Dhan Patel :- Senior Product Manager Information Governance :- IBM Software Group

The Essentials of Data Discovery: Do You Know Where Your Data Is?

Important Disclaimer

THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY.

WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED.

IN ADDITION, THIS INFORMATION IS BASED ON IBM'S CURRENT PRODUCT PLANS AND STRATEGY, WHICH ARE SUBJECT TO CHANGE BY IBM WITHOUT NOTICE.

IBM SHALL NOT BE RESPONSIBLE FOR ANY DAMAGES ARISING OUT OF THE USE OF, OR OTHERWISE RELATED TO, THIS PRESENTATION OR ANY OTHER DOCUMENTATION.

NOTHING CONTAINED IN THIS PRESENTATION IS INTENDED TO, OR SHALL HAVE THE EFFECT OF:

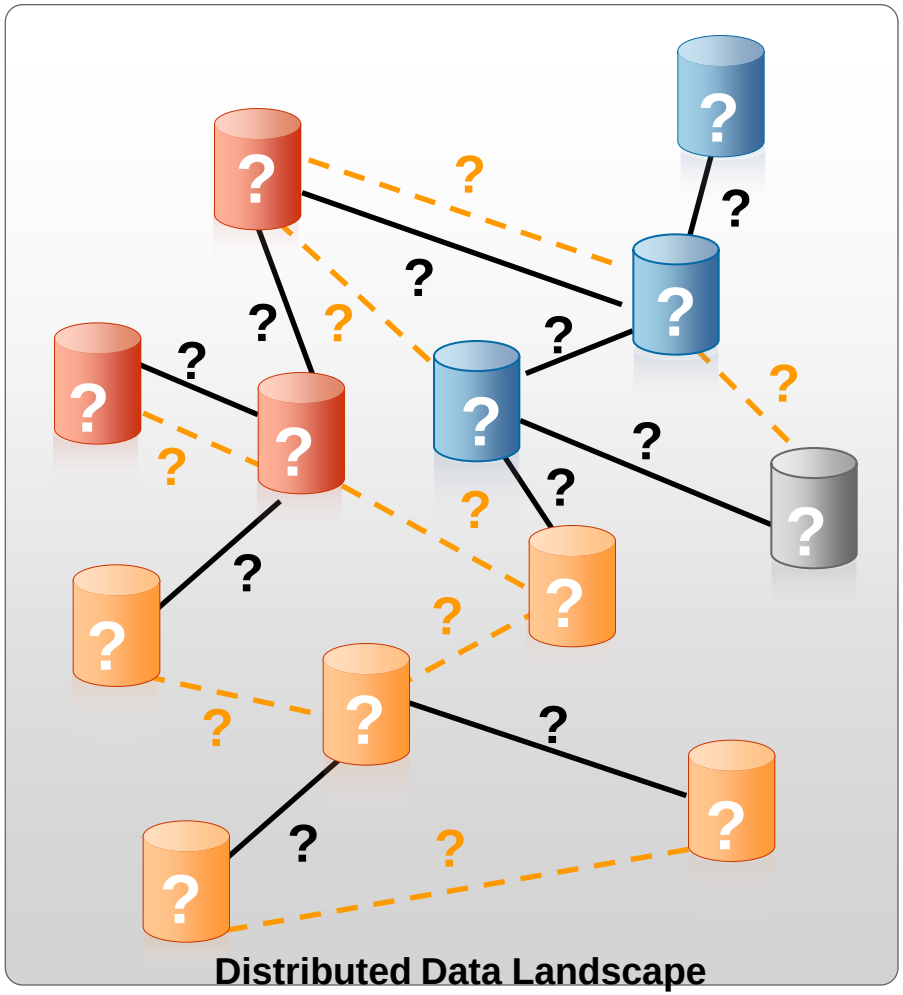
- CREATING ANY WARRANTY OR REPRESENTATION FROM IBM (OR ITS AFFILIATES OR ITS OR THEIR SUPPLIERS AND/OR LICENSORS); OR
- ALTERING THE TERMS AND CONDITIONS OF THE APPLICABLE LICENSE AGREEMENT GOVERNING THE USE OF IBM SOFTWARE.

Agenda

- Overview
- Discovery Deep Dive
  - Referential Integrity & Business Object discovery
  - Transformation discovery
  - Sensitive/Critical Data discovery
  - Discovery for Data Consolidation
- Customer Case Studies

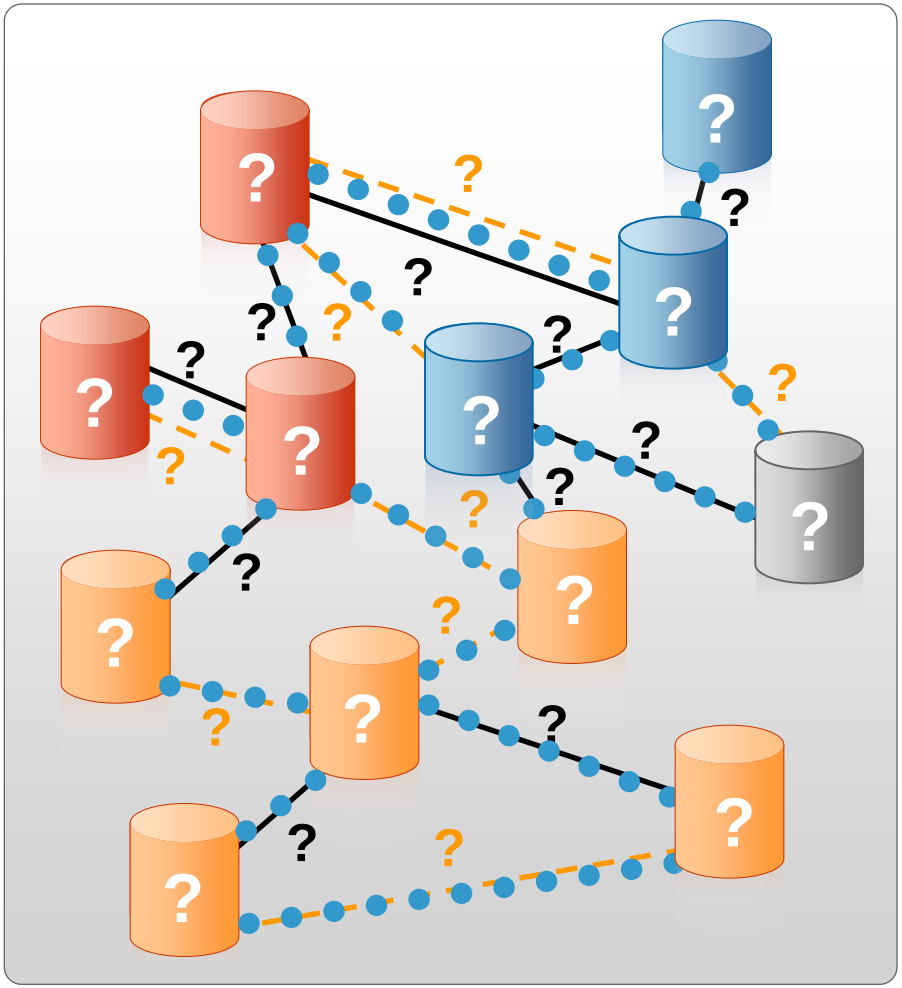
Overview

You can't --------- what you don't understand



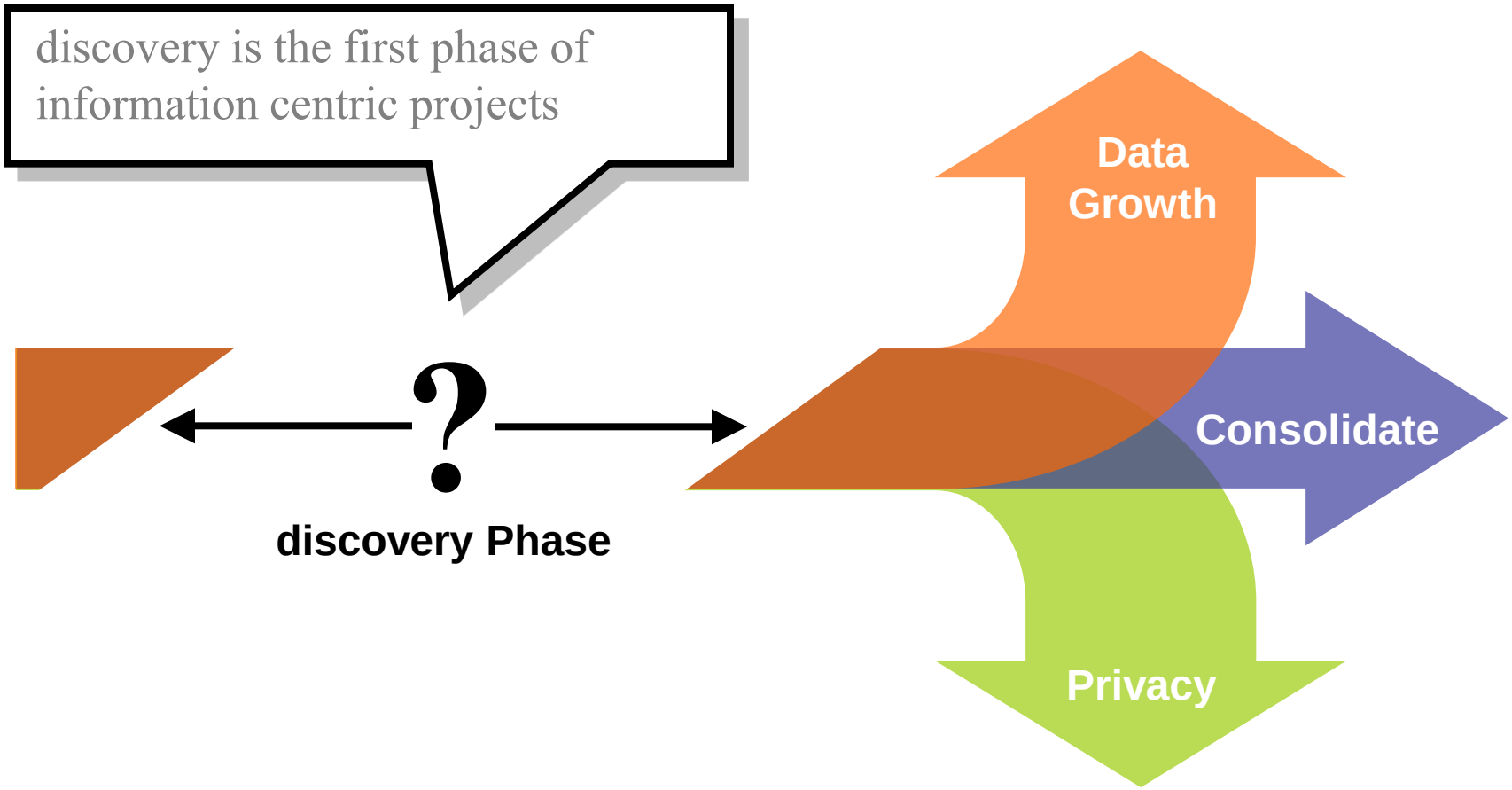
- Increasingly distributed
- Complex, poorly documented data & relationships within & across sources
- Data not understood because:
  - Corporate memory is poor
  - Documentation is poor or nonexistent
  - Logical relationships (enforced through application logic or business rules) are ***hidden***

