BlackOak ANALYTICS

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 Acxiom Corporation

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

The Value Proposition for MDM

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

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

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

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

MDM as Part of Data Governance (DG)

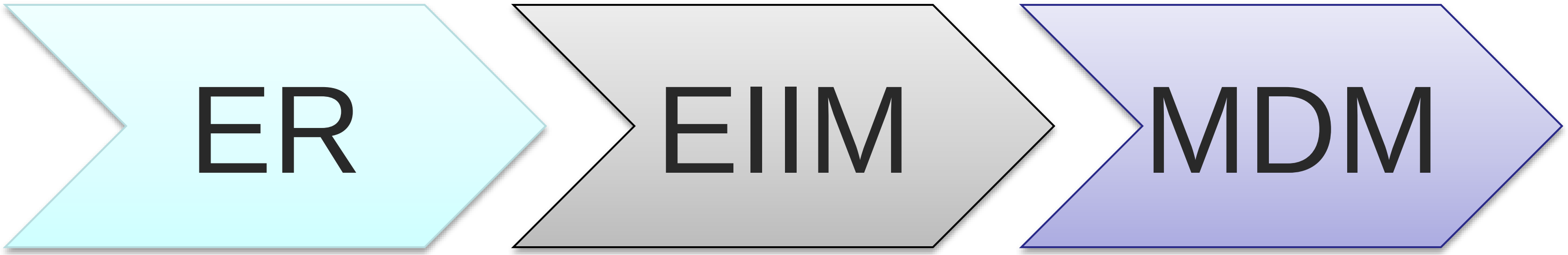
- DG is a program for managing information as an enterprise asset
- DG provides a single-point of communication and control over information in the enterprise
- DG has created new management roles devoted to data and information
  - CDO, Chief Data Officer
  - Data Stewards
  - MDM and Reference Data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

Technical Foundations of MDM

Entity Resolution, Entity Identity Information Management, and MDM

Three Related Concepts

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



ER → EIIM → MDM

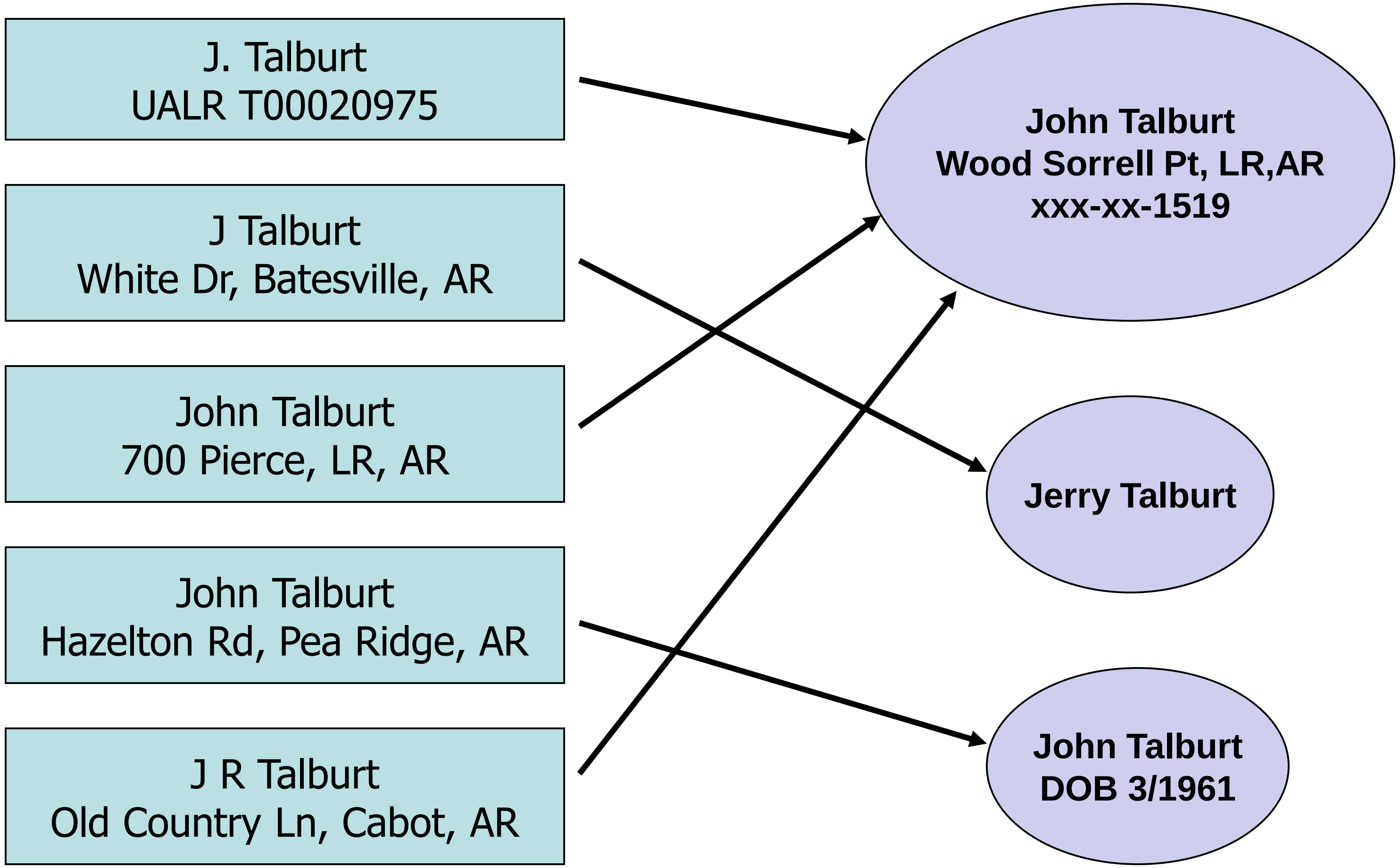
Entity Resolution (ER)

- The process of determining whether two references in an information system are referring to the same real-world object or to different objects (Talburt, 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**"

Which belong together?



J. Talburt
UALR T00020975

J Talburt
White Dr, Batesville, AR

John Talburt
700 Pierce, LR, AR

John Talburt
Hazelton Rd, Pea Ridge, AR

J R Talburt
Old Country Ln, Cabot, AR

John Talburt
Wood Sorrell Pt, LR,AR
xxx-xx-1519

Jerry Talburt

John Talburt
DOB 3/1961

Entity Identity Information Management (EIIM)

