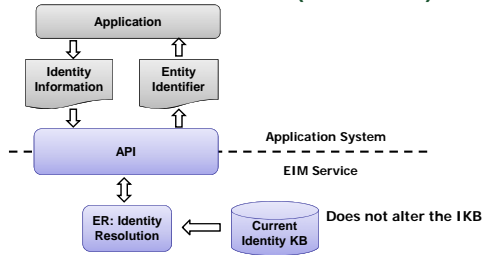
---

---

---

---

### Resolve & Retrieve Phase (Interactive)



© Black Oak Analytics

28



### 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

© Black Oak Analytics

29



### Pair- and Cluster-level Review Indicators

- Pair-Level
  - In Boolean (deterministic) systems – “Soft rules”
  - In Scoring (probabilistic) systems – “Review threshold”
- Cluster-Level
  - Cluster Entropy
  - Conflict Rules & Rationality Checks

© Black Oak Analytics

30



### Example: Rationality Check at the Cluster Level

Source 1 Passes QA checks at Source Level

Source 2 Passes QA checks at Source Level

EHR

Cluster 123

- Name: John Doe; Annual Check-up: 2015/01/15;
- Name: John Doe; Deceased: 2010/04/08;

Does not make sense at the Cluster Level

Unfortunately, many organizations only perform QA at the record level, and not at the Cluster level

© Black Oak Analytics 31

---

---

---

---

---

---

---

---

### MDM in the World of Big Data

New IT Paradigms

© Black Oak Analytics 32

---

---

---

---

---

---

---

---

### 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

© Black Oak Analytics 33

---

---

---

---

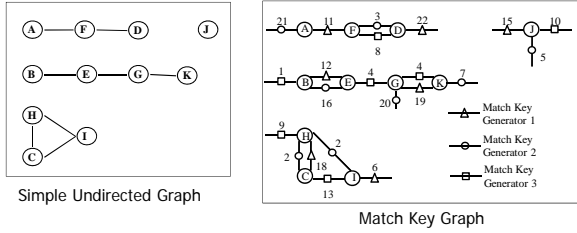
---

---

---

---

## Entity Resolution is a (Noisy) Graph Problem



© Black Oak Analytics

34




---

---

---

---

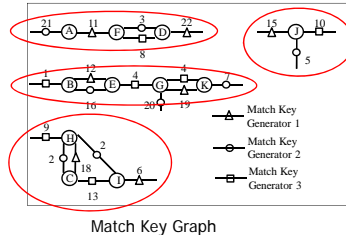
---

---

---

---

## Goal: Find the Connected Components



© Black Oak Analytics

35



Through a process  
called the  
"Transitive Closure"  
of the graph

---

---

---

---

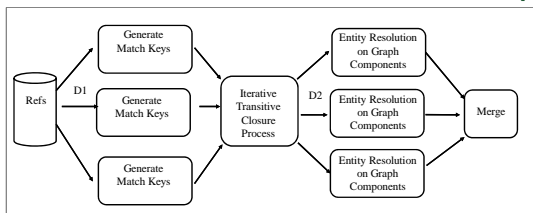
---

---

---

---

## Pre-Resolution Transitive Closure in Hadoop M/R



© Black Oak Analytics

36




---

---

---

---

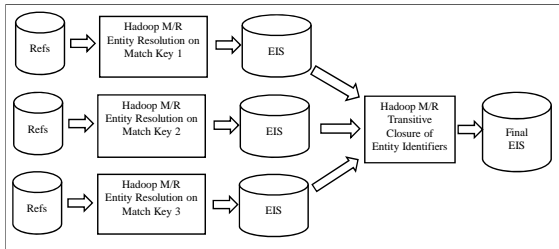
---

---

---

---

### Post-Resolution Transitive Closure

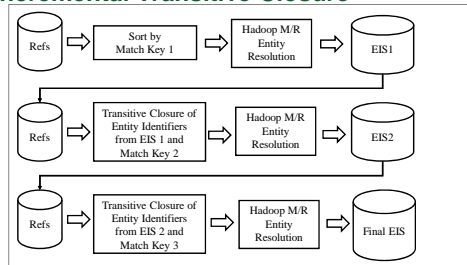


© Black Oak Analytics

37



### Incremental Transitive Closure



© Black Oak Analytics

38



### Questions and Discussion

© Black Oak Analytics

39

