Structured Data Security Methodology

Discovering Sensitive Data in Structured Data Sources

ISACA®
Serving IT Governance Professionals
San Francisco Chapter

EXEROS®

Agenda

Agenda

- Sensitive Data Security
  - Introduction
  - Find before you Fix
- Current Approaches
- Framework and Methodology
- Q & A

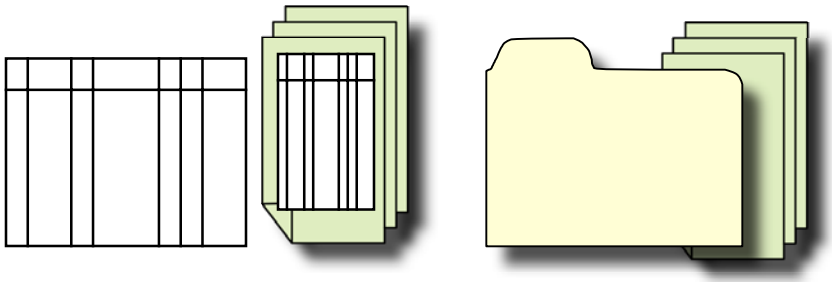
Sensitive Data Security

Uptrend in Data Breaches

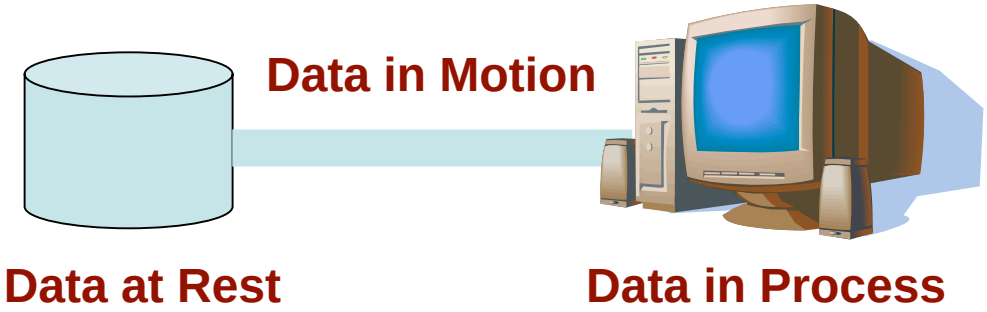
- Privacy Rights Clearinghouse – /www.privacyrights.org/ar/chrondatabreaches.htm
  - **2005: 13 / month → 2006: 62 / month -→ 2007: 74 / month**
  - Between January, 2005 – August, 2007 An estimated **165,891,898** records compromised!
  - Equal Opportunity Problem
  - Most industries affected – Financial Services, Payment Card Issuers, Education, Retail, Healthcare, Government…

Categories of Sensitive Data

- Based on Form
  - Structured
  - Unstructured



- Based on Location
  - Data in Motion
  - Data in Process
  - Data at Rest

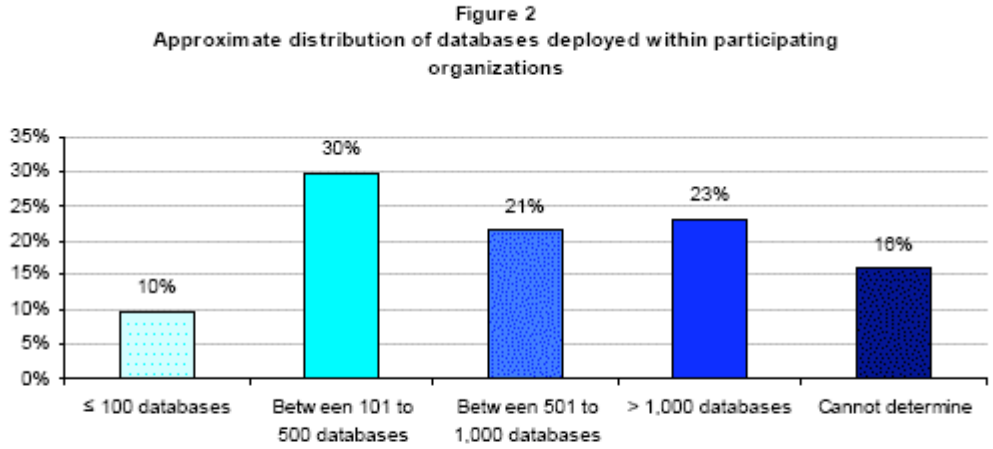


Introduction – Database Proliferation

Ponemon Institute LLC

June, 2007 Survey

Figure 2: Database deployment

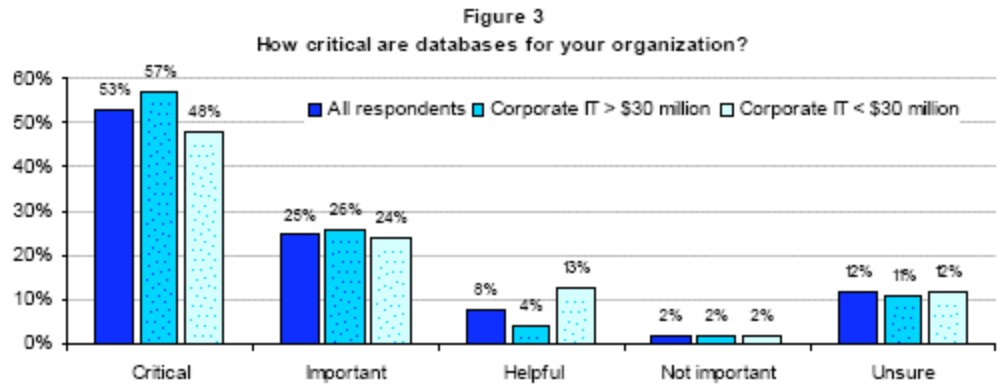


Ninety percent of surveyed organizations reported deployment of more than 100 databases within their organization. Twenty-three percent of organizations reported more than 1,000 databases.

Introduction – Database Importance

June, 2007 Survey

Figure 3: Database importance

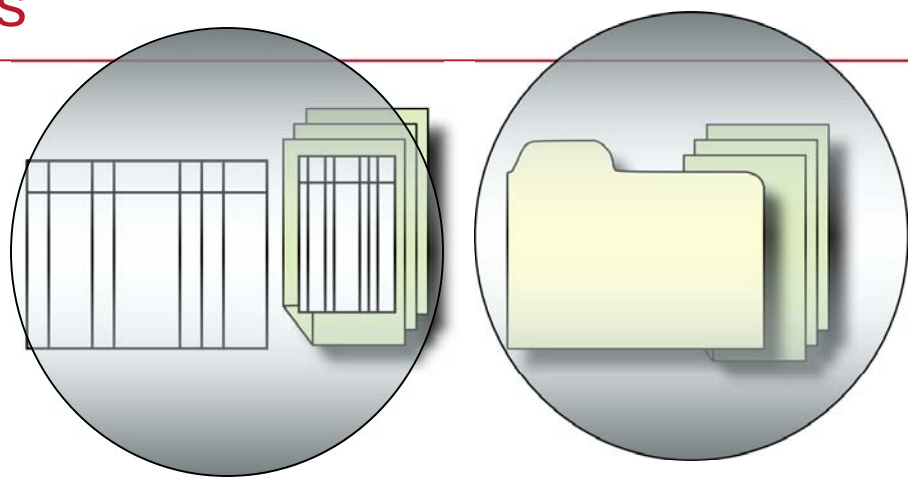


Organizations value their databases enormously. Fifty-three percent ranked database importance as critical, while another twenty-five percent responded with a rating of important. Perhaps more telling is the fact that less than two percent state that databases are not important to their business.