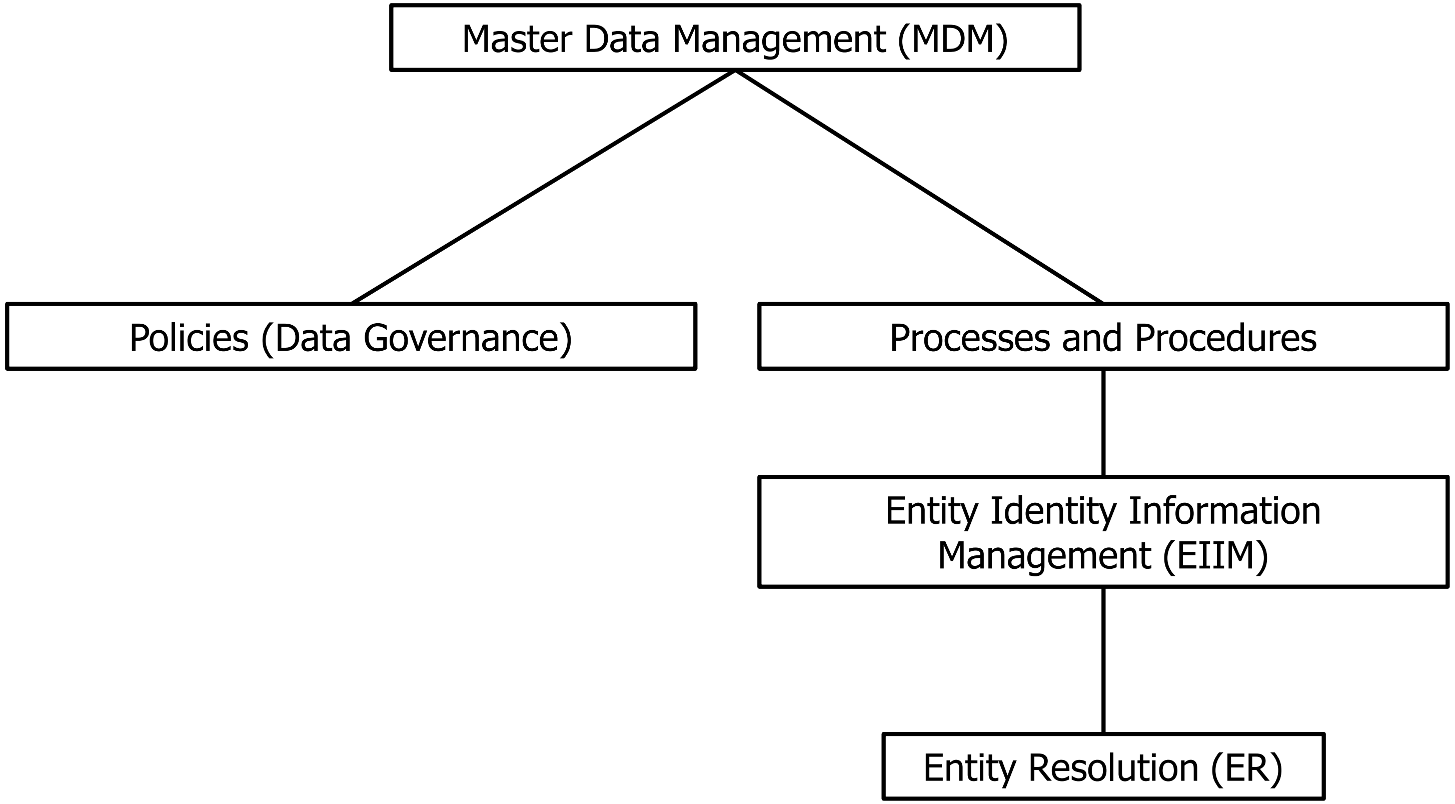
- An extension of ER in two dimensions
  - Knowledge management
    - Creating, storing, and managing the information that represents the identity of an entity
    - Entity Identity Structure (EIS)
  - Temporal
    - Maintain persistent entity identifiers over time, i.e. process to process
- Essential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

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*

Hierarchy of Support


Master Data Management (MDM)

Policies (Data Governance)

Processes and Procedures

Entity Identity Information Management (EIIM)

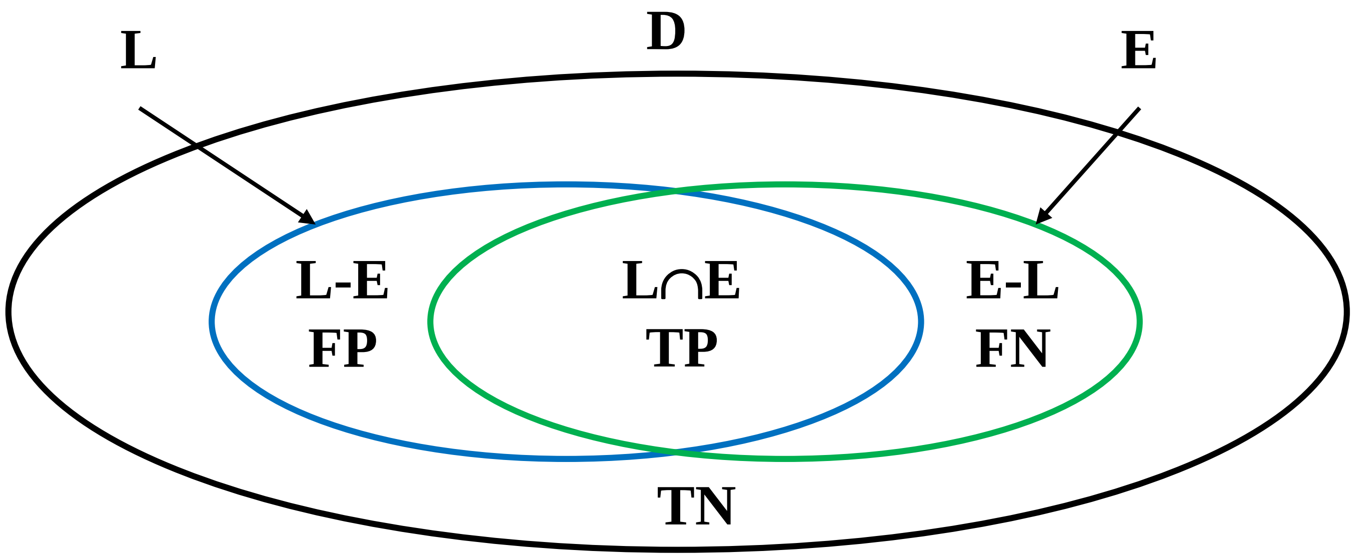
Entity Resolution (ER)

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

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

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

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”