Data Discovery: An Invaluable Data Analysis Tool



- Automated analysis of data and relationships for ***complete understanding*** of data assets:
  - ***Characterizes*** data elements ***within*** a Source
  - Identifies ***relationships*** that link data elements into "business entities" ***within*** a source
    - Customer, counterparty, invoice
  - Identifies ***complex logic*** that relates ***multiple*** sources

Poor Understanding = Unpredictable Project Deployment



Data Discovery:
Automation that accelerates time to value:

Data Growth Management:

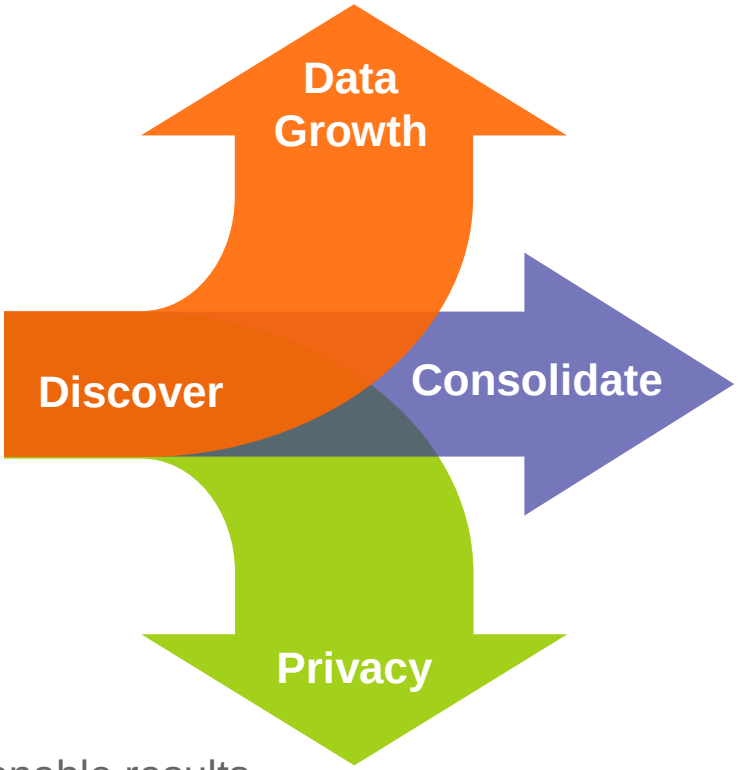
- Automates discovery of referential integrity and business objects

Data Consolidation, Integration & Migration :

- Discovers transformation and business logic between data sources
- Prototypes empty targets from the combination of many data sources

Data Privacy:

- Discovers hidden sensitive data



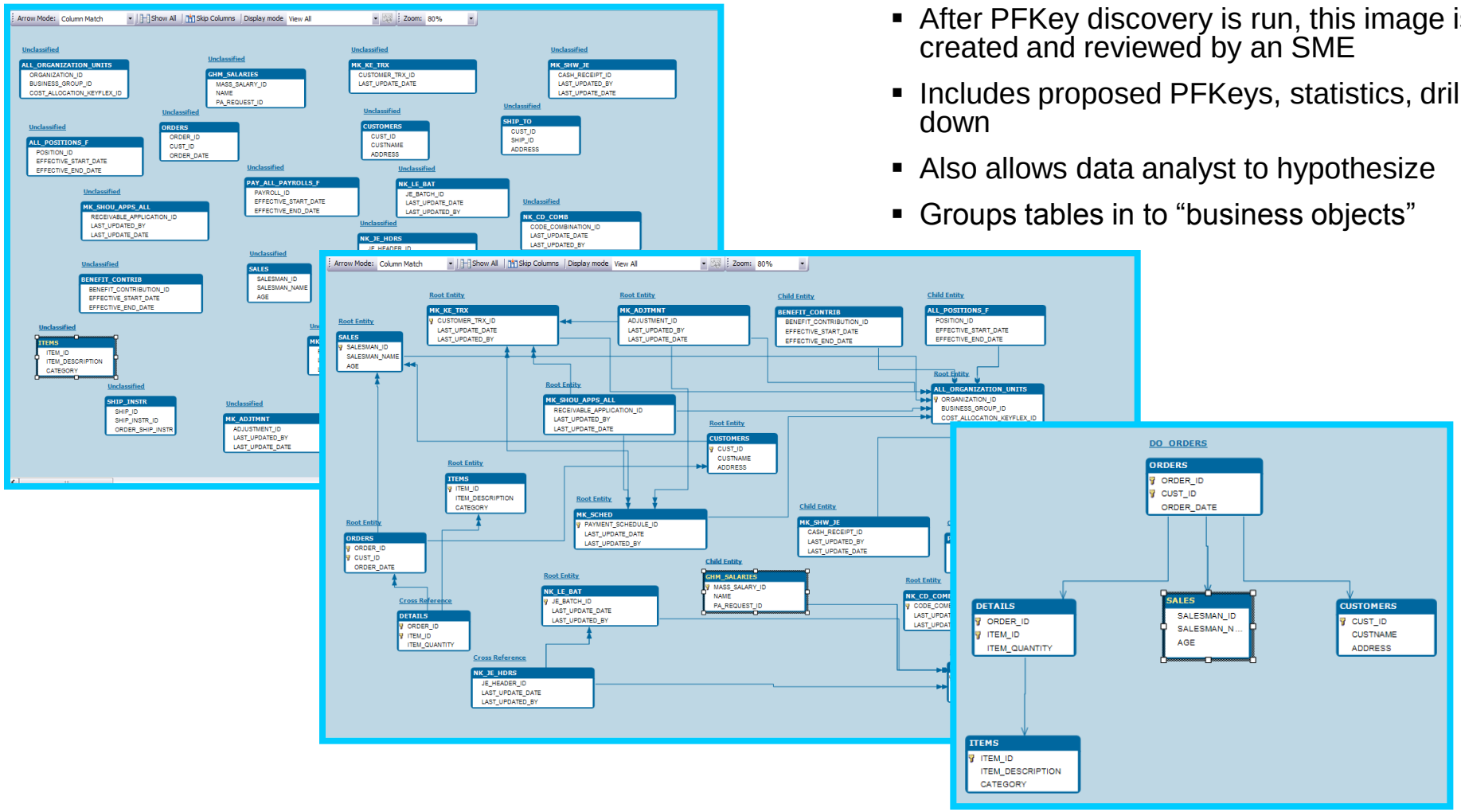
What is unique?

- Analyzes data values and patterns and produces actionable results
- Discovers complex relationships within and between data sources
- Patented approach with greatest level of automation in the industry