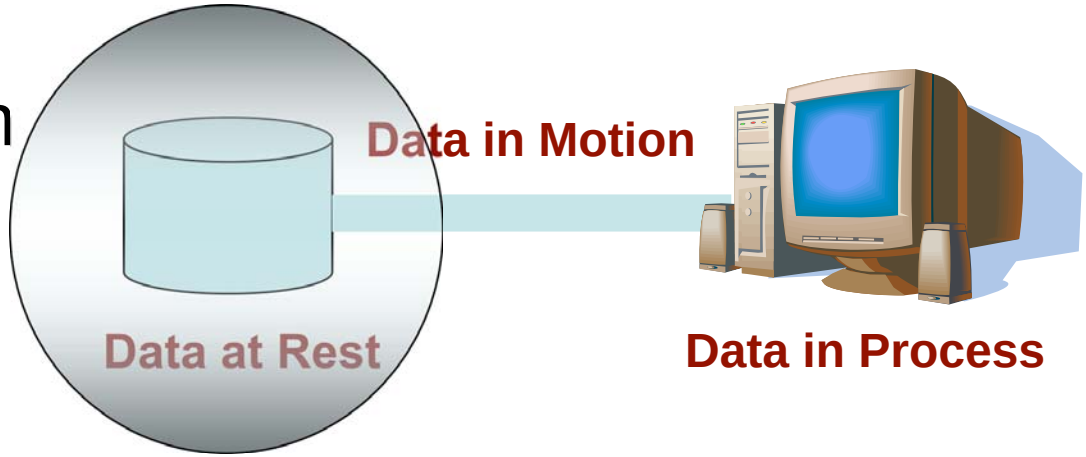
Categories of Sensitive Data

Protect Early to Prevent Loss

- Based on Form
  - Structured
  - Unstructured



- Based on Location
  - Data in Motion
  - Data in Process
  - Data at Rest



Introduction

- What is “Sensitive Data”?
  - Non-personal Public Information (NPI)
    - Account Number, PIN, Mother’s Maiden Name…
  - Private Customer Data
    - Customer ID, Security Question, Security Question Answer…
  - Private Employee Data
    - Employee ID, Name, DOB, Salary, Benefits Info…
  - Patient / Patient Care Data
    - Patient NPI, Disease Codes, Treatment Codes…

Introduction

- Why is Data “Sensitive” ?
  - Breach can compromise an individual’s identity
    - Card Holder
      - Possible Outcome - Credit Card Fraud
    - Patient
      - Possible Outcome - Denial of Insurance / Employment
    - Employee
      - Possible Outcome – Termination of Employment
  - Breach can compromise a Company’s Strategy
  - Violation of Privacy Laws
  - Lawsuits, Penalties, Brand Notoriety, Business shutdown

Business Drivers

- Business Drivers for Locating and Securing Sensitive Data:
  - Compliance Laws
    - Examples: GLBA, SOX
  - Privacy Laws
    - Examples: HIPAA, Company privacy laws
  - Privacy Standards
    - Examples: PCI DSS
  - Risk to Business due to Breach

Requirements

- Business
  - Establish Repeatable Audit Process
  - Validate/Confirm Compliance
  - Determine Non-Compliance
  - Report and Analyze Incidents
  - Remediate
  - De-Identify (Protect)

Requirements

- Technical
  - Repeatable
  - Measurable
  - Accurate
  - Scalable
  - Comprehensive Coverage

Find before you Fix
Sensitive Data Security – Step 1

- IDENTIFY Sensitive Data Elements (SDEs)
  - What ?
    - What is it?

Find before you Fix
Sensitive Data Security – Step 2

- LOCATE / DISCOVER Sensitive Data
  - What?
    - What are they called and what are their values?
  - Where?
    - Where are they located?
  - How?
    - How do they exist? (In what form do they exist?)

Find before you Fix
Sensitive Data Security – Step 3

- PROTECT Sensitive Data
  - Who can access ?
    - Remediation Policy
  - What can they access ?
    - Remediation Policy
  - When can they access ?
    - Remediation Policy
  - How (in what form) do they access ?
    - De-Identify

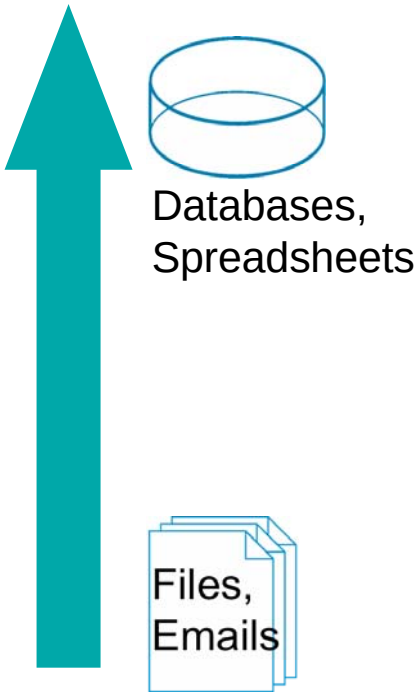
Find before you Fix
Sensitive Data Security starts with Discovery

- **Cannot Protect what you can't Find**

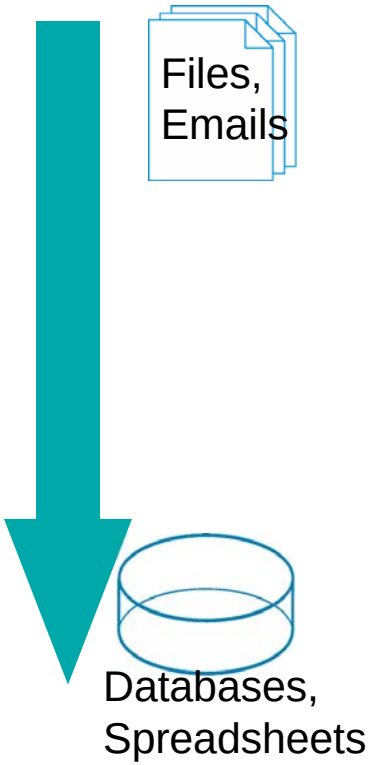
  - **Sensitive Data Discovery - the initial, critical step**