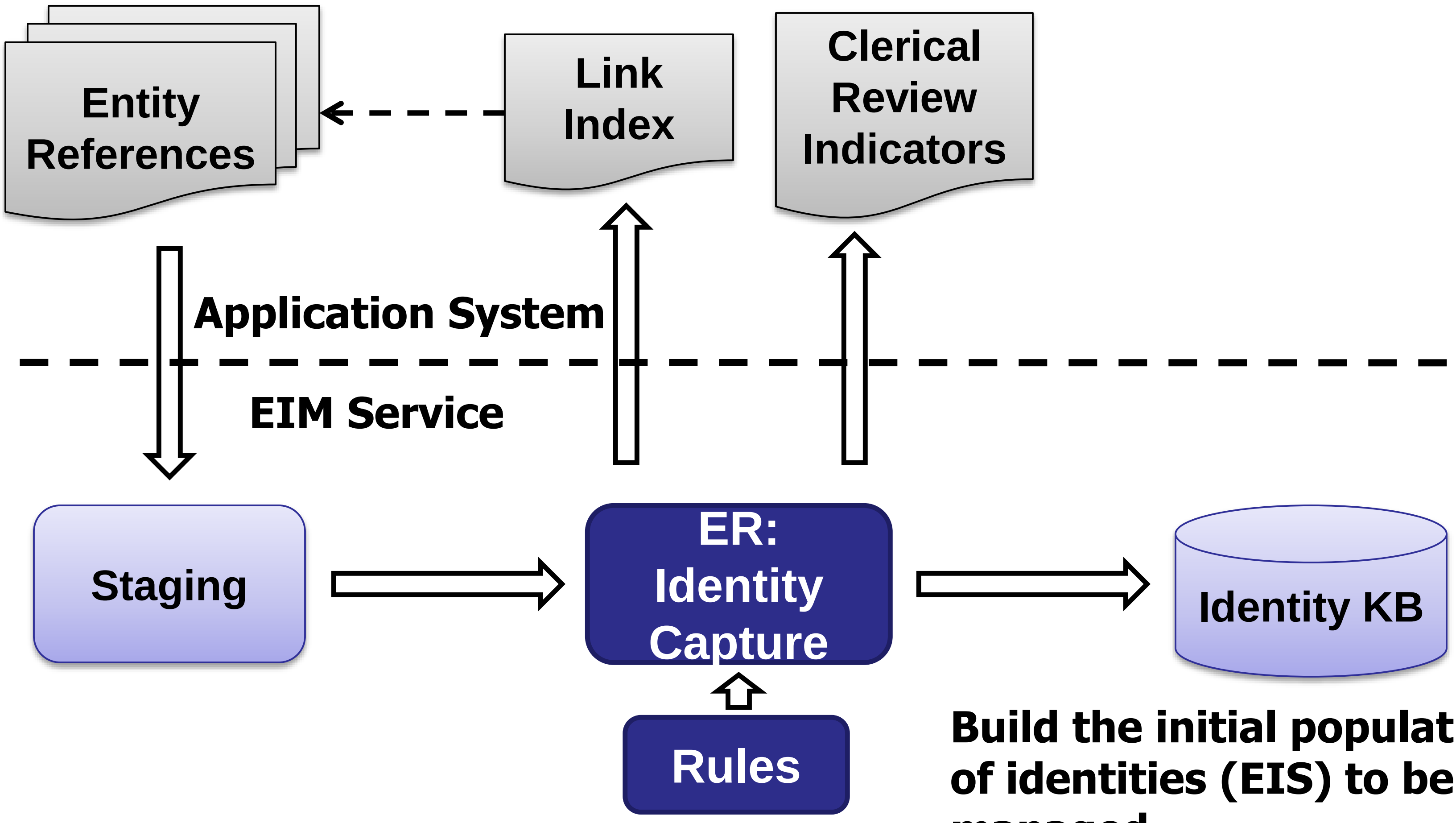
MDM Life Cycle Management

The CSRUD Model

CSRUD 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

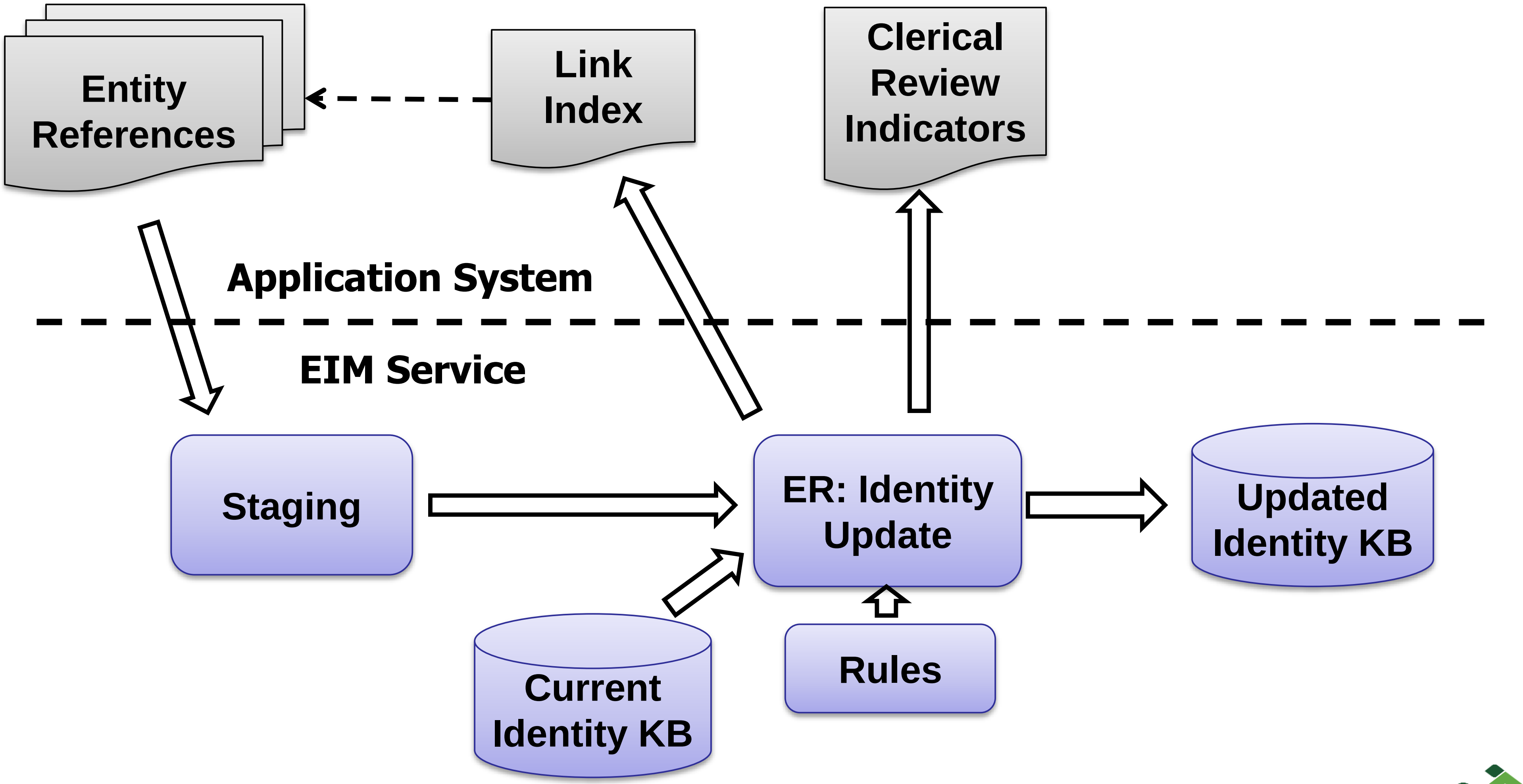