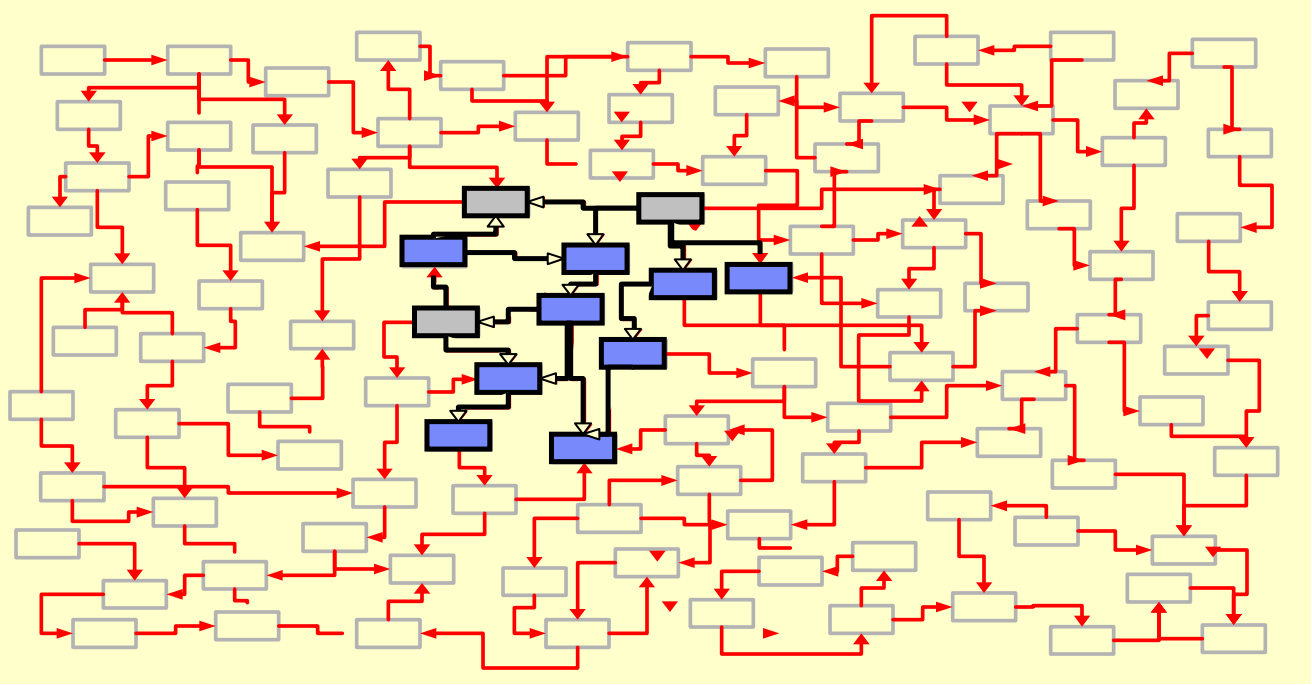
Discovery Deep Dive

Referential Integrity & Business Object Discovery

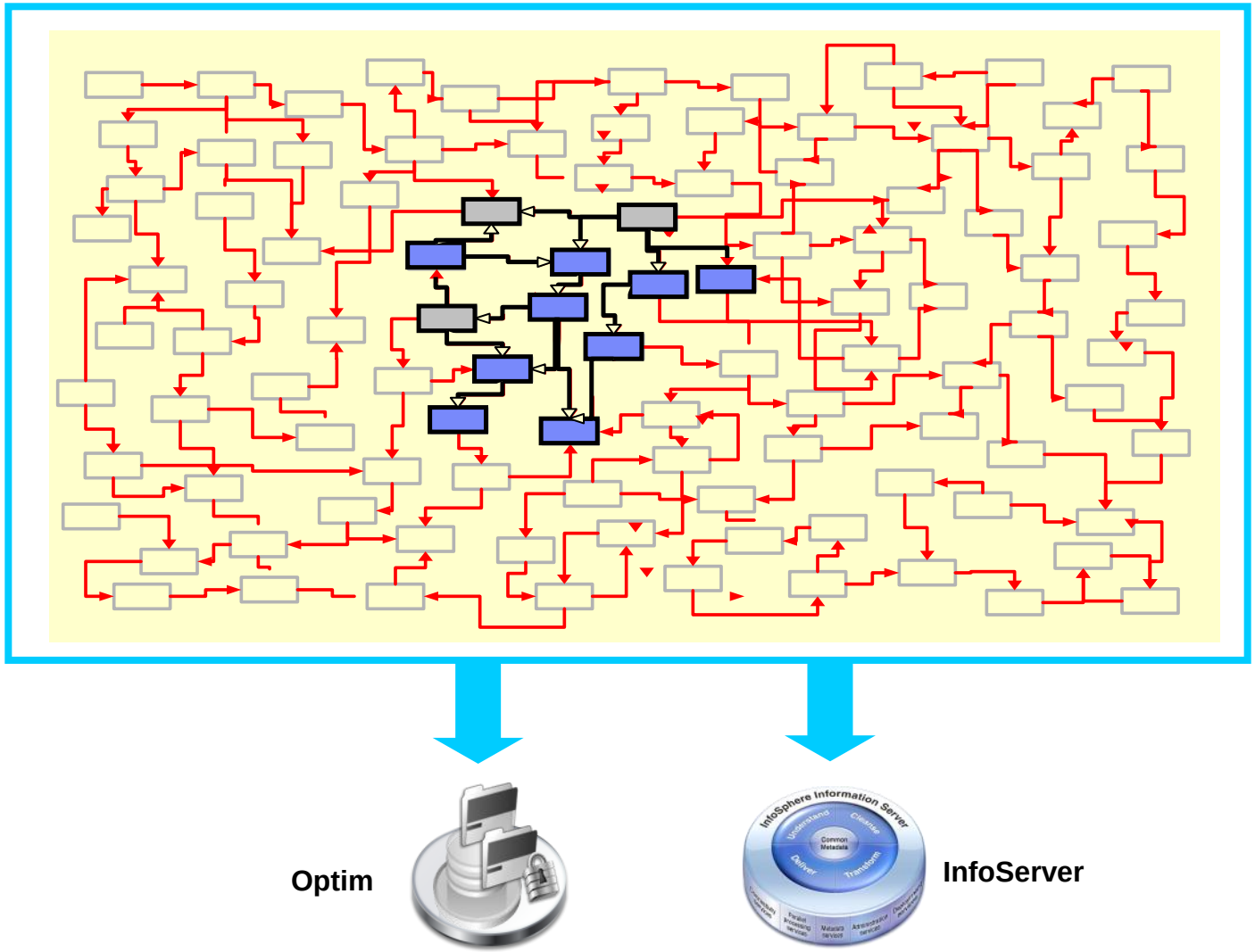
- 21 tables but what is the referential integrity?
- After PFKey discovery is run, this image is created and reviewed by an SME
- Includes proposed PFKeys, statistics, drill down
- Also allows data analyst to hypothesize
- Groups tables in to “business objects”



Referential Integrity & Business Object Discovery: From knowing nothing to knowing what you need to know

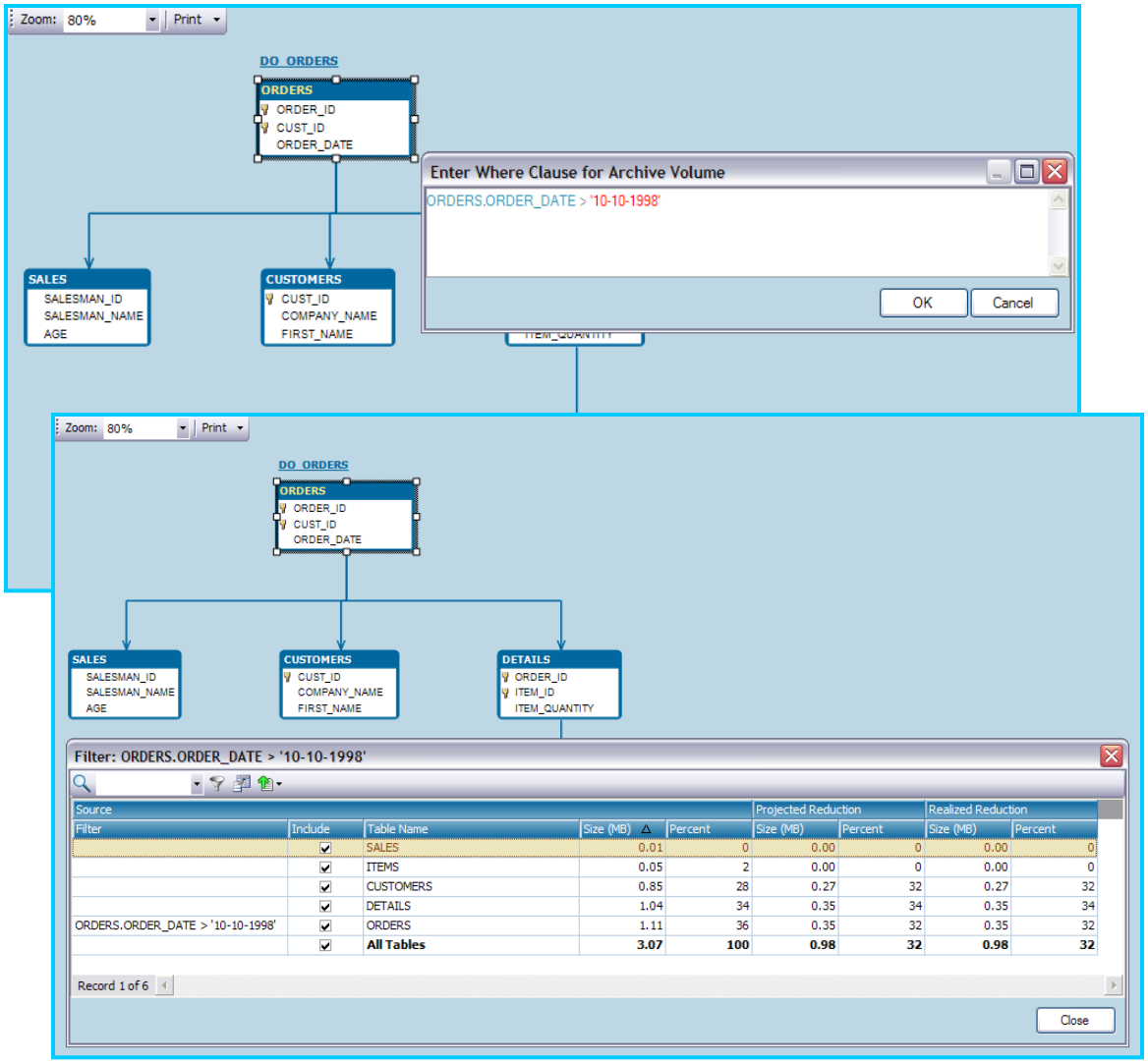


Discover Complete Business Object for Optim Projects: Archiving, Test Data, Application Retirement (Optim)