Market Has Focused on Unstructured Data

Criticality of Data



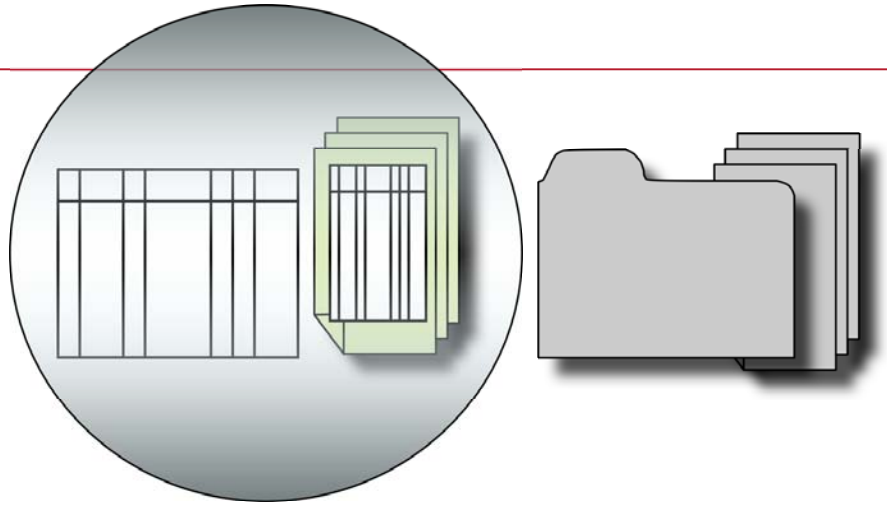
Availability of Sensitive Data Discovery Solutions



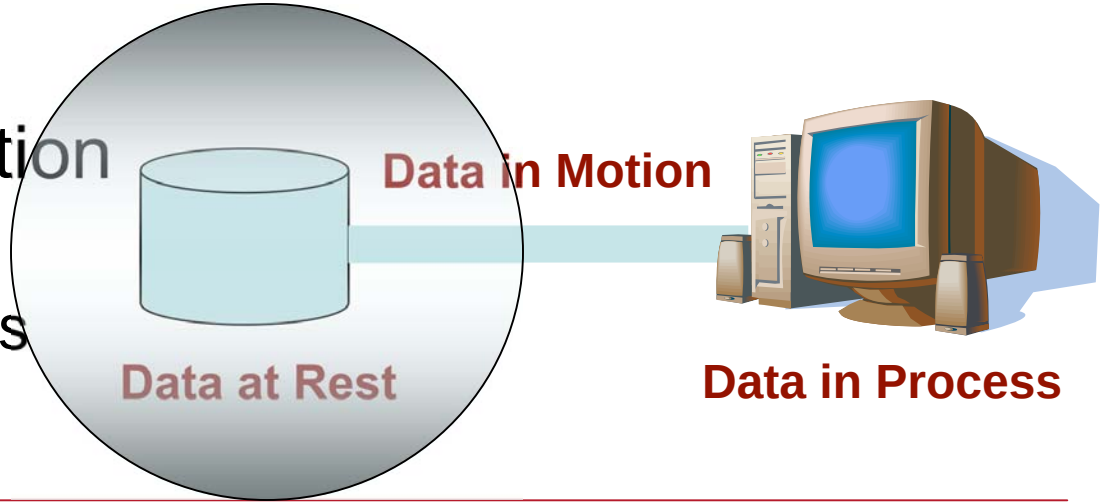
- Market has evolved in reverse of criticality
- Why?
  - Discovering sensitive data in databases is hard!

Categories of Sensitive Data – Structured Data at Rest

- Based on Form
  - Structured
  - Unstructured



- Based on Location
  - Data in Motion
  - Data in Process
  - Data at Rest



Current Approaches for Discovering Structured Sensitive Data

Current Approaches

- Ask the Expert
- Metadata Analysis
- Data Profiling
- Search / Scan

- Very little to no automation is used

Limitations of Current Approaches

- Ask the Expert
  - Need an expert per application
  - Expert must know application data inside-out
  - Expert may leave company
  - Inconsistency - Different Experts may come back with different results

Limitations of Current Approaches

- Metadata Analysis
  - Metadata misleading and incomplete
  - SSN column may be empty!
  - NOTES column may have SSNs!

Limitations of Current Approaches

- Data Profiling
  - Good starting point
  - Manual effort required to complete analysis

Limitations of Current Approaches

- Search / Scan
  - Cannot find hidden sensitive data
  - Need to discover relationships to make result meaningful
  - Search for last 4 digits of SSN will match any 4-digit number columns.

Limitations of Current Approaches

- Need a comprehensive, holistic approach
  - To catch-up with pending needs
    - (PCI-compliance, Sox-compliance etc.)
  - To address current needs
    - Periodic Audits and Compliance Reporting
  - To pro-actively address future needs
    - New Regulations/Laws/Standards
    - New Sensitive Data Elements
    - Changes to existing sensitive data elements
    - Changes to applications/application logic

A Framework & Methodology for Sensitive Data Discovery in Structured Data

Framework Goals

- Manageable
- Easy to Implement
- Repeatable
- Accurate Results in Reasonable Time

- Comprehensive Methodology to
  - Validate Known Sensitive Data
  - Discover Unknown Sensitive Data
  - Assess the impact of New Sensitive Data Elements / Changes to existing ones

Methodology Steps - Overview

- Identify Sensitive Data Elements
  - PCI List for example
- Collect sample set of values
  - Valid values eliminate false positives
- Classify Sensitive Data Elements
- Discover Obvious Sensitive Data (iterate)
- Discover Hidden Sensitive Data (iterate)
- Enable Sensitive Data Flow or Lineage Analysis
- Enable Remediation

Methodology Steps - Overview

- Identify Sensitive Data Elements
  - PCI List for example
- Collect sample set of values
  - Valid values eliminate false positives
- Classify Sensitive Data Elements
- Discover Obvious Sensitive Data (iterate)
- Discover Hidden Sensitive Data (iterate)
- Enable Sensitive Data Flow or Lineage Analysis
- Enable Remediation

Forms of Structured Sensitive Data

- Based on Context
  - Some data are “born” sensitive while others are “made” sensitive
  - Independent Sensitive Data
  - Dependent Sensitive Data

- Based on Form of Existence
  - Obviously Sensitive Data
  - Hidden Sensitive Data

Independent Sensitive Data

- Inherently Sensitive
- Do not need separate context to make them sensitive
- “Single” key to the jewel box
- Uniqueness / High Selectivity

Dependent Sensitive Data

- Not inherently sensitive
- Requires context to “make” them sensitive
- “Multiple” keys to the jewel box
- Non-unique

Obviously Sensitive Data

- Clearly Identifiable / Recognizable
- Full-Valued

Hidden Sensitive Data

- Non-obvious
- Can be combined with other data to become sensitive
- Is derivable from Obviously Sensitive data
- Obviously sensitive data can be derived
- Types
  - Partial
  - Encoded
  - Re-Used / Overloaded / Multi-purposed

Types of Hidden Sensitive Data

- Partial
  - Examples: Last 4 digits of an SSN
  - Employee ID = First Initial+Last Initial+Last 4 digits of SSN
- Encoded
  - Demographic Codes
  - Accountholder Credit Score Ranking
  - Diseases and Disease Codes

DiseaseName	CodeValue
HIV positive	12
HIV negative	10
Cancer5Years	13
MajorSurgery5Years	14

Types of Hidden Sensitive Data

- Re-used / Overloaded / Multi-Purposed
  - Sensitive prior to some date or event; non-sensitive after
  - Non-sensitive prior to some date or event; sensitive after
  - One type of sensitive data prior to some date or event; different type after
  - Different types of sensitive data elements in the same field

