



Automated SmartData Discovery and Classification (ASDAC), and Data Lineage

March 2021



WhamTech approach to data discovery, governance and lineage

- There are great INFORMATIONAL data catalog, data governance and/or data lineage tools available, e.g., Alation and Collibra, and if a customer is JUST seeking such a tool, we would recommend they consider these, however:
 - Such INFORMATIONAL tools need to be integrated with other APPLICABLE data-related tools such as access control, applied data governance, data transformation/ETL, data security, master data management, data integration, query execution, graph database, etc. - can take time and cost, and ultimately, not be seamless, performant or scalable
 - WhamTech SmartData Fabric® can complement and leverage such INFORMATIONAL tools, and APPLY their INFORMATION to rigorous and full data management
- However, if a customer is seeking a tool that provides both INFORMATION and APPLICATION, then SmartData Fabric® can provide both to the full range of data management from discovery to delivery, to insight
 - ACTIVE data catalog, ACTIVE data governance (combined with ACTIVE access control) and ACTIVE data lineage in addition to the rest of the data management process



Customer Challenges

- Incomplete understanding of all data sources and data in the enterprise
- Concerns for locations and associated risk of personal data for PCI, data theft risk, CCPA/CCPR, GDPR, etc.
- Need a means to secure personal data, but still allow it to be used for operations
- Make metadata and data governance information available through a data catalog
- Use the data catalog to impose data governance and rigorous enterprise-wide access control with cross-domain AD/LDAP, self-serve apps and execute SQL queries
- Track and audit data from discovery to delivery to apps
- Apply data management and master data management to create an enterprise virtual data warehouse
- Analyze and report on data usage, and tie-in with User Behavior Analytics (UBA)
- All the above in near real-time – little to no latency