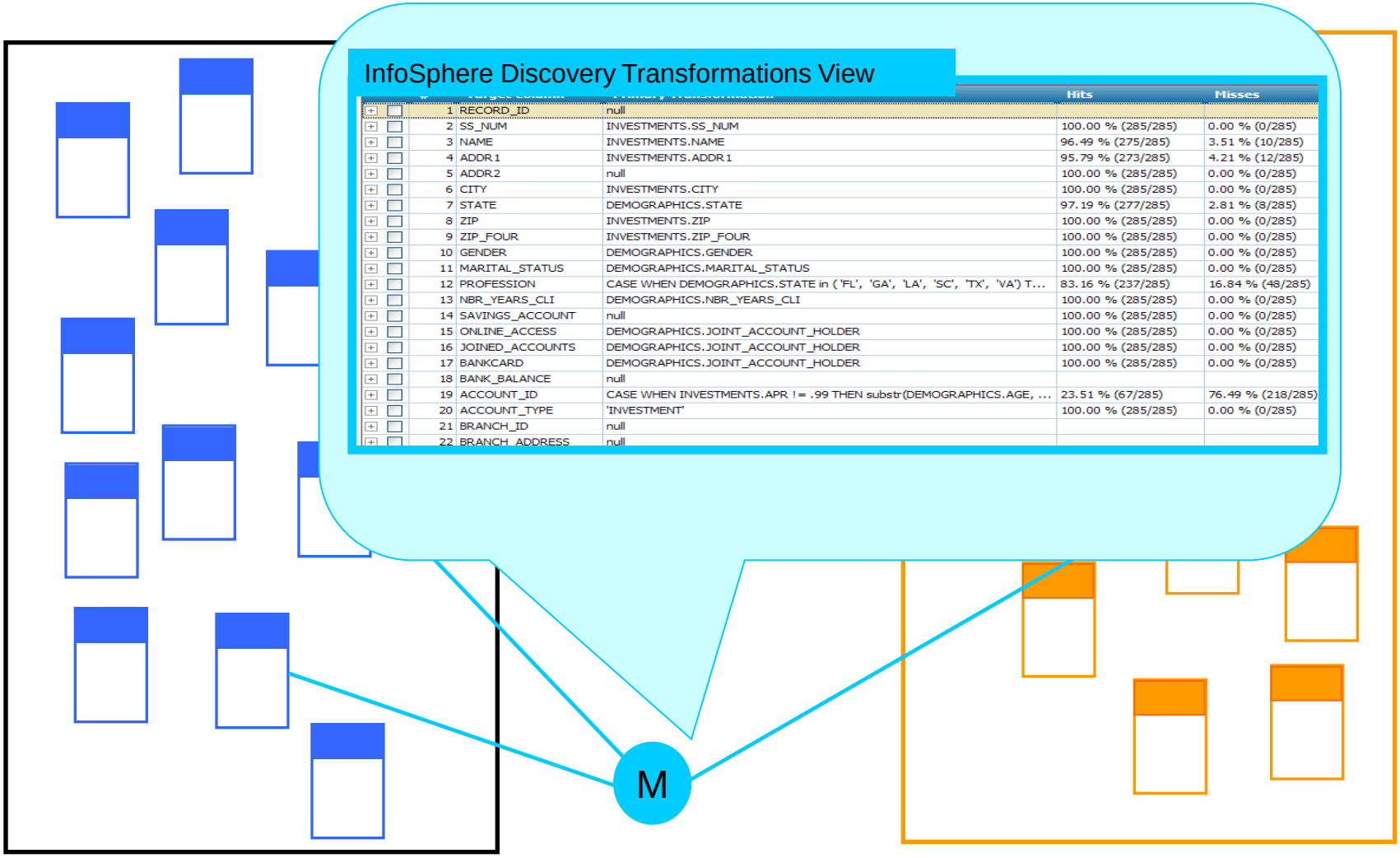


Business Object Volume Analysis

- Prototyping archive strategy
  - Define a selection condition
  - Identify tables that will be kept on source or archived away.
  - How much volume do I archive away?



The Crown Jewels: Transformation Discovery



			Hits	Misses
1	RECORD_ID	null		
2	SS_NUM	INVESTMENTS.SS_NUM	100.00 % (285/285)	0.00 % (0/285)
3	NAME	INVESTMENTS.NAME	96.49 % (275/285)	3.51 % (10/285)
4	ADDR1	INVESTMENTS.ADDR1	95.79 % (273/285)	4.21 % (12/285)
5	ADDR2	null	100.00 % (285/285)	0.00 % (0/285)
6	CITY	INVESTMENTS.CITY	100.00 % (285/285)	0.00 % (0/285)
7	STATE	DEMOGRAPHICS.STATE	97.19 % (277/285)	2.81 % (8/285)
8	ZIP	INVESTMENTS.ZIP	100.00 % (285/285)	0.00 % (0/285)
9	ZIP_FOUR	INVESTMENTS.ZIP_FOUR	100.00 % (285/285)	0.00 % (0/285)
10	GENDER	DEMOGRAPHICS.GENDER	100.00 % (285/285)	0.00 % (0/285)
11	MARITAL_STATUS	DEMOGRAPHICS.MARITAL_STATUS	100.00 % (285/285)	0.00 % (0/285)
12	PROFESSION	CASE WHEN DEMOGRAPHICS.STATE in ('FL', 'GA', 'LA', 'SC', 'TX', 'VA') T...	83.16 % (237/285)	16.84 % (48/285)
13	NBR_YEARS_CLI	DEMOGRAPHICS.NBR_YEARS_CLI	100.00 % (285/285)	0.00 % (0/285)
14	SAVINGS_ACCOUNT	null	100.00 % (285/285)	0.00 % (0/285)
15	ONLINE_ACCESS	DEMOGRAPHICS.JOINT_ACCOUNT_HOLDER	100.00 % (285/285)	0.00 % (0/285)
16	JOINED_ACCOUNTS	DEMOGRAPHICS.JOINT_ACCOUNT_HOLDER	100.00 % (285/285)	0.00 % (0/285)
17	BANKCARD	DEMOGRAPHICS.JOINT_ACCOUNT_HOLDER	100.00 % (285/285)	0.00 % (0/285)
18	BANK_BALANCE	null		
19	ACCOUNT_ID	CASE WHEN INVESTMENTS.APR != .99 THEN substr(DEMOGRAPHICS.AGE, ...	23.51 % (67/285)	76.49 % (218/285)
20	ACCOUNT_TYPE	'INVESTMENT'	100.00 % (285/285)	0.00 % (0/285)
21	BRANCH_ID	null		
22	BRANCH_ADDRESS	null		

Transformation Discovery By Comparing Data: Aligns Rows Based on Correlations

Step 1: Discovery Engine analyzes the **data values** to **automatically discover** the columns that align rows across data sources:

Member (Table 1) = ID (Table 25)

Table 1

Row	Member	SS #	Age	Phone	Sex
1	595846226	123-45-6789	15	(123) 456-7890	M
2	567472596	138-27-1604	8	(138) 271-6037	F
3	540450091	154-86-4196	22	(154) 864-1961	M
4	514714372	173-44-7900	55	(173) 447-8996	F
5	490204164	194-26-1648	4	(194) 261-6476	F
6	466861109	217-57-3046	66	(217) 573-0453	M
•	•	•	•	•	•
•	•	•	•	•	•
987,623	444629628	243-68-1812	25	(243) 681-8107	F
987,624	423456789	272-92-3629	87	(272) 923-6280	M

Table 25

ID	Demo1
514714372	3
444629628	3
540450091	2
567472596	1
423456789	2
490204164	1
•	•
•	•
595846226	0
466861109	0

Transformation Discovery By Comparing Data : Column Mapping

Step 2: With rows now aligned, analyzes the **data values** to **automatically discover** complex business rules and transformations:

Case: 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
= Demo1

Table 1

Row	Member	SS #	Age	Phone	Sex
1	595846226	123-45-6789	15	(123) 456-7890	M
2	567472596	138-27-1604	8	(138) 271-6037	F
3	540450091	154-86-4196	22	(154) 864-1961	M
4	514714372	173-44-7900	55	(173) 447-8996	F
5	490204164	194-26-1648	4	(194) 261-6476	F
6	466861109	217-57-3046	66	(217) 573-0453	M
•	•	•	•	•	•
•	•	•	•	•	•
987,623	444629628	243-68-1812	25	(243) 681-8107	F
987,624	423456789	272-92-3629	87	(272) 923-6280	M

Table 25

ID	Demo1
595846226	0
567472596	1
540450091	2
514714372	3
490204164	1
466861109	0
•	•
•	•
444629628	3
423456789	2

Transformation Discovery By Comparing Data : Measuring Mapping Rule Strength

Step 3: With business rules now discovered, analyzes the **data values** to automate discovery of **unknown data inconsistencies**

- Moves QA early in the process

Case: 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
= Demo1

Table 1

Row	Member	SS #	Age	Phone	Sex
1	595846226	123-45-6789	15	(123) 456-7890	M
2	567472596	138-27-1604	8	(138) 271-6037	F
3	540450091	154-86-4196	22	(154) 864-1961	M
4	514714372	173-44-7900	55	(173) 447-8996	F
5	490204164	194-26-1648	4	(194) 261-6476	F
6	466861109	217-57-3046	66	(217) 573-0453	M
•	•	•	•	•	•
•	•	•	•	•	•
987,623	444629628	243-68-1812	25	(243) 681-8107	F
987,624	423456789	272-92-3629	87	(272) 923-6280	M

Table 25

ID	Demo1
595846226	0
567472596	1
540450091	2
514714372	3
490204164	1
466861109	0
•	•
•	•
444629628	3
423456789	2