Sensitive Data Element Categories

	Independent	Dependent
Obvious	**data that is sensitive standalone and exists as full values** ***Example: social security numbers appearing in its entirety***	**data that requires additional context to make it sensitive and exists as full values** ***Example: expiration date, Mother's maiden name.***
Hidden	**data that is sensitive standalone and exists as partial or as variations of full values.** ***Example: Last 4 digits of a social security number; Account number that includes the last 4 digits of social security number.***	**data requiring additional context to make it sensitive and exists as partial or variations of full values.** ***Example: Credit ranking***

A Data-Driven Methodology Overview

- Phase I
  - Discover Obviously Sensitive Data
  - Data Matching
- Phase II
  - Discover Hidden Sensitive Data
  - Data Mapping
- Phase III
  - Enable Data Flow Analysis
  - Data Lineage

A Data-Driven Methodology Overview

- Phase I
  - Identify and Locate full-valued matches
- Phase II
  - Types of Sensitive Data discovered:
    - Correlations and Fuzzy Matches
    - Encoded
    - Partial
    - Overloaded
- Phase III
  - Enable Lineage based on data mapping
  - Store Matching / Mapping results in a Metadata Repository
  - Enable Data Flow Analysis / Impact Analysis

Data-Driven Methodology Goals

- Comprehensive Coverage
- Process Enablement
  - Measurable
  - Repeatable
  - Accurate
  - Scalable
  - Fast

Sensitive Data Element Types and applicable Methodology Phases

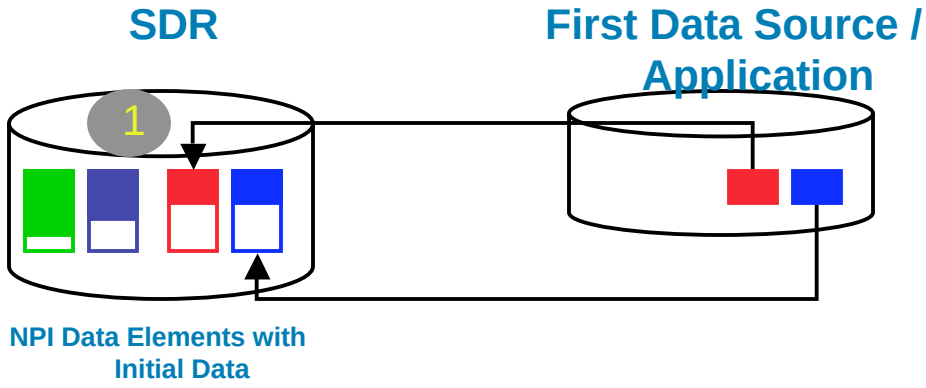
	INDEPENDENT	DEPENDENT
OBVIOUSLY SENSITIVE	PHASE 1	PHASE 1
HIDDEN SENSITIVE	PHASE 2	PHASE 2

Phase I

SDE list for PCI DSS (for example)

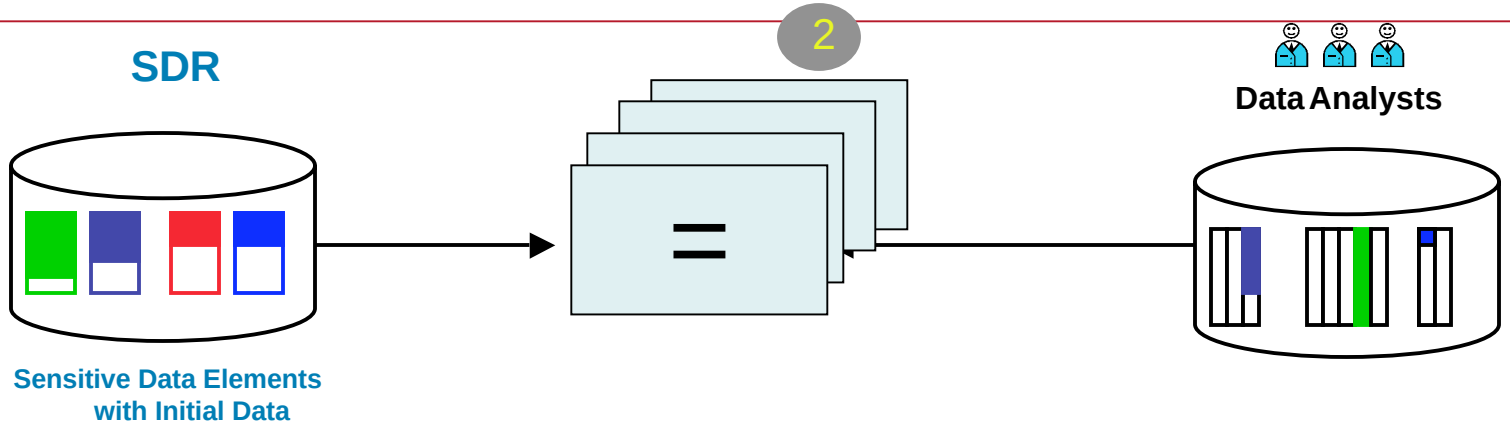
- Cardholder Account Number
- Cardholder Name
- Cardholder Address
- Cardholder Phone Number
- PIN
- Cardholder SSN
- CVV
- Track 1 data

1. Create Sensitive Data Repository (SDR)



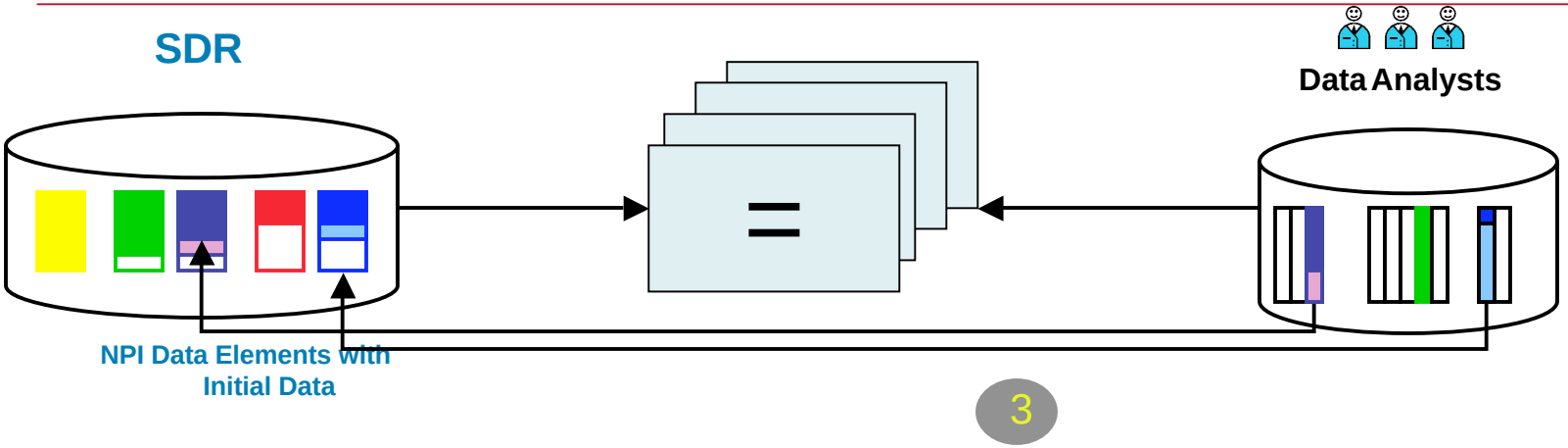
- Separate Column for each Sensitive Data Element
- Populate by
  - Generating typical sensitive data
  - Loading data from initial sources