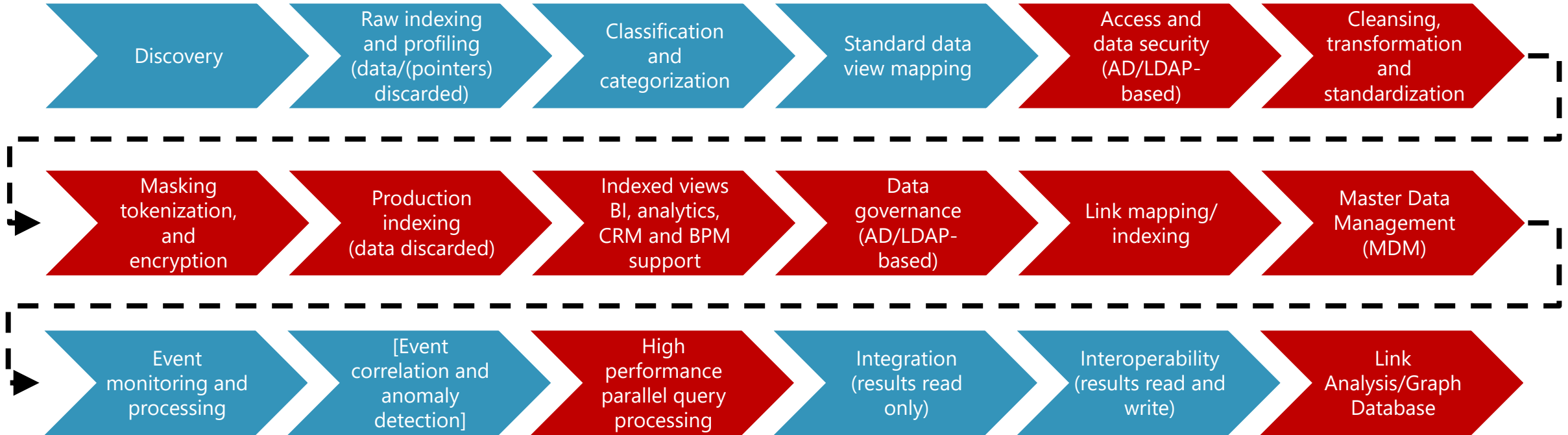


SmartData Fabric® (SDF) Solution

- Run ASDAC* to automatically discover devices and data sources in the network
- Build raw indexes on data sources with appropriate credentials – do not store data or pointers
- Use data profiles generated from raw indexes to automatically obtain metadata, semantically identify and classify data, and manually, for now, develop and test data transforms
- Map indexes to a virtual standard (industry) data model – multiple virtual views available
- Handover to next stage, near real-time production indexing and federated adapter configuration
- Using data transforms, build cleansed and standardized production indexes with unique record numbers (URNs) and indexed views, and make available through federated adapters and standard virtual views to standard apps
- Use multiple adapters and indexes to build SDF and execute SQL with rigorous data governance and cross-domain access control through standard drivers and APIs, including RBAC-based dynamic data masking, tokenization and/or encryption
- Provide results to standard apps with data lineage and any other metadata
- Combine SDF activity logs with server logs for complete reporting of data and app use, and user behavior - can also restrict what days and times a role or user can submit queries
- Interact with a User Behavior Analytics (UBA) app to monitor, analyze and limited user access