Hit Rate: 98%

Transformation Discovery By Comparing Data: Discoverable Transformations

- **Scalar**
- One to one
- Substring
- Concatenation
- Constants
- Tokens

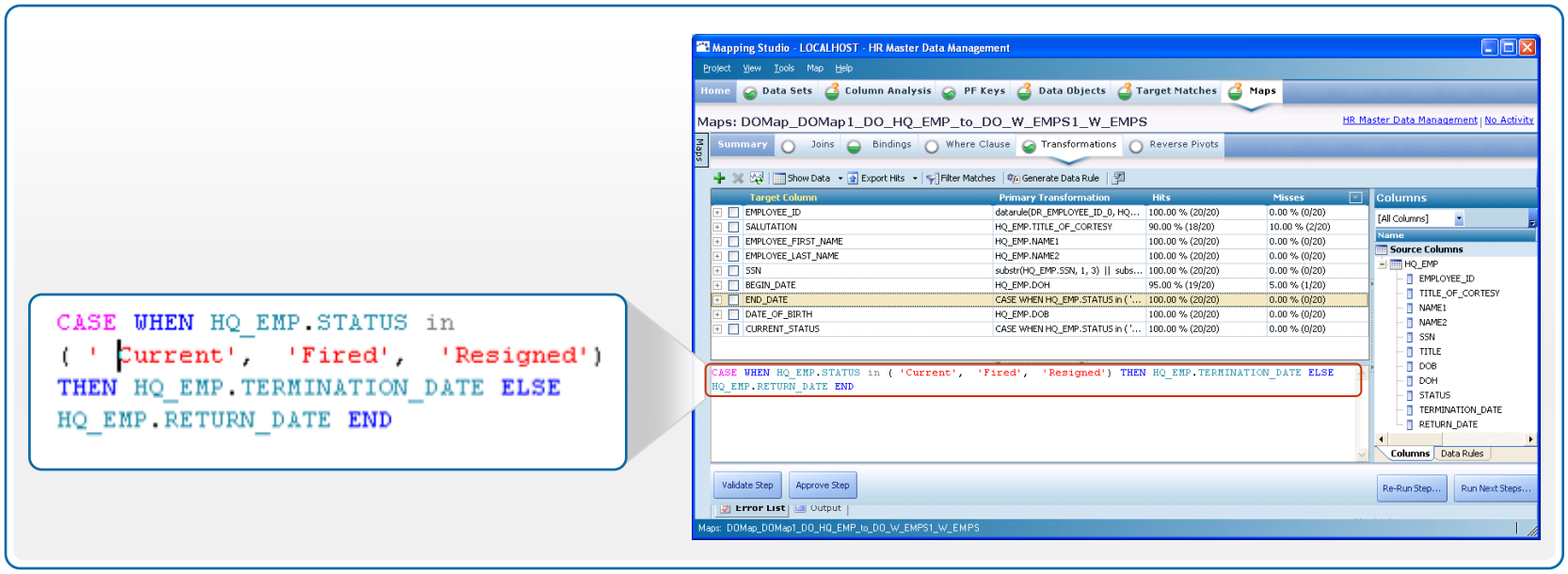
- **Conditional Logic**
- Case Statements
- Equality/Inequality
- Null Conditions
- In/Not In
- Conjunctions

- **Joins**
- Inner
- Left Outer

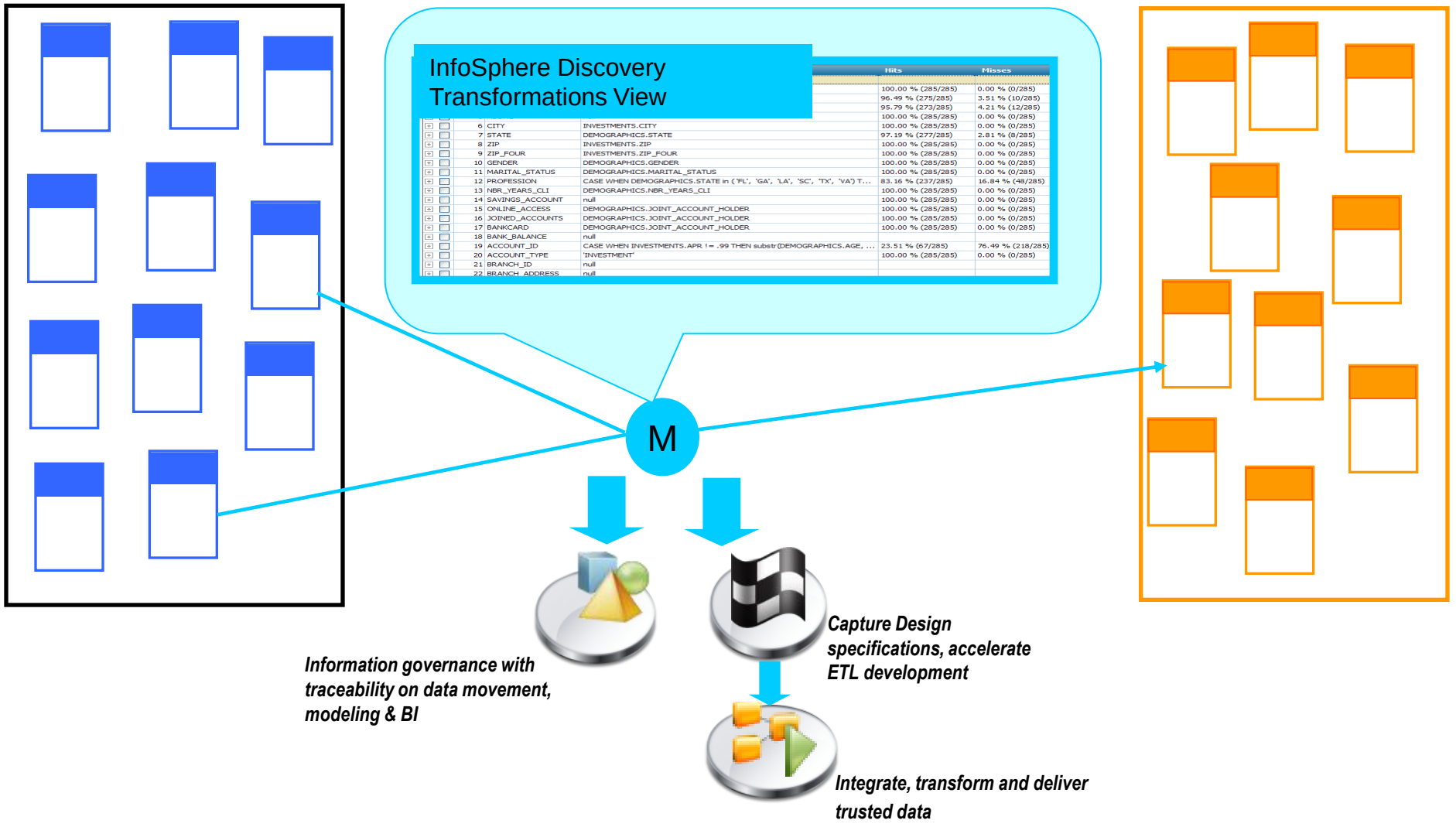
- **Aggregation**
- Sum
- Average
- Minimum
- Maximum

- **Column Arithmetic**
- Add
- Subtract
- Multiply
- Divide

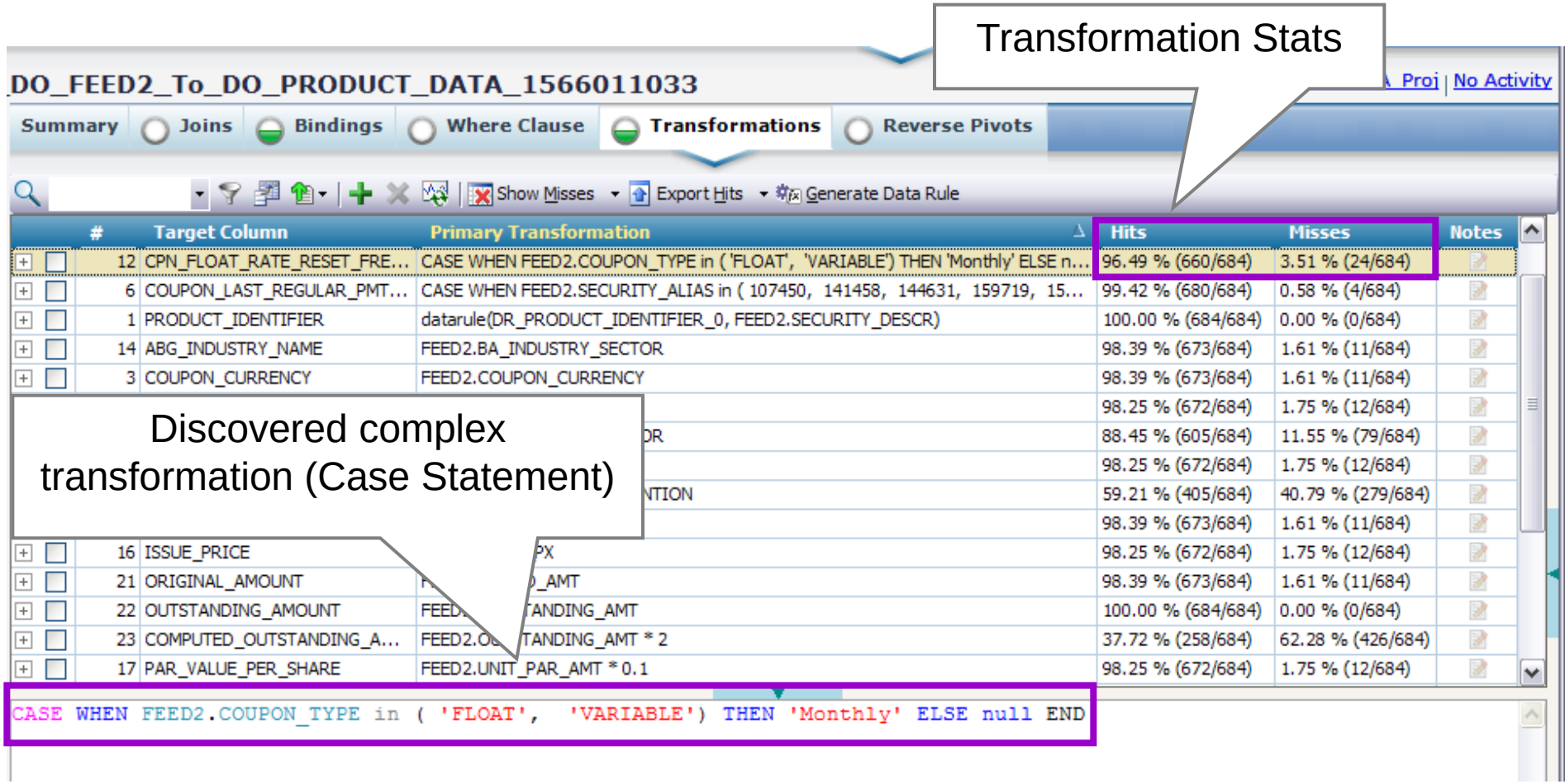
- **Reverse Pivot**
- **Cross-Reference**
- **Custom Data Rules**