Build the initial population of identities (EIS) to be managed

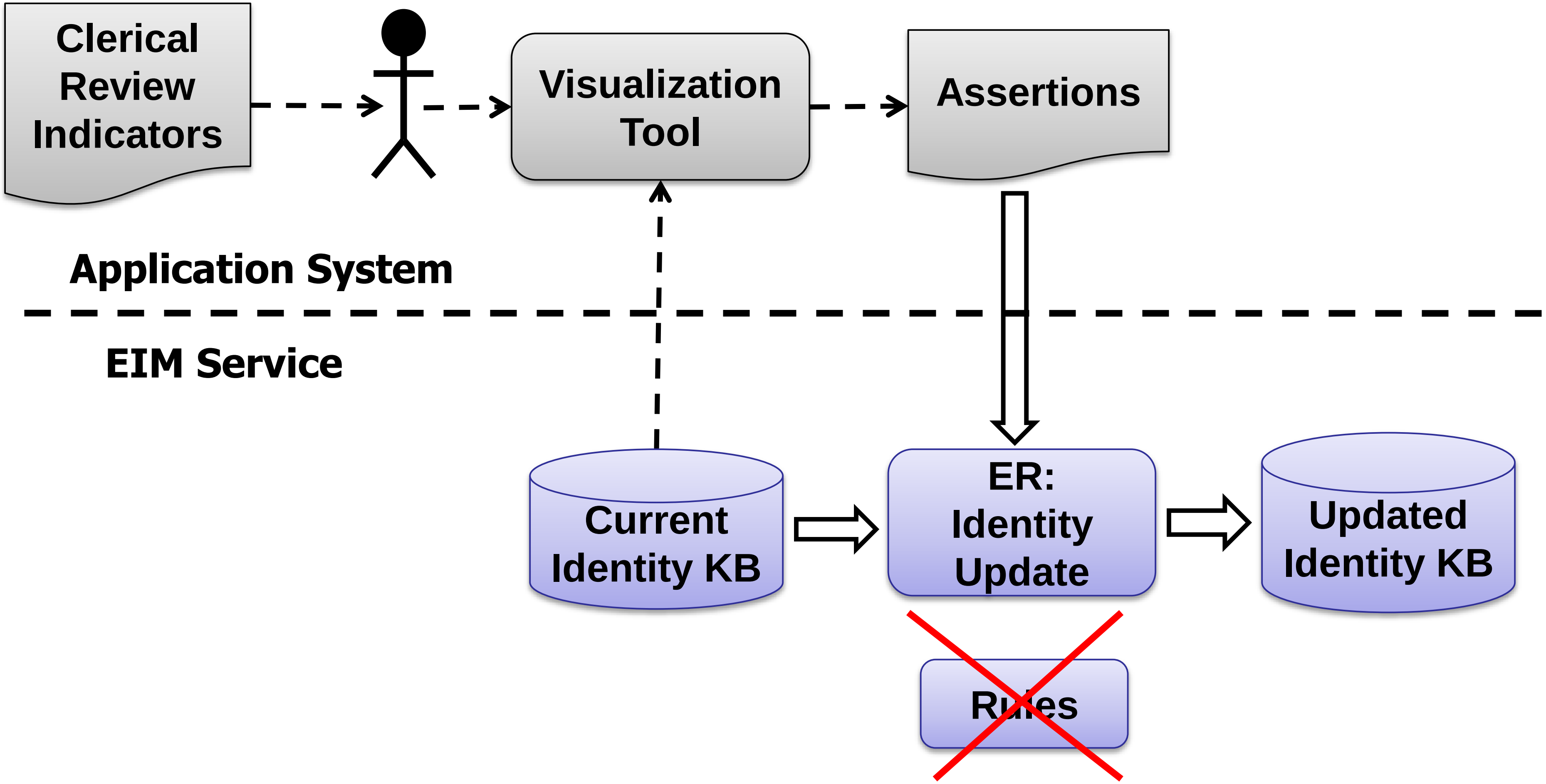
Store & Share Phase

- 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.

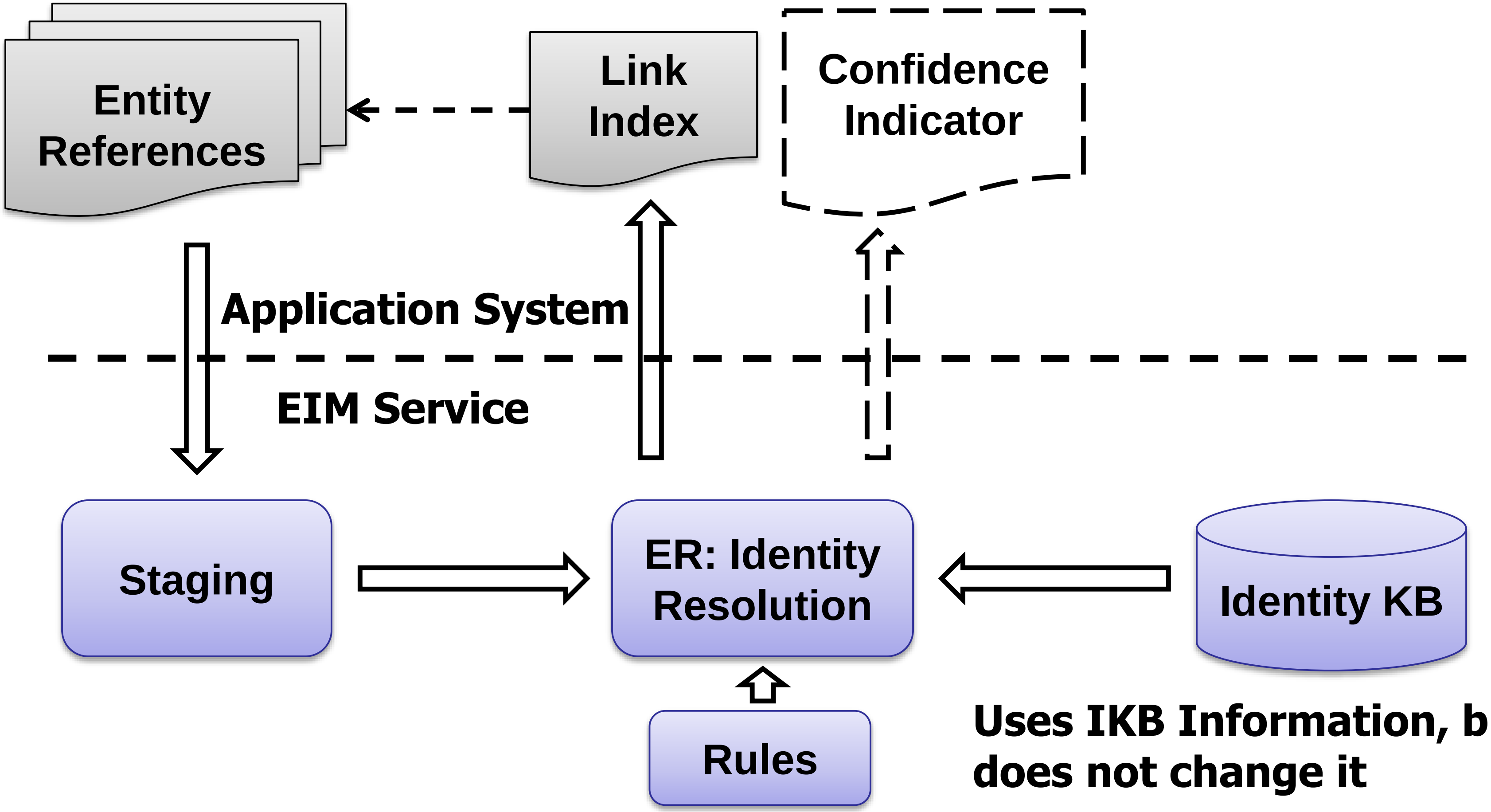
Update Phase (Automated)