SmartData Fabric® for data virtualization, federation and integration

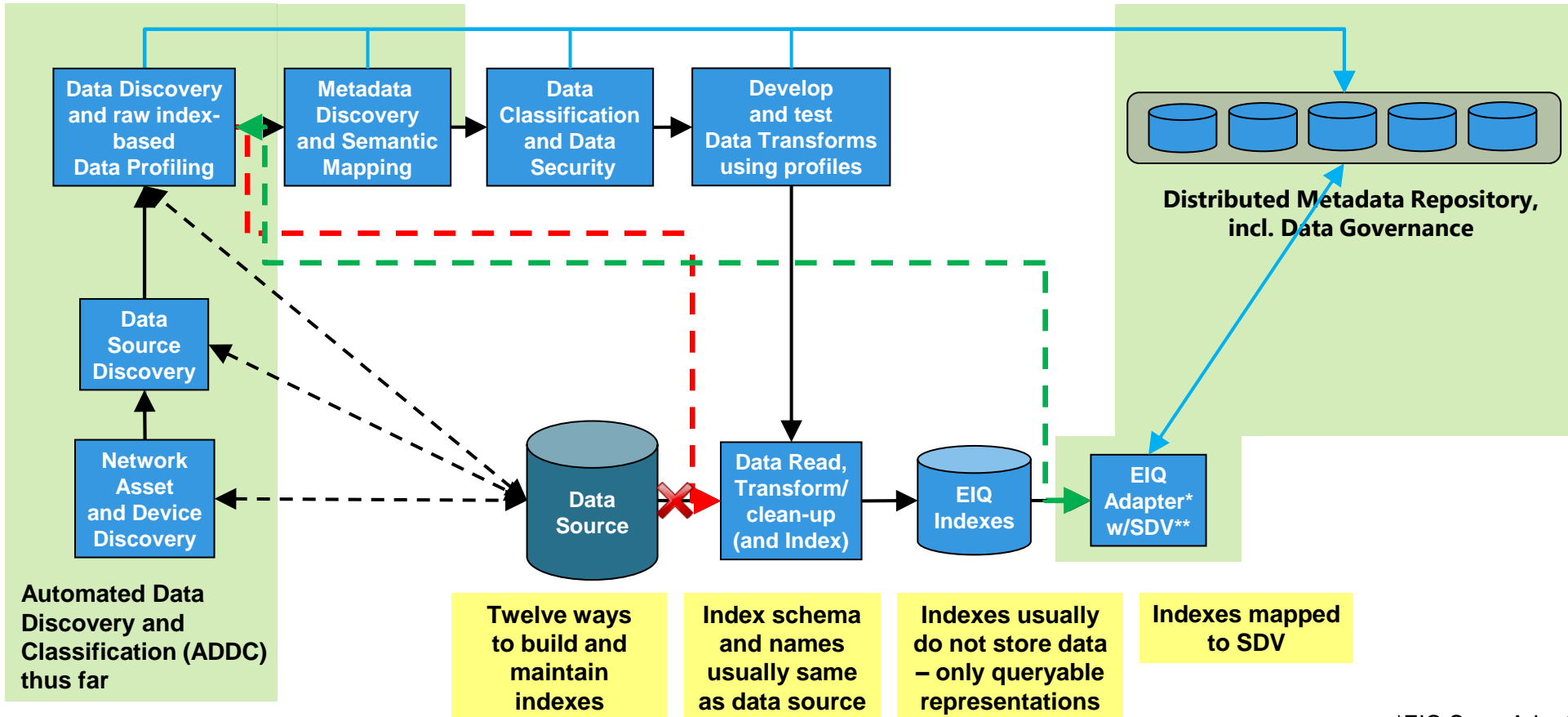


WhamTech key differentiators addressed upfront



WHAMTECH®

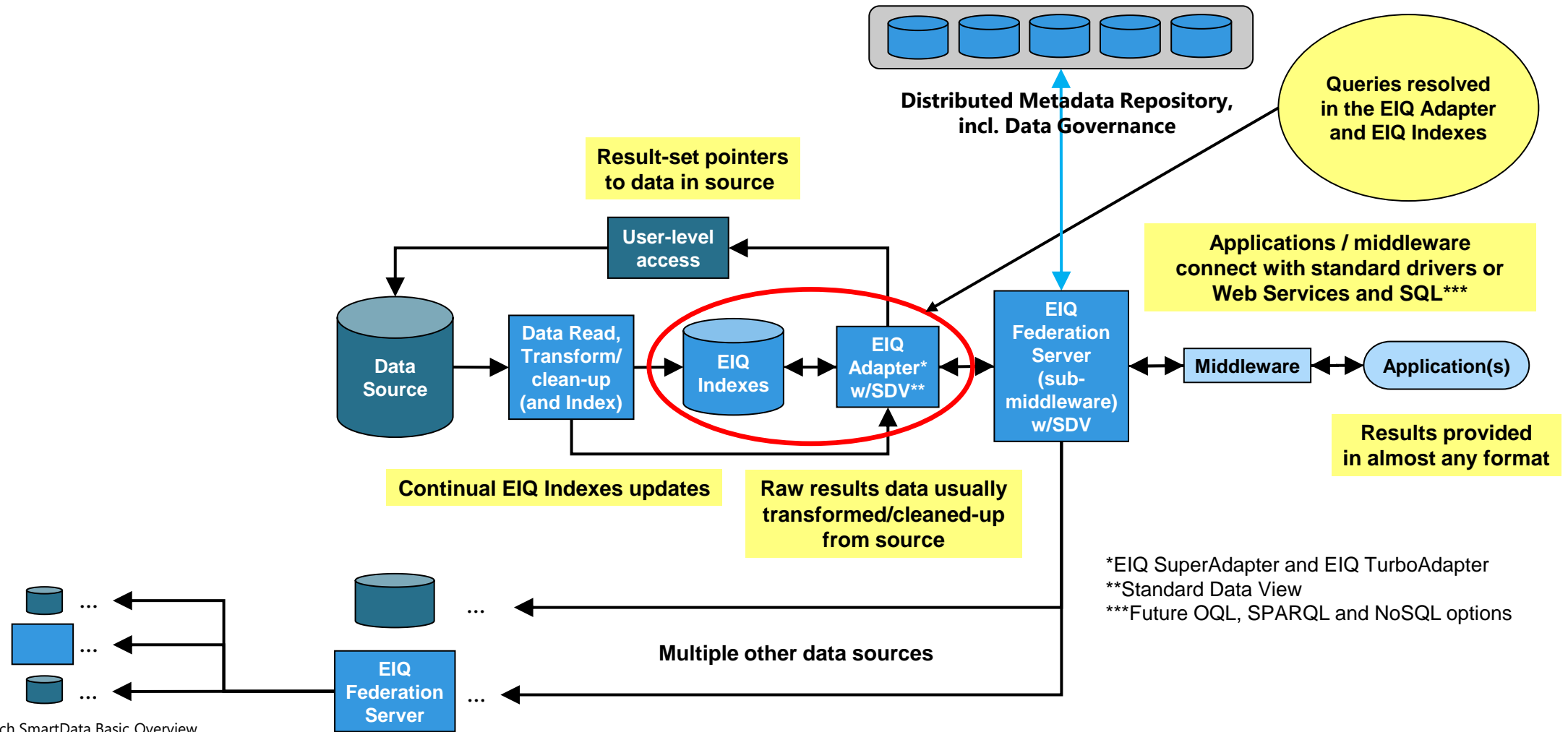
Initial EIQ Adapter configuration, index build and data view mapping