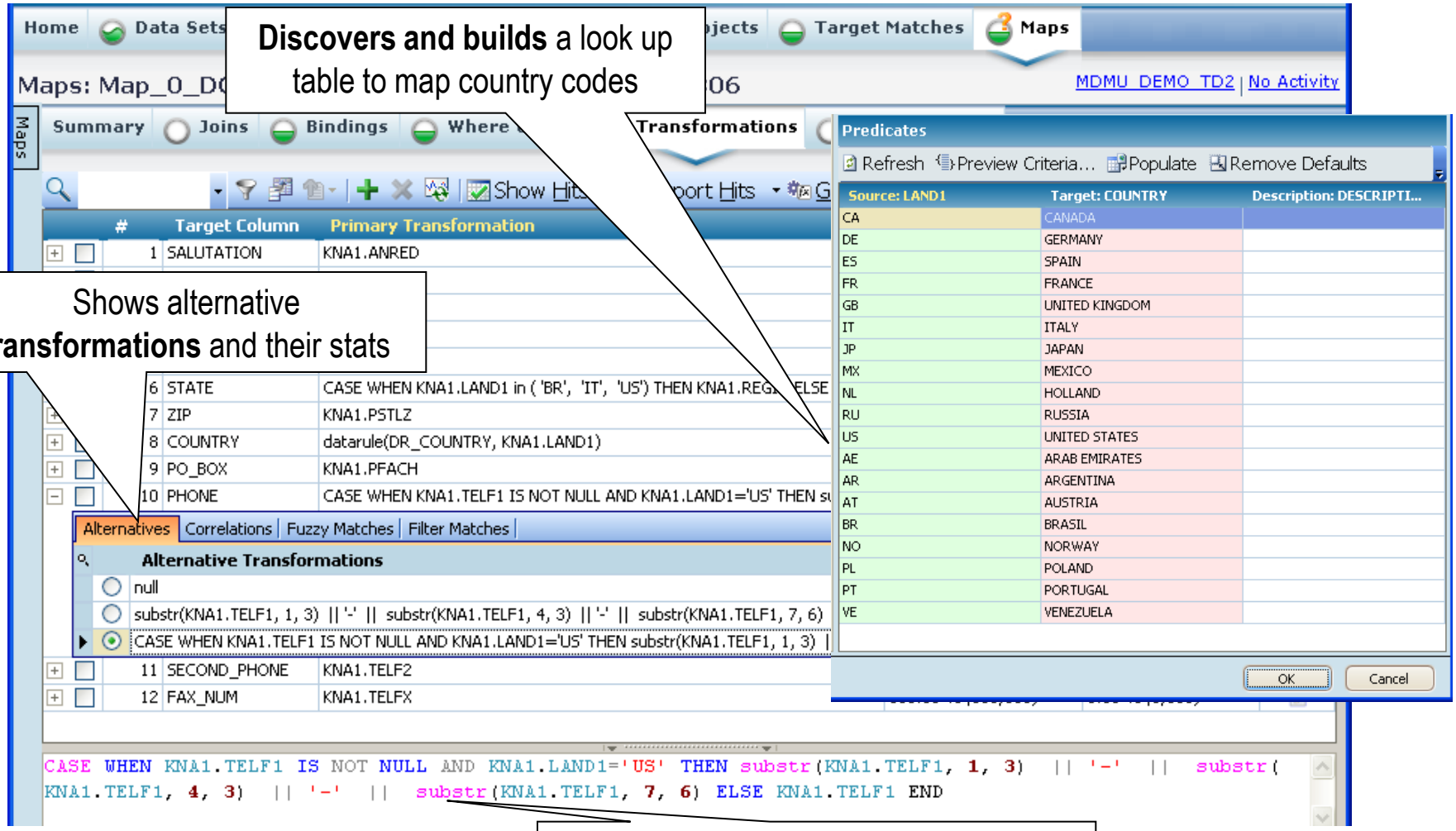
Discovery for ETL and Lineage Metadata Initiatives



Transformation Discovery Example



Transformation Discovery Example



Pattern Based Sensitive Data Discovery Example: SSN

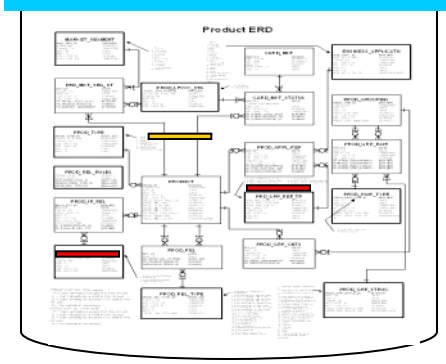
InfoSphere Discovery Classified Columns View

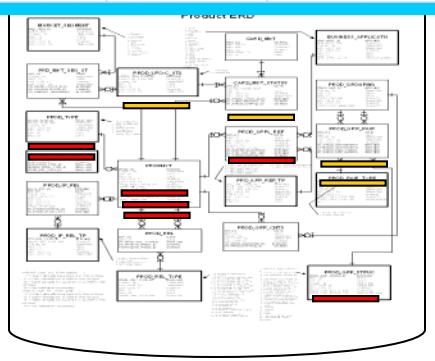
MC	MasterCard	Yes	Yes	No
NEA	Notes Email Address	No	Yes	No
NIF	Spanish NIF	No	Yes	No
NINO	UK NINO	No	Yes	No
PAN	Passport Number	No	Yes	No
PN	Personal Name	No	Yes	No
SIN	Canadian SIN	No	Yes	No
SSN	US Social Security Number	Yes	Yes	No

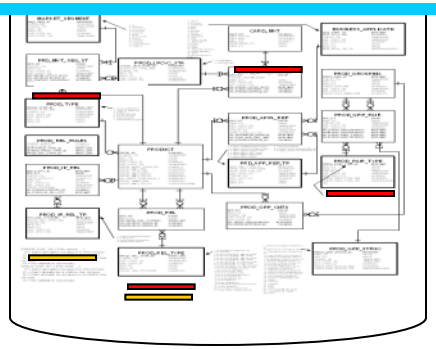
Assigned | Excluded

Column Metadata			Column Classification			Statistics		
Name	Table	Data Set	Hit Rate	Ap...	Method	Cardinality	Selectivity	Null Count
COMMENTS	ALL_POSITIONS_F	Data Set 1	50%	☐	Discovered	4	0.00	3641
MRC_ACCTD_AMOUNT	MK_ADJTMNT	Data Set 1	35.2941%	☐	Discovered	12	0.04	257
MRC_EXCHANGE_RATE	MK_KE_TRX_70	Data Set 1	52.5%	☐	Discovered	45	0.00	9397
MRC_EXCHANGE_RATE	MK_SCHED	Data Set 1	33.2075%	☐	Discovered	47	0.00	9590
MRC_ACCTD_EARNED_DISC_TAKEN	MK_SHOU_APPS_ALL	Data Set 1	100%	☐	Discovered	1	0.00	200
MRC_ACCTD_UNEARNED_DISC_T...	MK_SHOU_APPS_ALL	Data Set 1	100%	☐	Discovered	1	0.00	200
ID_PIN	NK_JE_HDRS	Data Set 1	100%	☐	Discovered	23	0.00	0
BILLING_ID	PRFL_3_CUST	Data Set 1	39.777%	☐	Discovered	269	1.00	0

URL	URL	No	Yes	No
USPHN	US Phone Number	Yes	Yes	No
USSC	US State Code	No	Yes	No







Beyond “Scanning for Pattern”: additional sensitive data discovery

- Discovery finds other types of sensitive data
  - Sensitive data that do not have distinctive pattern
  - Sensitive data that are partial or hidden (malicious)
  - Sensitive data embedded in free text

- How do we find them?
  - data similarity – exact value matching
  - data similarity -- fuzzy value matching
  - metadata similarity including known classification
  - data relationships

Scan for pattern: custom “sensitive”, custom “algorithm”

- Define “custom sensitive” data in Discovery

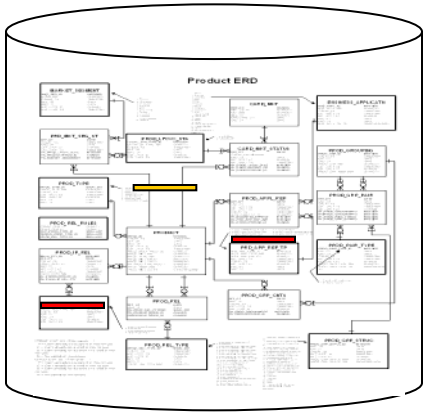
- How do we find custom sensitive data?
  - User defined algorithm -- Once deployed, behave the same as “built-in”
    - Algorithm is run as part of profiling
    - Hit/miss metrics and data view will be available on custom sensitive.
  - data similarity – exact value matching
  - data similarity -- fuzzy value matching
  - metadata similarity including known classification
  - data relationships

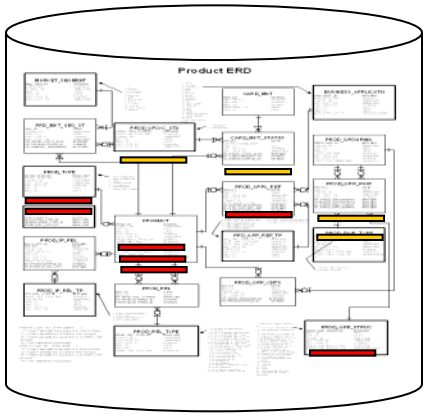
From "Sensitive Data" Discovery" to "Critical Data" Discovery: Using Discovery to connect Business to IT

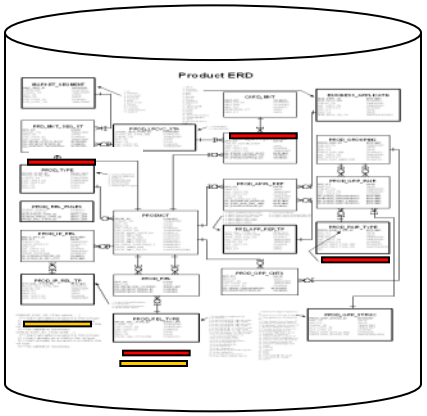
- Discover "critical data elements"(CDEs) and map them to Business Glossary
  - Data driven term mapping

- Once we find these CDEs, explore its surroundings to identify other data relevant to the business of interest.