Update Phase (Manual)

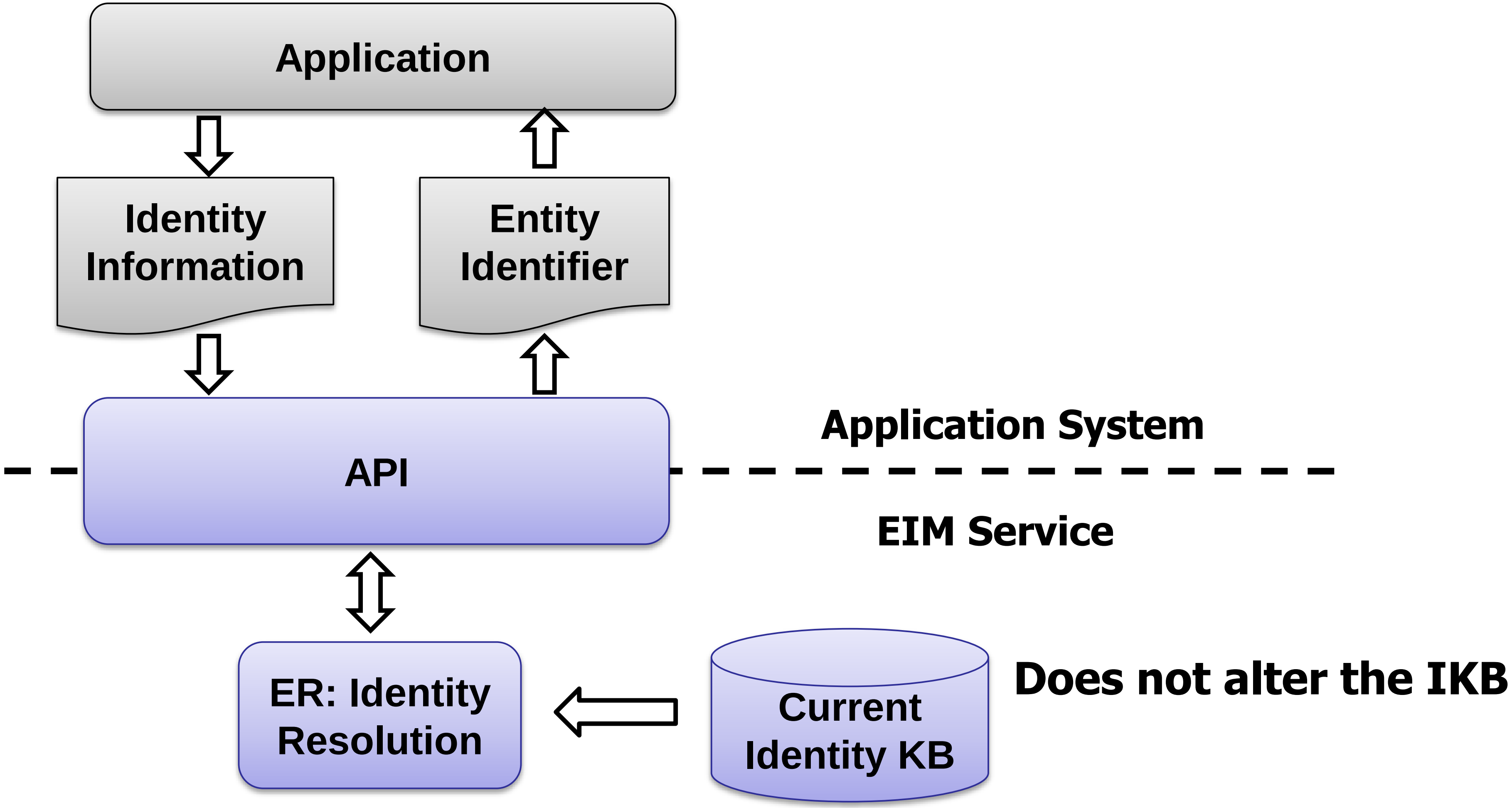


Resolve and Retrieve (Batch)



Uses IKB Information, but does not change it

Resolve & Retrieve Phase (Interactive)