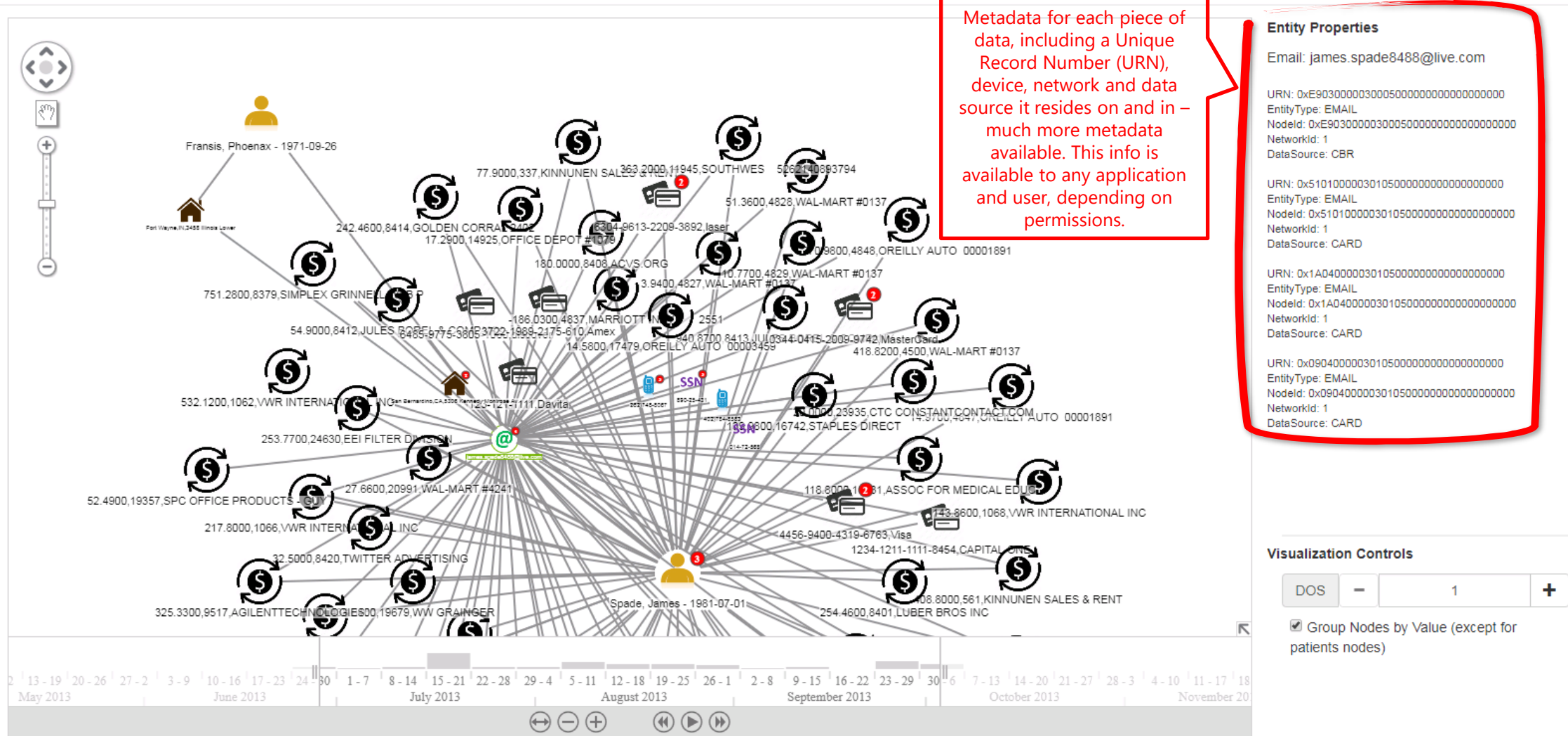
*EIQ SuperAdapter and EIQ TurboAdapter
**Standard Data View