Sensitive/Critical Data Discovery For Compliance and Test Data Management Projects







Term mapping

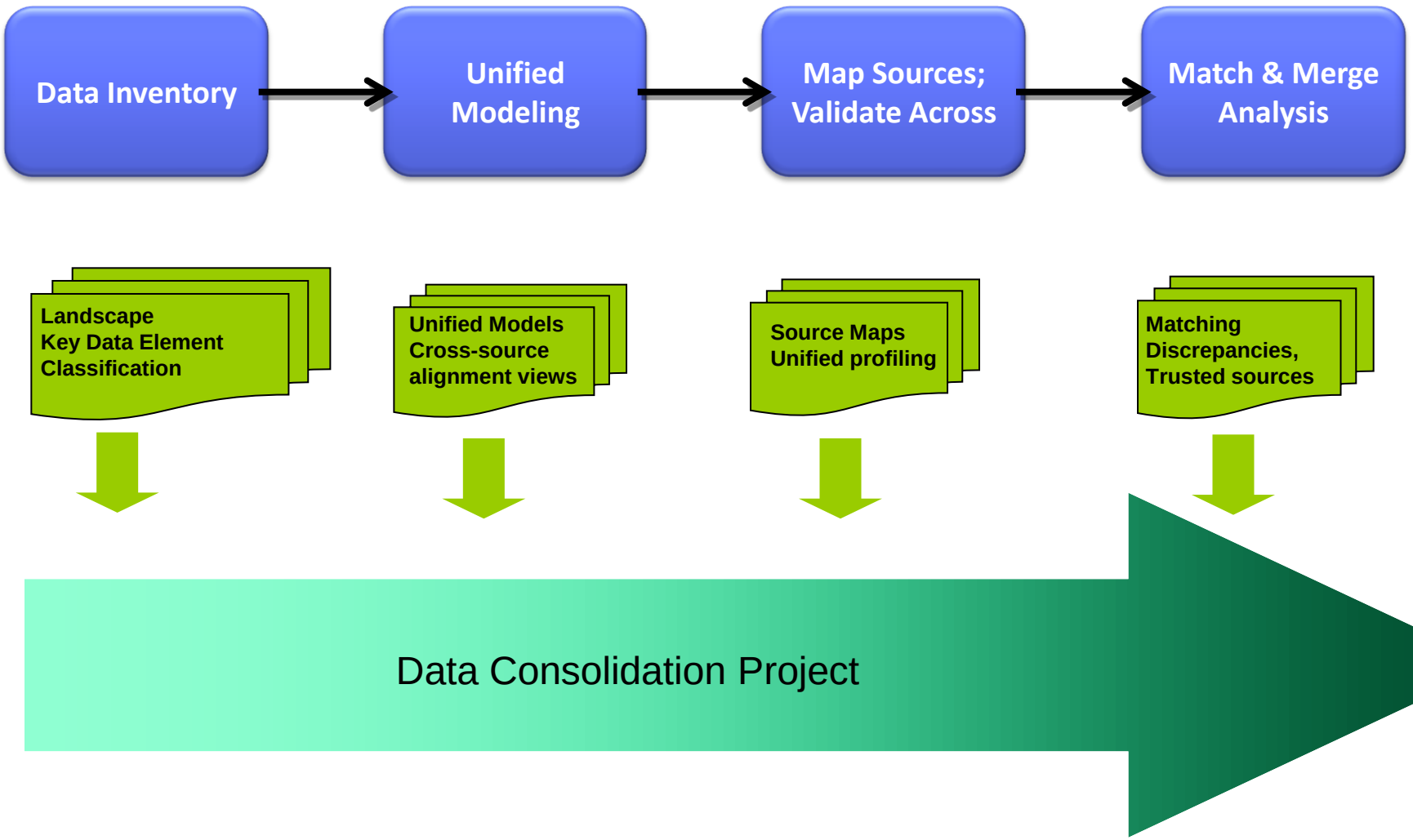
Sensitive data domains with default masking

Business Glossary

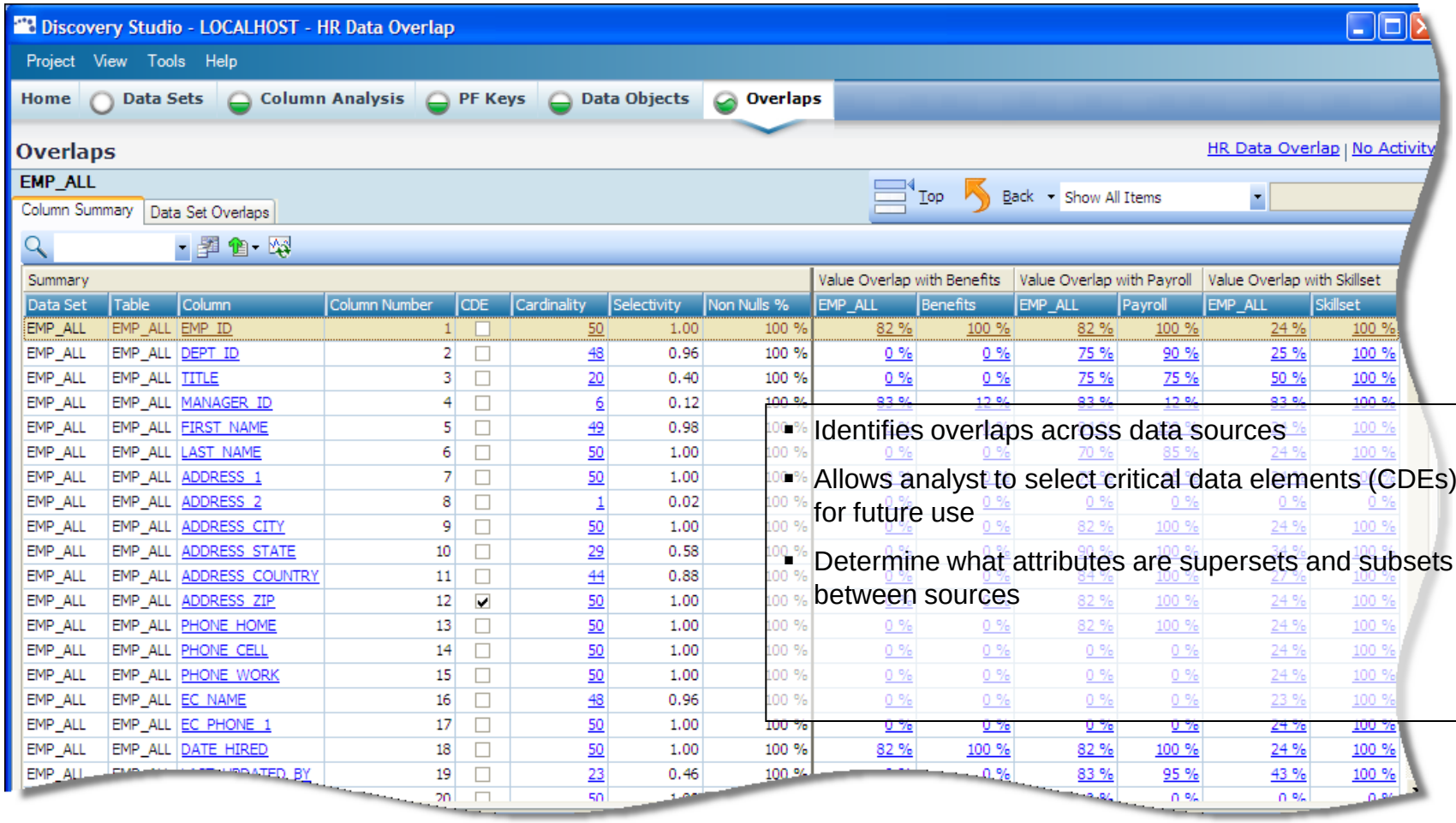
Connect business to IT

Masking privacy data in test data extract

Complete methodology for Consolidation Prototyping



Data Inventory: Cross System Overlap



- Identifies overlaps across data sources
- Allows analyst to select critical data elements (CDEs) for future use
- Determine what attributes are supersets and subsets between sources

Data Inventory: Critical Data Element Alignment Across sources

Target Table Schema | Source Mapping | Unified Column Analysis | Match and Merge Analysis

Total Source Row Count: 189

Preview Data | Value Frequency | All Mappings | Export Table

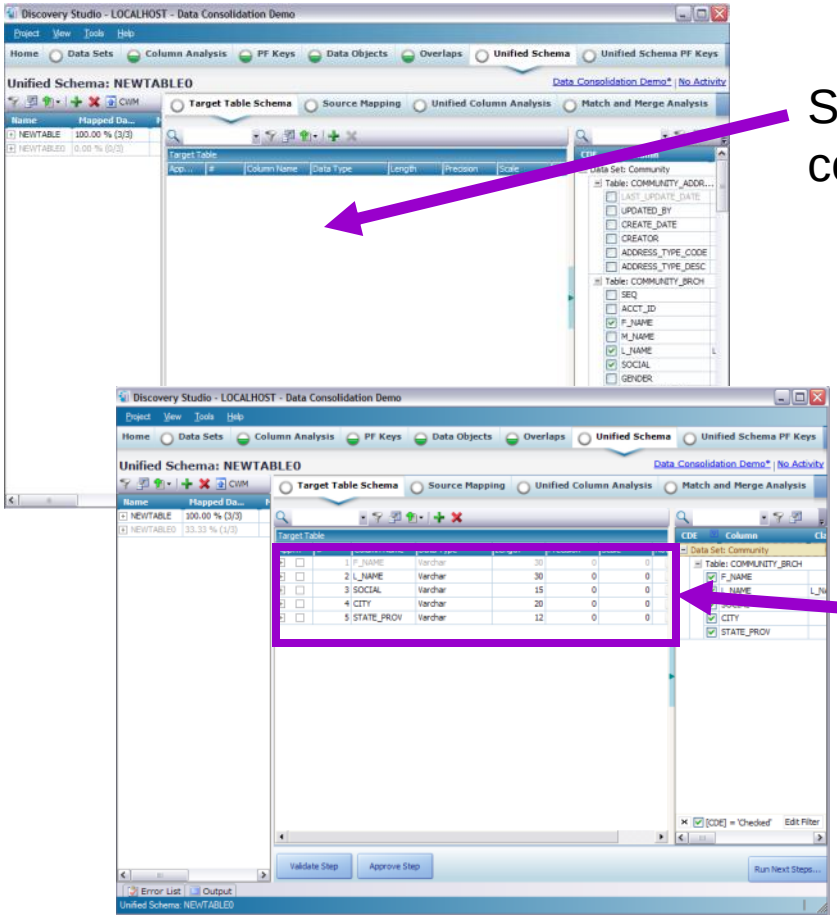
Target Table			Statistics				
#	Name	Data Type	Data Sets	Cardinality	Selectivity	Min	Max
1	F_NAME	varchar(30)	100.00 % (3/3)	108	0.5714285...	ABBY	YVONNE
2	L_NAME	varchar(30)	100.00 % (3/3)	89	0.4708994...	AARON	WELCH
3	SOCIAL	varchar(15)	100.00 % (3/3)	145	0.7671957...	103-58-1068	396694138