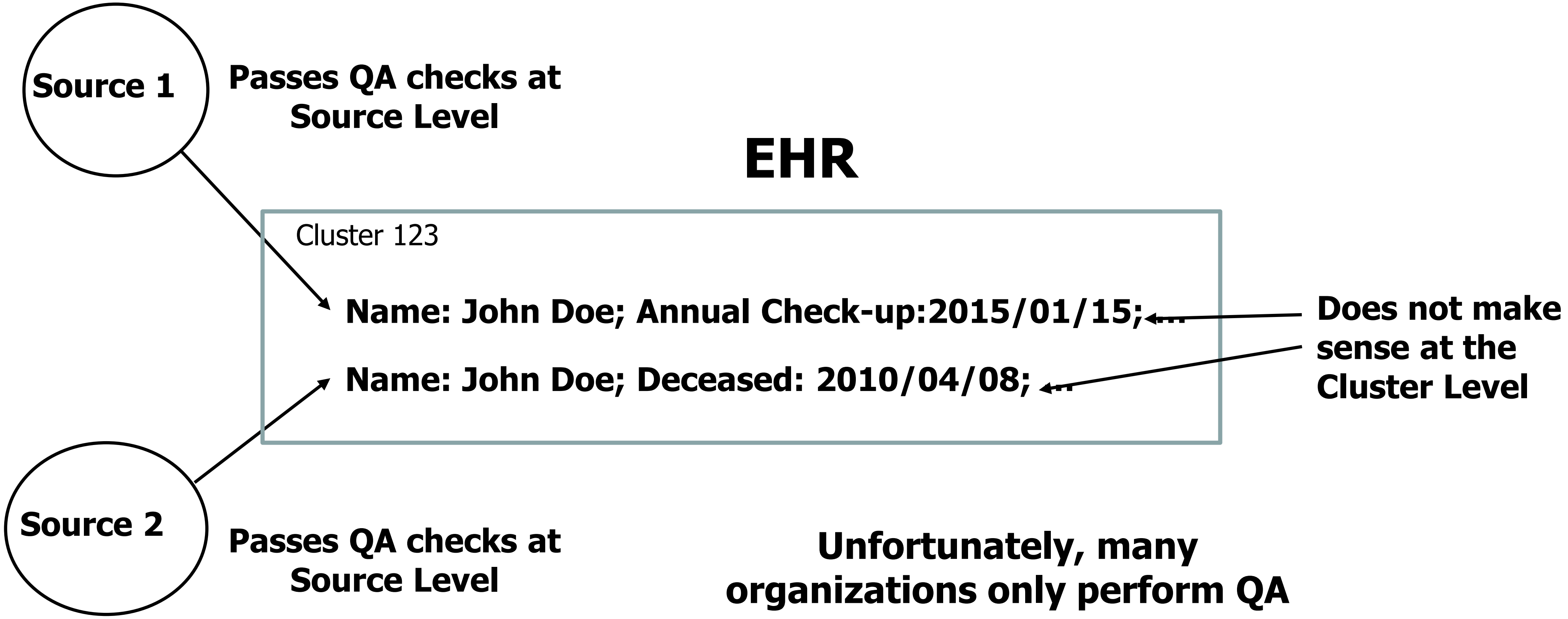
Dispose (Retire) Phase

- 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

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

Example: Rationality Check at the Cluster Level



Source 1

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

Source 2

Passes QA checks at Source Level

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

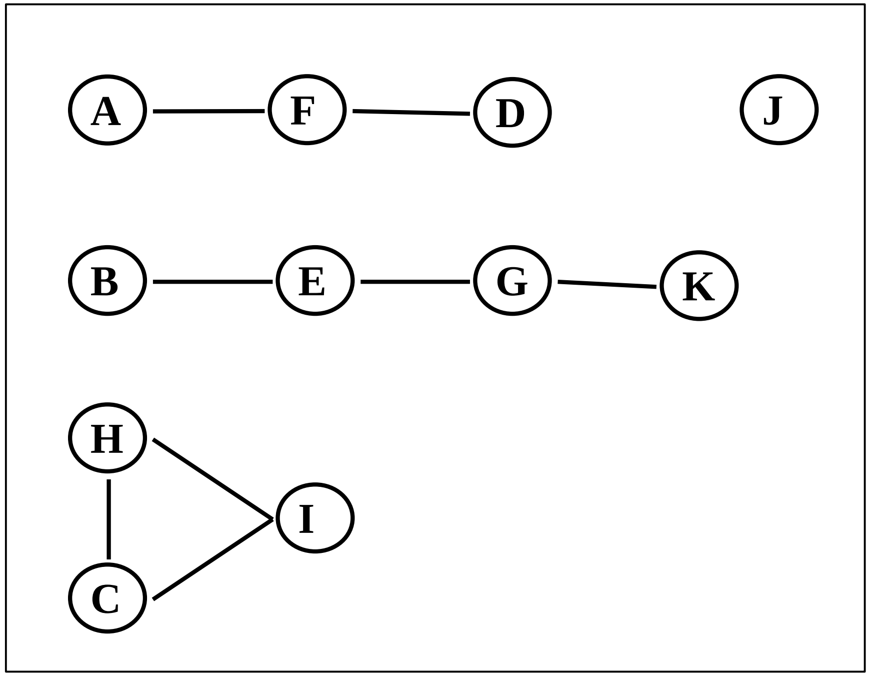
MDM in the World of Big Data

New IT Paradigms

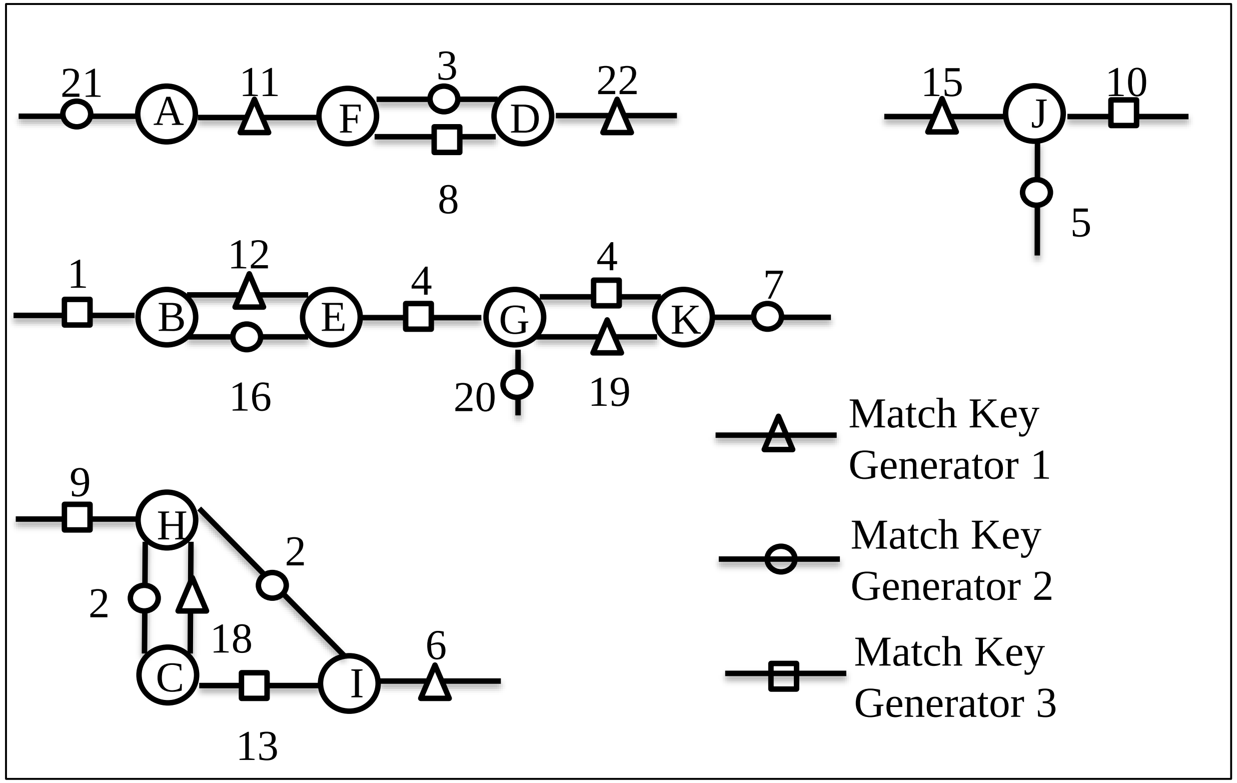
New IT Paradigm of Big Data

- 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