2. Match against new source



- Use Concurrency and Parallelism for rapid analysis
- Analyze Results
  - Highly matched columns contain Sensitive Data
  - Columns with few matches are candidate Sensitive Data columns and require review by SME
    - Even a few matches are enough to point to where sensitive data is
    - As SDR grows, the accuracy of matching will increase and the need for SME review will decrease

3. Enrich SDR or Add new Data Element



- Enrich SDR with sensitive data from Data Source or
- Add new Sensitive Data Element (SDE)

Phase I Methodology - Example

Phase 1: Analyze the **<u>data values</u>** to **<u>automatically discover</u>** the Obviously Sensitive Data across disparate datasets

Account_Number = ACCT_NO (Table 26)

Cardholder_Name = Name (Table 26)

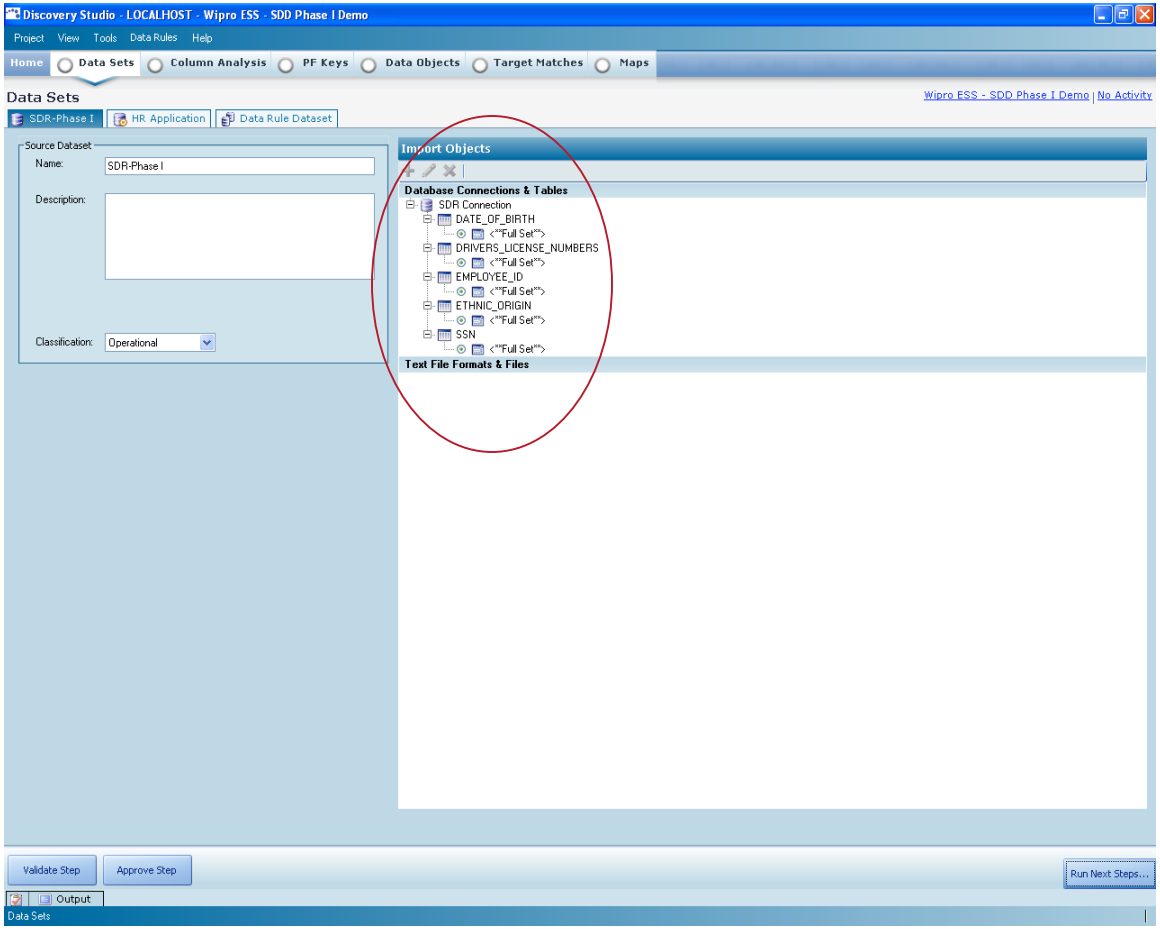
Sensitive Data Repository

Account_Number	Cardholder_Name	PIN
5958 4622 6327 9607	Jack Bauer	5890
5674 7259 6957 4267	Juan Valdez	7637
5404 5009 1482 2859	Robert Mey	1991
5147 1437 2793 3485	Mona Kavitzer	8696
4902 0416 4582 5681	Alfred E. Newman	4746
4668 6110 9867 2753	Rupert Murdoch	4453
⋮	⋮	⋮
4446 2962 8837 1957	Ralph Wiggum	
4234 5678 8389 8582	Bill Tiden	