Alternate use of raw indexes to initially build EIQ Indexes

EIQ Adapter index update, query and results retrieval



Example of metadata and data traceability in a graph view of financial services data



Metadata for each piece of data, including a Unique Record Number (URN), device, network and data source it resides on and in – much more metadata available. This info is available to any application and user, depending on permissions.

Entity Properties

Email: james.spade8488@live.com

URN: 0xE903000003000500000000000000000000
 EntityType: EMAIL
 Nodeld: 0xE9030000030005000000000000000000
 NetworkId: 1
 DataSource: CBR

URN: 0x5101000003010500000000000000000000
 EntityType: EMAIL
 Nodeld: 0x51010000030105000000000000000000
 NetworkId: 1
 DataSource: CARD

URN: 0x1A04000003010500000000000000000000
 EntityType: EMAIL
 Nodeld: 0x1A040000030105000000000000000000
 NetworkId: 1
 DataSource: CARD

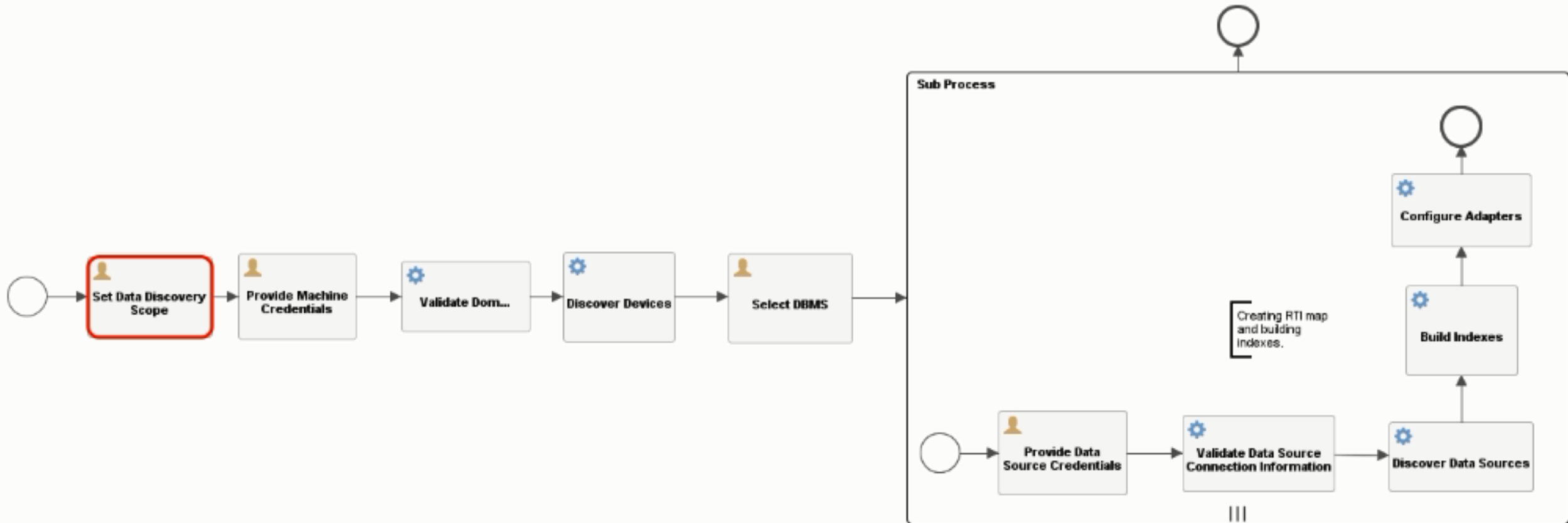
URN: 0x0904000003010500000000000000000000
 EntityType: EMAIL
 Nodeld: 0x09040000030105000000000000000000
 NetworkId: 1
 DataSource: CARD