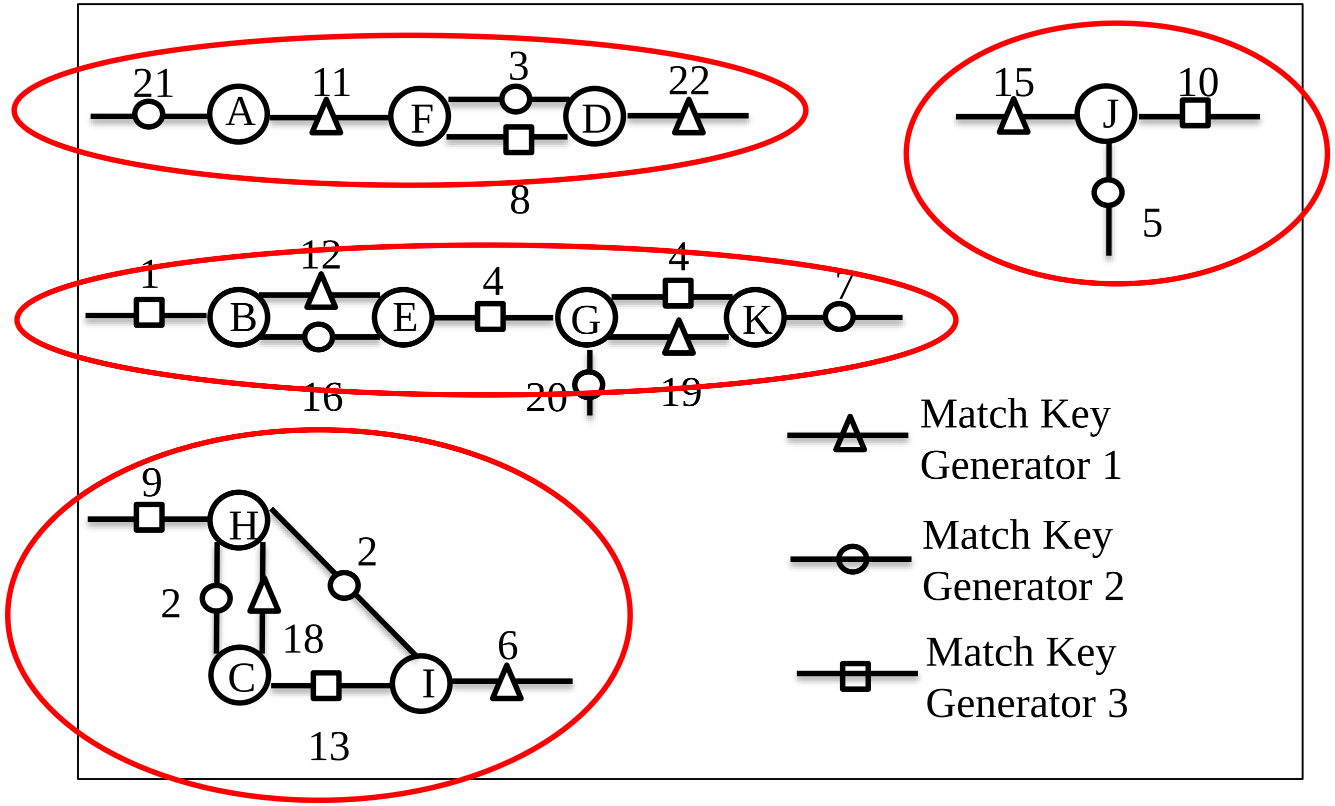


Simple Undirected Graph



Match Key Graph

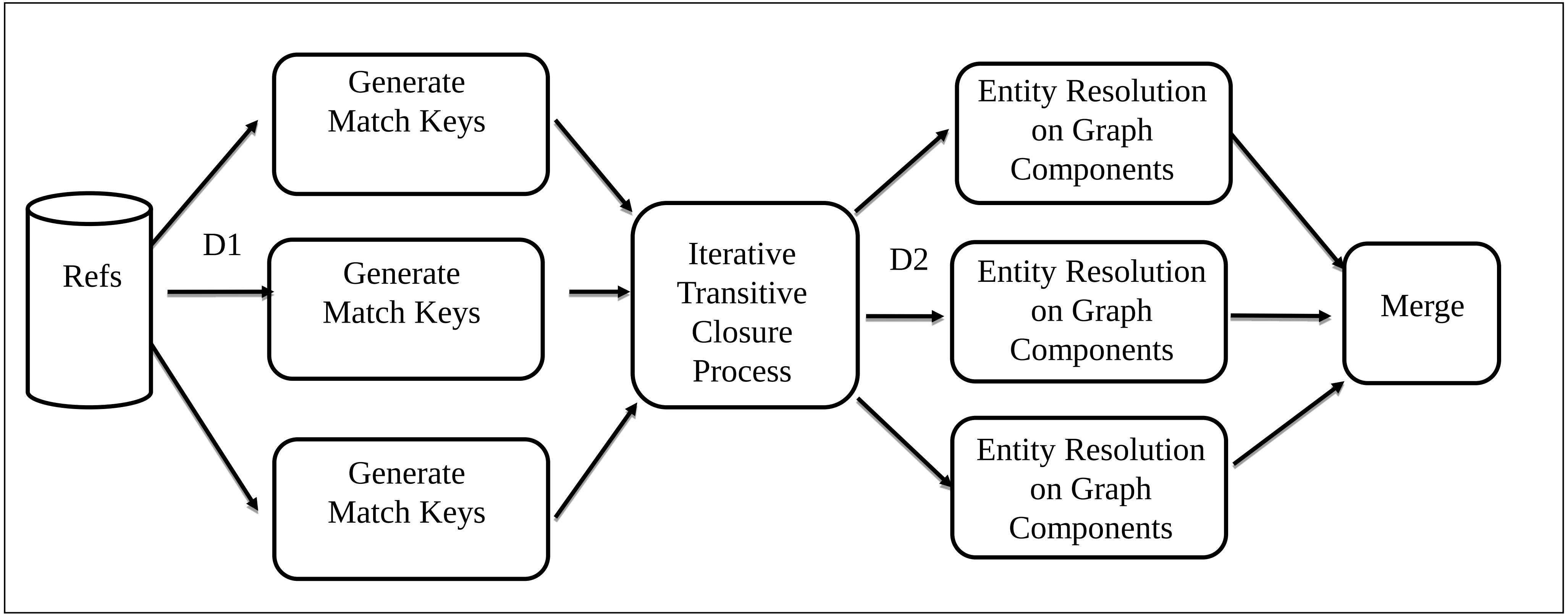
Goal: Find the Connected Components



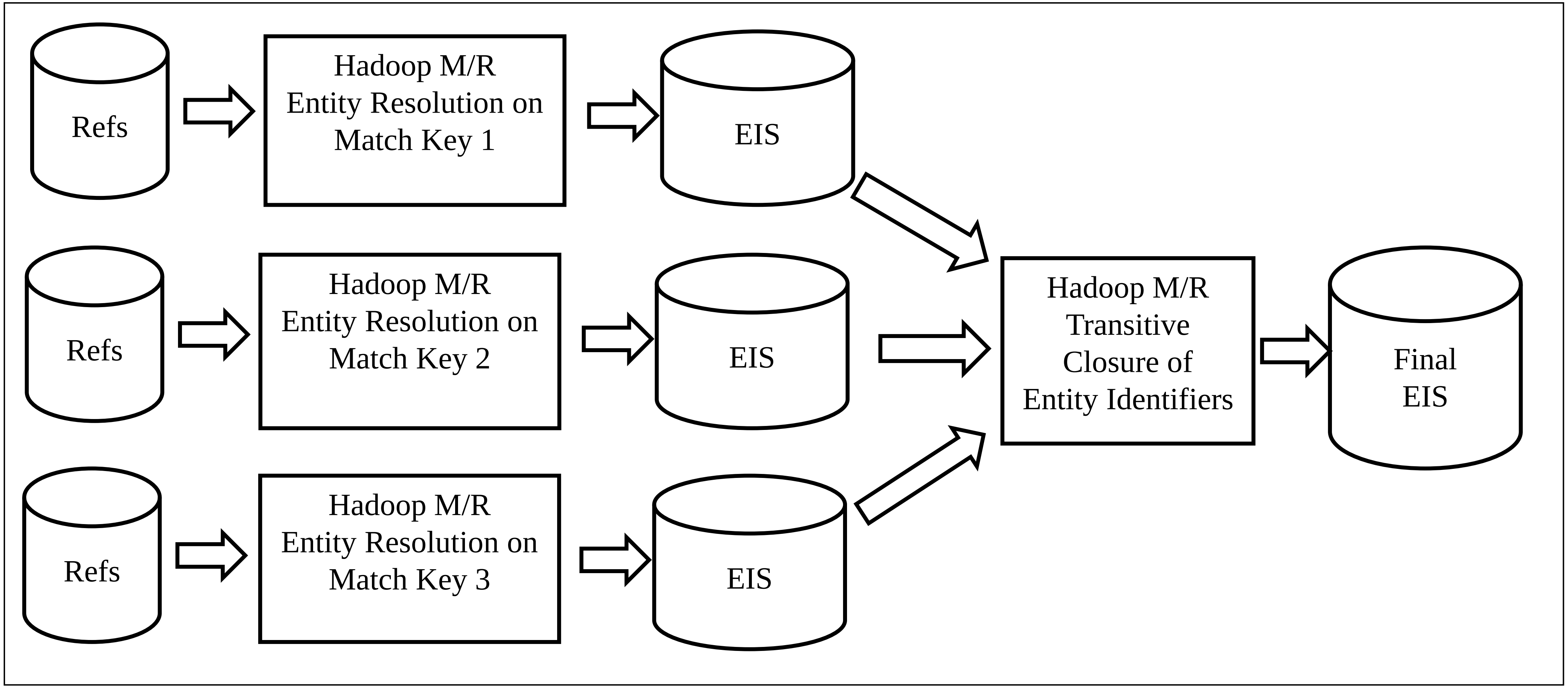
Match Key Graph

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

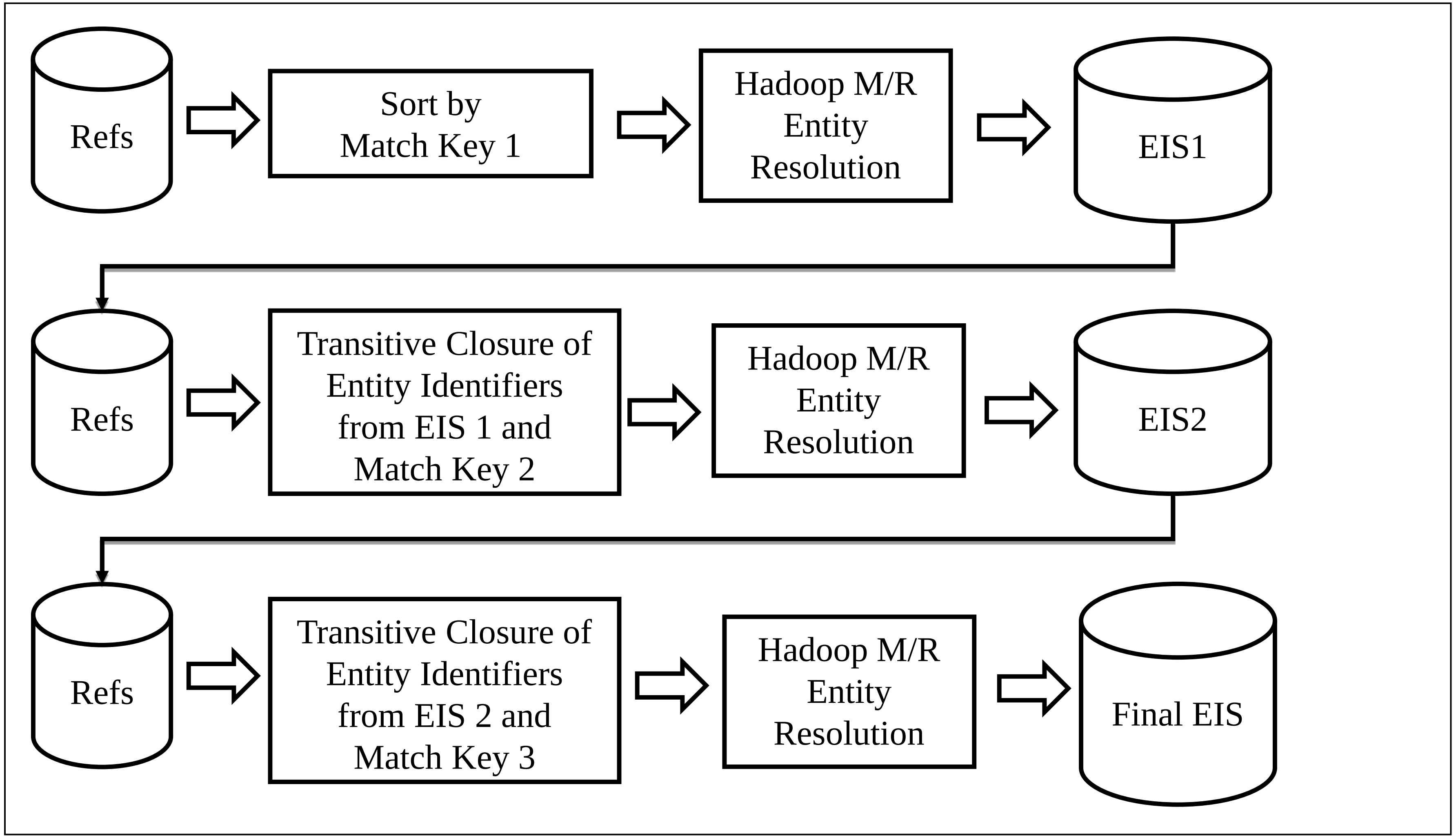
Pre-Resolution Transitive Closure in Hadoop M/R


Refs
D1
Generate Match Keys
Generate Match Keys
Generate Match Keys
Iterative Transitive Closure Process
D2
Entity Resolution on Graph Components
Entity Resolution on Graph Components
Entity Resolution on Graph Components
Merge

Post-Resolution Transitive Closure


Refs
Hadoop M/R Entity Resolution on Match Key 1
EIS
Refs
Hadoop M/R Entity Resolution on Match Key 2
EIS
Refs
Hadoop M/R Entity Resolution on Match Key 3
EIS
Hadoop M/R Transitive Closure of Entity Identifiers
Final EIS

Incremental Transitive Closure


Refs
Sort by Match Key 1
Hadoop M/R Entity Resolution
EIS1
Refs
Transitive Closure of Entity Identifiers from EIS 1 and Match Key 2
Hadoop M/R Entity Resolution
EIS2
Refs
Transitive Closure of Entity Identifiers from EIS 2 and Match Key 3
Hadoop M/R Entity Resolution
Final EIS

Questions and Discussion