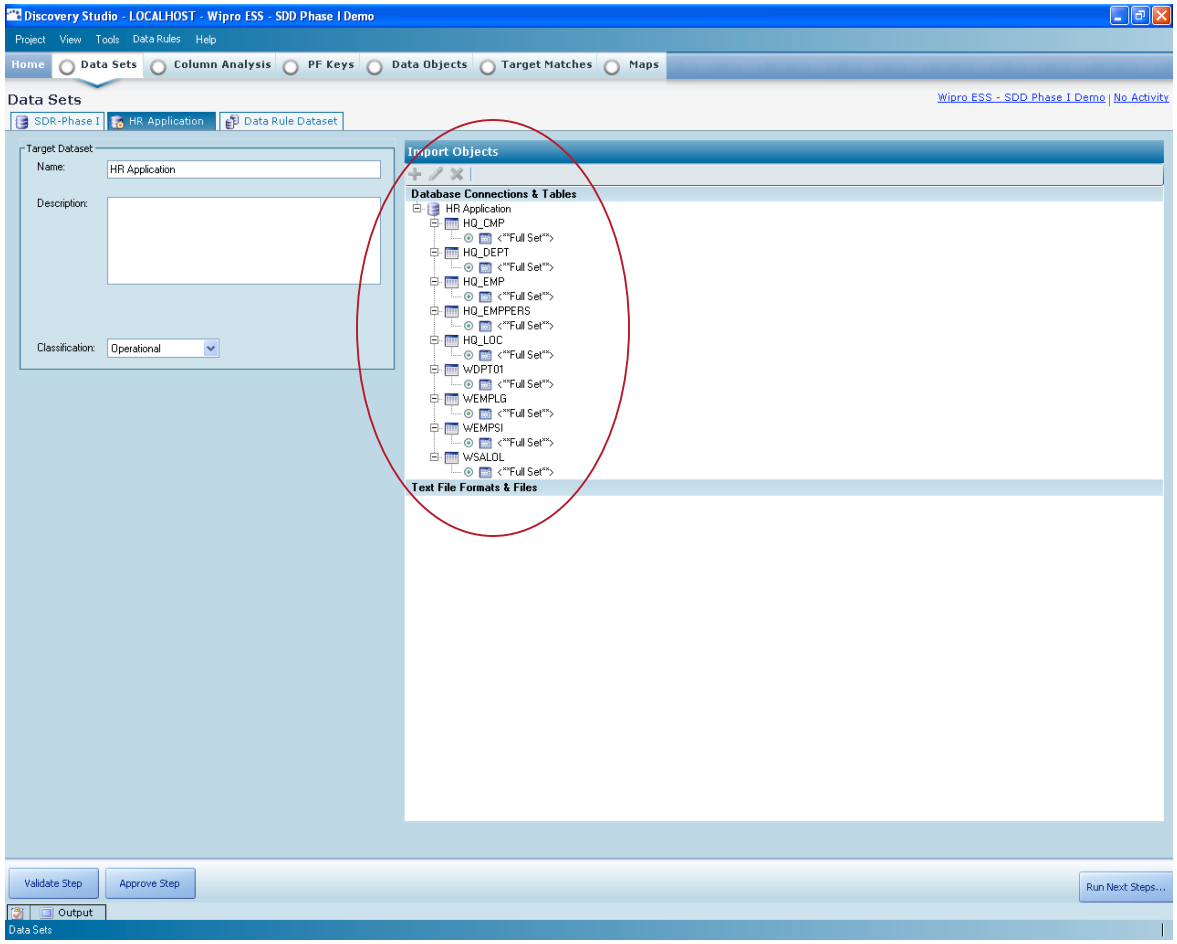
Table 26

Row	ACCT_NO	Name	State	SSN	Score
1	5958 4622 6327 9607	Robert Mey	CA	111-22-3333	678
2	5674 7259 6957 4267	Jack Bauer	UT	777-44-3321	730
3	5404 5009 1482 2589	Juan Valdez	CA	789-55-4433	590
4	5147 1437 2793 3485	Mona Kavitzer	UT	677-38-2121	423
5	4902 0416 4582 5681	Rupert Murdoch	UT	097-44-7378	760
6	4668 6110 9867 2753	Alfred E. Newman	UT	532-48-4884	666
⋮	⋮	⋮	⋮	⋮	⋮
987,623	4446 2962 8837 1957	Bill Tilden	CA	666-22-4444	590
987,624	4234 5678 8389 8582	Ralph Wiggum	CA	999-88-7755	300

Phase I Methodology – Sample Results



Phase I Methodology – Sample Results



Phase I Methodology – Sample Results

Show All Target Matches

Show Hits | Sort By Rate

Column Δ

	Source		Target			Row Hit Rate		Value Hit Rate		Selectivity	
Type	Table	Column Δ	Table	Column	Data Rule	Source	Target ∇	Source	Target	Source	Target
Column: BIRTH_DATE											
One-To-One	DATE_OF_BIRTH	BIRTH_DATE	WEMPLG	DATE_OF_BIRTH		54% (32/59)	28% (32/116)	53% (31/58)	27% (31/115)	98% (58/59)	99% (115/116)
One-To-One	DATE_OF_BIRTH	BIRTH_DATE	HQ_EMPPERS	DOB		100% (59/59)	26% (59/230)	100% (58/58)	27% (58/214)	98% (58/59)	93% (214/230)
One-To-One	DATE_OF_BIRTH	BIRTH_DATE	WEMPSI	DATE_OF_BIRTH		46% (27/59)	25% (26/102)	45% (26/58)	25% (26/102)	98% (58/59)	100% (102/102)
Column: EID											
One-To-One	EMPLOYEE_ID	EID	HQ_CMP	EMP_ID		94% (62/66)	27% (62/230)	94% (62/66)	27% (62/230)	100% (66/66)	100% (230/230)
One-To-One	EMPLOYEE_ID	EID	HQ_EMPPERS	EMPID		91% (60/66)	26% (60/230)	91% (60/66)	27% (60/224)	100% (66/66)	97% (224/230)
One-To-One	EMPLOYEE_ID	EID	HQ_EMP	EMPLOYEE_ID		100% (66/66)	26% (66/250)	100% (66/66)	26% (66/250)	100% (66/66)	100% (250/250)
Column: LICENSE											
One-To-One	DRIVERS_LICENSE_NUMBERS	LICENSE	HQ_EMPPERS	DRIVLICNO		100% (88/88)	38% (88/230)	100% (88/88)	43% (88/203)	100% (88/88)	88% (203/230)
One-To-One	DRIVERS_LICENSE_NUMBERS	LICENSE	WEMPSI	SID		42% (37/88)	36% (37/102)	42% (37/88)	37% (37/100)	100% (88/88)	98% (100/102)
Column: RACE											
One-To-Many	ETHNIC_ORIGIN	RACE	HQ_EMPPERS	EO		100% (3/3)	35% (81/230)	100% (3/3)	38% (3/8)	100% (3/3)	3% (8/230)
Column: SOCIAL_SECURITY_NUMBER											
One-To-One	SSN	SOCIAL_SECURITY_NUMBER	HQ_EMPPERS	SSN		100% (230/230)	100% (230/230)	100% (230/230)	100% (230/230)	100% (230/230)	100% (230/230)

Record 1 of 10

Close

Phase I Methodology- Overview

1. Create Sensitive Data Repository (SDR)
2. Match SDR values against each data source
3. Enrich SDR with each new source or Add new Sensitive Data Element (SDE).
4. Repeat steps 2 – 3 with all data sources until
   - No new Sensitive Data Element (SDE) discovered, and/or
   - Results are satisfactory
5. Delete / Truncate SDR

Phase II

Phase II Methodology - Example

- **Align the data sets**

- **Find encoded data**
- **Find partial data**
- **Find overloaded data**

CASE WHEN Score > 700 THEN 'A'
WHEN Score > 600 THEN 'B'... = CD

Table 26

Row	ACCT_NO	Name	State	SSN	Score
1	5958 4622 6327 9607	Robert Mey	CA	111-22-3333	678
2	5674 7259 6957 4267	Jack Bauer	UT	777-44-3321	730
3	5404 5009 1482 2589	Juan Valdez	CA	789-55-4433	590
4	5147 1437 2793 3485	Mona Kavitzer	UT	677-38-2121	423
5	4902 0416 4582 5681	Rupert Murdoch	UT	097-44-7378	760
6	4668 6110 9867 2753	Alfred E. Newman	UT	532-48-4884	666
•	•	•	•	•	•
•	•	•	•	•	•
•	•	•	•	•	•
987,623	4446 2962 8837 1957	Bill Tilden	CA	666-22-4444	590
987,624	4234 5678 8389 8582	Ralph Wiggum	CA	999-88-7755	300

Table 27

ACCT #	FN	SEC_ID	CD
5958 4622 6327 9607	Robert	111-22-3333	B
5674 7259 6957 4267	Jack	UT98765	A
5404 5009 1482 1991	Juan	789-55-4433	C
5147 1437 2793 3485	Mona	UT66666	D
4902 0416 4582 5681	Rupert	UT91823	A
4668 6110 9867 2753	Alfred	UT76521	B
•	•	•	•
•	•	•	•
•	•	•	•
4446 2962 8837 1957	Bill	666-22-4444	C
4234 5678 8389 8582	Ralph	999-88-7755	F