Mapping		Statistics				
Data Set	Transformation	Cardinality	Selectivity	Min	Max	Complete...
Target Table		145	0.76719576...	103-58-1068	396694138	
Community	COMMUNITY_BRCH.SOCIAL	52	1	103-58-1068	387-29-3234	100.00 % (5...
Region	REGION_BRCH.SSN	51	1	106619513	396694138	100.00 % (5...
CRM	CRM_BRCH_1A.TAX_ID	86	1	103-58-1068	396-69-4138	100.00 % (8...

4	CITY	varchar(20)	100.00 % (3/3)	96	0.5079365...	ADONA	YEADON
5	STATE_PROV	varchar(12)	100.00 % (3/3)	42	0.2222222...	AL	WV

- Provides profiling data for the union of all of the sources mapped to the virtual target
- Shows profiling results for individual source attribute and for the union of all sources

Unified Modeling: Bottoms Up Modeling of Target Schema

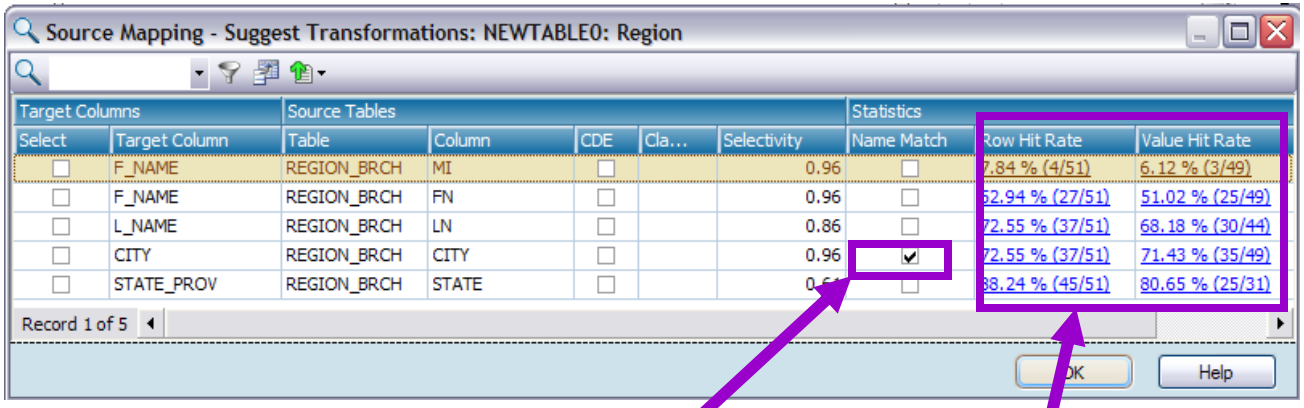


- Import industry standard model for each target table

  OR

- Model target table “bottoms up” by dragging and dropping columns from sources
- Using source attributes populates the virtual target with data

Sources to Target Mapping Discovery: using both data and metadata



Name matches are also shown

Overlap percentages between a data source to the virtual target

- Map sources to the target schema based on name and value matches (from the overlap analysis)
- Each source mapped adds more data to the “virtual target” and enables more value matching from the overlap analysis

Match and Merge Analysis

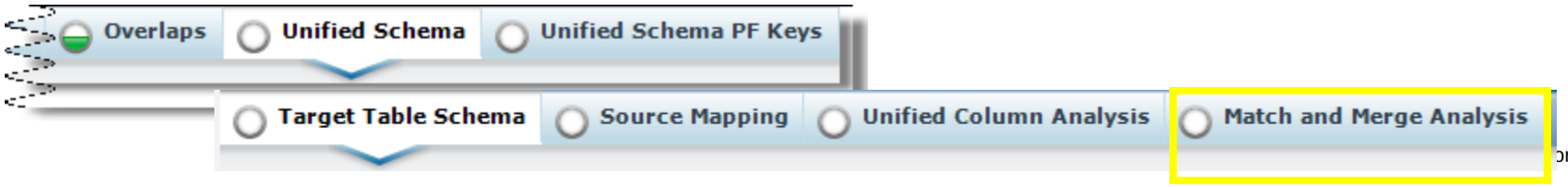
Data Set	F_NAME	L_NAME	SOCIAL	CITY	STATE_PROV
<Merged>	SERGIO	ELLIOTT	230-24-7823	PARKERS PRAIRIE	GA
COMMUNITY	SERGIO	ELLIOTT	230-24-7823	PARKERS PRAIRIE	GA
COMMUNITY	ELEANOR	ELLIOTT	295-78-2505	LOOMIS	PA
COMMUNITY	MARSHAL	ELLIOTT	272-72-7078	NORTH OAKS	CA
COMMUNITY	ROGER	ELLIOTT	218-28-1759	ORCHARD HEIGHTS	NV
CRM	SERGIO	ELLIOTT	230-24-7823	PARKERS PRAIRIE	GA
CRM	ELEANOR	ELLIOTT	295-78-2505	LOOMIS	PA
CRM	MARSHAL	ELLIOTT	272-72-7078	NORTH OAKS	CA
CRM	ROGER	ELLIOTT	218-28-1759	ORCHARD HEIGHTS	NV
REGION	SERGIO	ELLIOTT	230-24-7823	PARKERS PRAIRIE	GA
REGION	ELEANOR	ELLIOTT	295-78-2505	LOOMIS	PA
REGION	MARSHAL	ELLIOTT	272-72-7078	NORTH OAKS	CA
REGION	ROGER	ELLIOTT	218-28-1759	ORCHARD HEIGHTS	NV

- Match and Merge analysis allows user to quickly prototype using different columns as the matching key across sources

Example of overmatching when using a match key of L_Name (conflicts in red)

Data Set	F_NAME	L_NAME	SOCIAL	CITY	STATE_PROV
<Merged>	SERGIO	ELLIOTT	230-24-7823	PARKERS PRAIRIE	GA
COMMUNITY	SERGIO	ELLIOTT	230-24-7823	PARKERS PRAIRIE	GA
CRM	SERGIO	ELLIOTT	230-24-7823	PARKERS PRAIRIE	GA
REGION	SERGIO	ELLIOTT	230-24-7823	PARKERS PRAIRIE	GA

Example of matching when match key being prototyped is Social Security Number



Customer Case Studies

Case Study: International Truck Manufacturer

Business Challenge:

- Improve customer experience
- Reduce number of financial systems
- Address data quality once for multiple large migration projects

Approach:

- Consolidate various systems into a single master

Technical Issues:

- 14 customer, 9 vendor and 6 materials data sources
- Manual data analysis put project at risk
  - Required 6 months to cross-analyze just 2 systems
  - Assumptions were often incorrect and documentation was spotty
  - Each analyzed attribute involved 3-7 business and data analysts
  - Manual combination of cross system queries was error prone
  - QA process for cross-system analysis was almost non-existent

Case Study Results: Using InfoSphere Discovery for Cross-System Analysis

- 12 months of work completed in only 1 month
  - ~$180K savings in first month alone using Exeros
  - Increased project predictability and lowered risk
- Assumptions about business rules and relationships are now tested and validated against millions of rows before decisions are made
- One person can answer questions in an hour that used to take three or more people several days to answer
- Less SME time is now required and less time is spent coordinating discussions between different SMEs

KeyBank

"After the product was purchased, the Key Bank team targeted applications requiring source-to-target mapping from the source systems to the enterprise data warehouse. **The previous approach**, requiring crafted queries and manual effort, **was expected to take nine staff months**. With the cross-system data analysis workbench, **Exeros (now InfoSphere) Discovery, the team was able to do the mapping automatically in two weeks –** a significant reduction in cost and effort but, more importantly, a reduction in time to business value."

Source: Data Governance From Policy to Practice
BeyeNETWORK Research Report Case Study**: Key Bank**
David Loshin and BeyeNetwork 2008

Case Study:
Data Lineage and Traceability

Chief Data Architect

Regional Commercial Bank

> “A lineage mapping project as part of our MDM program was **estimated to cost $120K** and take three months, **was completed for only $10K** in one week because of Discovery software.”

Case Study: IT Asset Master

- **Problem:**
- Couldn't answer simple questions:
  - Who owns an IT asset?
  - How much does it cost us to operate our current Infrastructure?
  - How much are we underutilizing our assets?

- **Project:**
- Consolidate 24 asset management systems into a single asset master

- **Manual results:**
- 7,520 person hours at $110/hour (~$825K) spent manually mapping
- Volume of data and complexity was too great
- Still couldn't answer the questions

- **InfoSphere Discovery results**:
- 2 weeks of elapsed time

Vice President

Charlotte, NC based Commercial bank

"One person using Discovery software, who knew nothing about our environment accomplished more in a two week pilot than 9 subject matter experts were able to complete over 9 months."

IBM InfoSphere™ Discovery – Invaluable Data Analysis Tool

- Accelerate deployment of your information agenda projects:
  - –Improves accuracy, predictability and repeatability
  - –Speeds project data analysis by as much as 10 times
  - –Minimize SME Time