Token 1(Name) = FN

CASE WHEN State = 'CA' THEN SSN... = SEC_ID

Phase II Methodology - Example

- **Align the data sets**
- **Find correlated data**

Score <-> Category Correlation

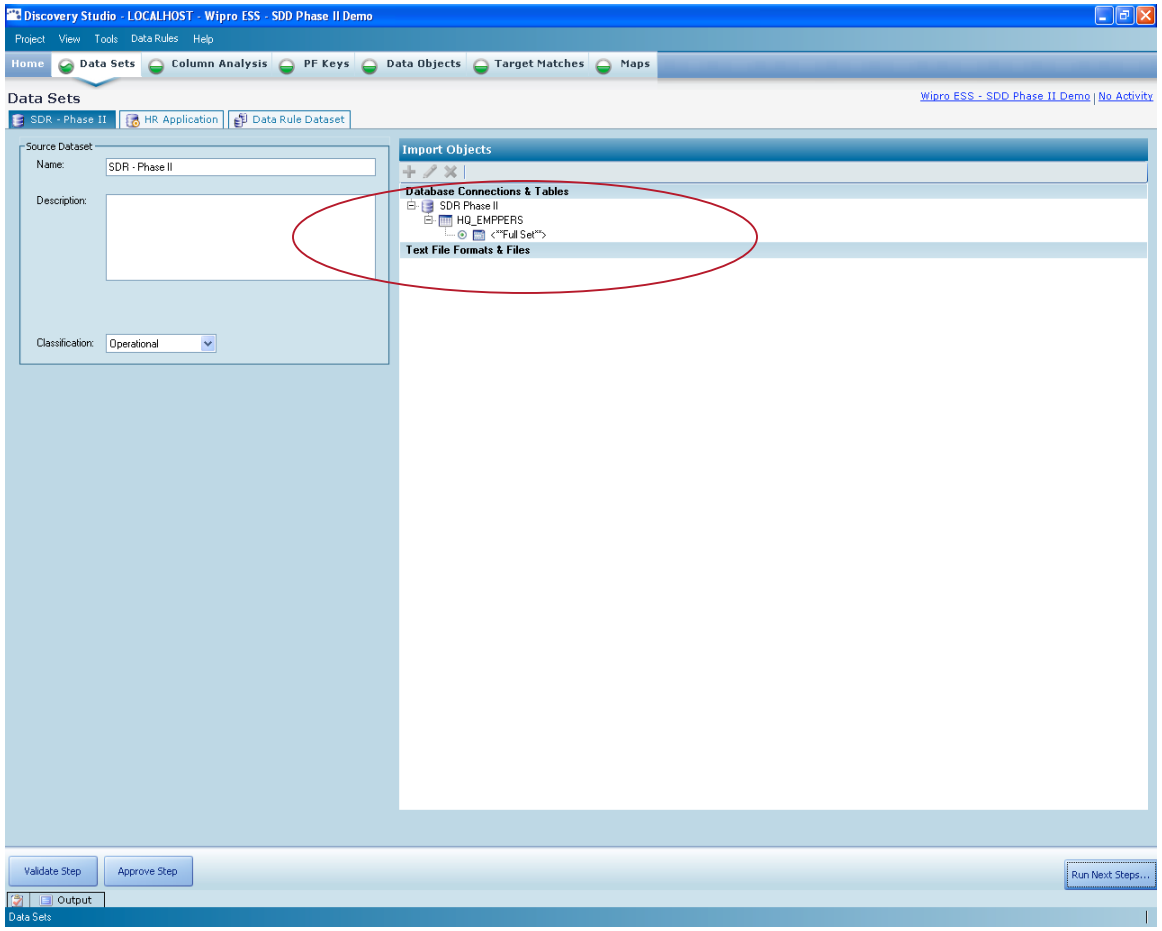
Table 26

Row	ACCT_NO	Name	State	SSN	Score
1	5958 4622 6327 9607	Robert Mey	CA	111-22-3333	678
2	5674 7259 6957 4267	Jack Bauer	UT	777-44-3321	730
3	5404 5009 1482 2589	Juan Valdez	CA	789-55-4433	590
4	5147 1437 2793 3485	Mona Kavitzer	UT	677-38-2121	423
5	4902 0416 4582 5681	Rupert Murdoch	UT	097-44-7378	760
6	4668 6110 9867 2753	Alfred E. Newman	UT	532-48-4884	666
•	•	•	•	•	•
•	•	•	•	•	•
•	•	•	•	•	•
987,623	4446 2962 8837 1957	Bill Tilden	CA	666-22-4444	590
987,624	4234 5678 8389 8582	Ralph Wiggum	CA	999-88-7755	300

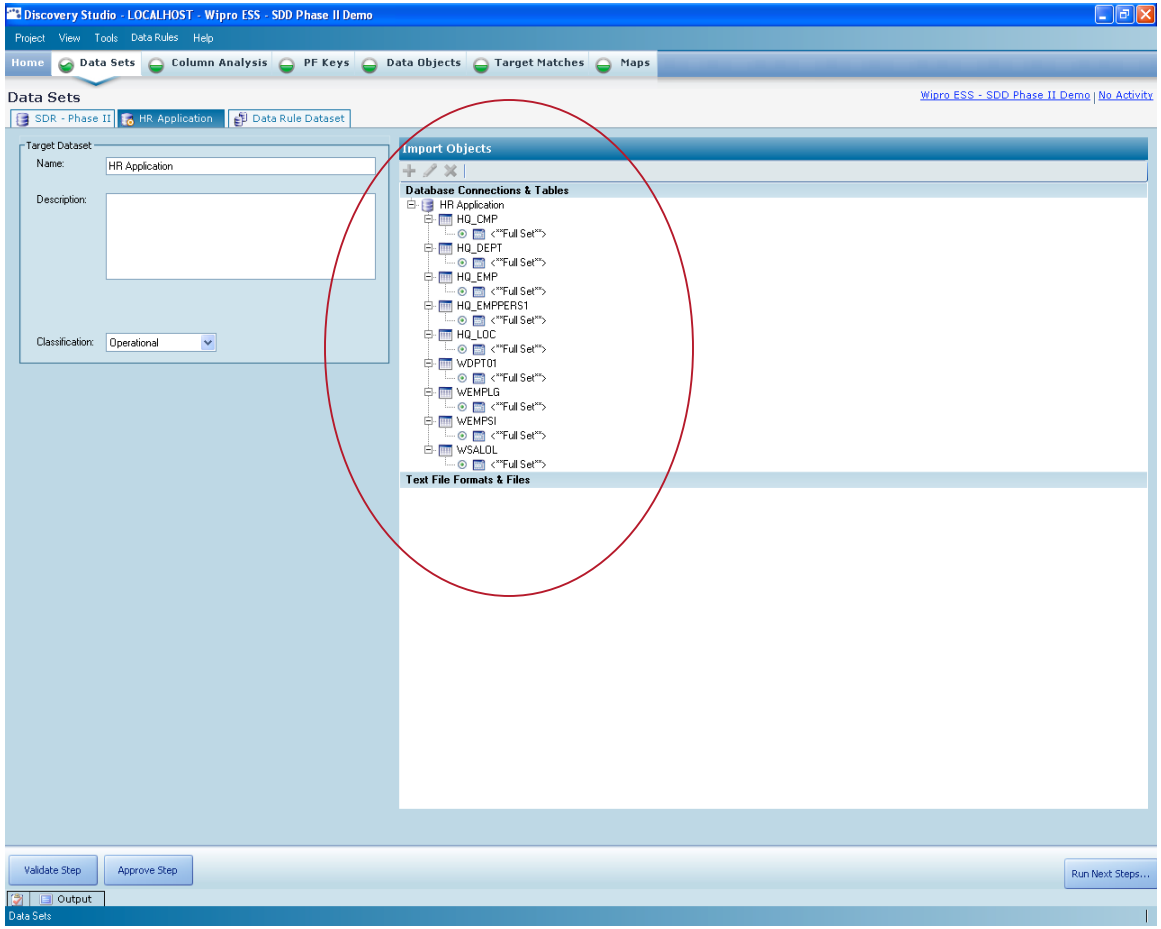
Table 72

ACCT #	Category
5958 4622 6327 9607	Gold
5674 7259 6957 4267	Diamond
5404 5009 1482 1991	Silver
5147 1437 2793 3485	Bronze
4902 0416 4582 5681	Diamond
4668 6110 9867 2753	Gold
•	•
•	•
•	•
4446 2962 8837 1957	Silver
4234 5678 8389 8582	N/A

Phase II Methodology - Sample Results



Phase II Methodology - Sample Results



Phase II Methodology - Sample Results

Report

Column

Source			Target				Transform	
Datasource	Table	...	Datasource	Table	Column	Row Hit Rate	Name	Expression
Column: SSN								
SDR Phase II	HQ_EMPPERS...	SSN	HR Application	WEMPLG (DM30DATA.WEMPLG)	SSN	100%(110/110)	Map...	substr(HQ_EMPPERS.SSN, 1, 3) || substr(HQ_EMPPERS.SSN, 5, 2) || substr(HQ_EMPPERS.SSN, 8, 4)
SDR Phase II	HQ_EMPPERS...	SSN	HR Application	WEMPSI (DM30DATA.WEMPSI)	ID	100%(96/96)	Map...	token(HQ_EMPPERS.SSN, 3)
SDR Phase II	HQ_EMPPERS...	SSN	HR Application	HQ_EMPPERS1 (DM30DATA.H...	SSN	100%(230/230)	Map...	HQ_EMPPERS.SSN
Column: PRIMARY_CONTACT_REL								
SDR Phase II	HQ_EMPPERS...	PRI...	HR Application	HQ_EMPPERS1 (DM30DATA.H...	PRIMARY_CONTACT_REL	100%(230/230)	Map...	HQ_EMPPERS.PRIMARY_CONTACT_REL
Column: PRIMARY_CONTACT_PH								
SDR Phase II	HQ_EMPPERS...	PRI...	HR Application	WEMPSI (DM30DATA.WEMPSI)	ECD	20%(20/96)	Map...	'ECD' || substr(HQ_EMPPERS.PRIMARY_CONTACT_PH, 10, 2)
SDR Phase II	HQ_EMPPERS...	PRI...	HR Application	HQ_EMPPERS1 (DM30DATA.H...	PRIMARY_CONTACT_PH	100%(230/230)	Map...	HQ_EMPPERS.PRIMARY_CONTACT_PH
Column: PRIMARY_CONTACT_NAME								
SDR Phase II	HQ_EMPPERS...	PRI...	HR Application	HQ_EMPPERS1 (DM30DATA.H...	PRIMARY_CONTACT_NAME	100%(230/230)	Map...	HQ_EMPPERS.PRIMARY_CONTACT_NAME
Column: GENDER								
SDR Phase II	HQ_EMPPERS...	GE...	HR Application	HQ_EMPPERS1 (DM30DATA.H...	GENDER	100%(230/230)	Map...	HQ_EMPPERS.GENDER
Column: EO								
SDR Phase II	HQ_EMPPERS...	EO	HR Application	HQ_EMPPERS1 (DM30DATA.H...	EO	100%(230/230)	Map...	HQ_EMPPERS.EO
Column: EMPID								
SDR Phase II	HQ_EMPPERS...	EMPID	HR Application	HQ_EMPPERS1 (DM30DATA.H...	EMPID	100%(230/230)	Map...	HQ_EMPPERS.EMPID
Column: DRIVLICNO								
SDR Phase II	HQ_EMPPERS...	DRI...	HR Application	WEMPSI (DM30DATA.WEMPSI)	SID	82%(79/96)	Map...	HQ_EMPPERS.DRIVLICNO
SDR Phase II	HQ_EMPPERS...	DRI...	HR Application	HQ_EMPPERS1 (DM30DATA.H...	DRIVLICNO	100%(230/230)	Map...	HQ_EMPPERS.DRIVLICNO
Column: DOH								
SDR Phase II	HQ_EMPPERS...	DOH	HR Application	WEMPLG (DM30DATA.WEMPLG)	BEGIN_DATE	97%(107/110)	Map...	HQ_EMPPERS.DOH
SDR Phase II	HQ_EMPPERS...	DOH	HR Application	HQ_EMPPERS1 (DM30DATA.H...	DOH	100%(230/230)	Map...	HQ_EMPPERS.DOH
Column: DOB								
SDR Phase II	HQ_EMPPERS...	DOB	HR Application	WEMPLG (DM30DATA.WEMPLG)	DATE_OF_BIRTH	100%(110/110)	Map...	HQ_EMPPERS.DOB
SDR Phase II	HQ_EMPPERS...	DOB	HR Application	WEMPLG (DM30DATA.WEMPLG)	EID	99%(109/110)	Map...	datarule(DR_EID_0, HQ_EMPPERS.DOB)
SDR Phase II	HQ_EMPPERS...	DOB	HR Application	WEMPSI (DM30DATA.WEMPSI)	DATE_OF_BIRTH	100%(96/96)	Map...	HQ_EMPPERS.DOB
SDR Phase II	HQ_EMPPERS...	DOB	HR Application	HQ_EMPPERS1 (DM30DATA.H...	DOB	100%(230/230)	Map...	HQ_EMPPERS.DOB

Close

Phase II Methodology - Overview

1. Include , in a new SDR, tables with Sensitive Data columns in application/s
   - Select sub-set of tables from Phase I results
   - (Optionally) prune one or more of these tables to contain
     - All independent and dependent sensitive data columns
     - Any candidate key columns
     - Other relevant columns
2. Map these SDR tables to other tables in application/other applications
3. Discover Hidden Sensitive Data:
   - Encoded values
   - Partial data
   - Overloaded columns
4. Repeat steps 1-3 per Application as needed until all hidden sensitive data discovered

Phase III

Phase III Methodology - Overview

1. Store discovered Phase I (Data Matching) and Phase II (Data Mapping) in a metadata repository
2. Enable Data Flow Analysis / Data Lineage Analysis / Impact Analysis

Methodology Highlights

- Automation
- Auto-refinement and Coverage
- Scalable with high-performance
- Repeatable, Measurable and Accurate
- Extensible

Q&A