Visualization Controls

DOS - 1 +

Group Nodes by Value (except for patients nodes)

BPM workflow on top of REST APIs










Screenshots of ASDAC in action



Launch the ASDAC app in a Web browser

Automated SmartData Discovery and Classification By  **SmartData Fabric™**

   Gavin

-  Process Dashboard
-  Task Inbox (0)
-  Process Status
-  Repository Explorer
-  Export Repository

Toggle Process Diagram

Clear Data Discovery Repository

Business Process Model and Notation (BPMN) Inbox

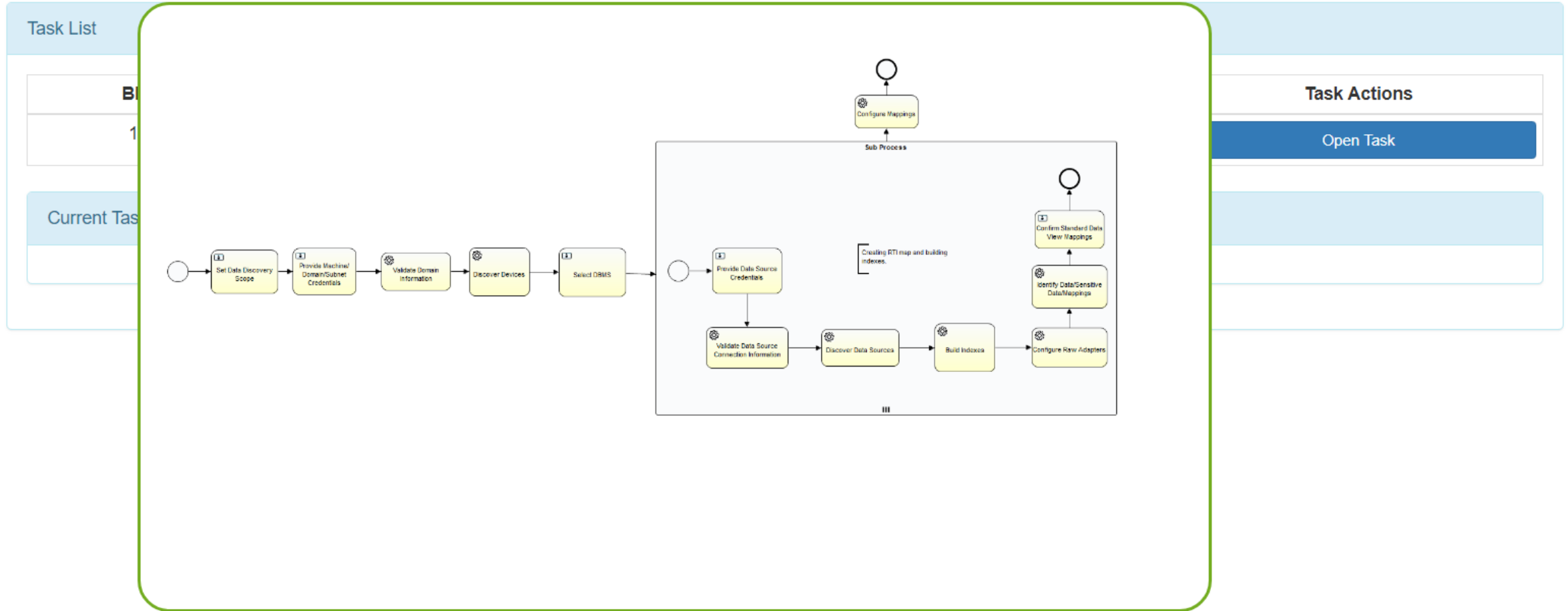
BPMN Process	Actions
Data Discovery and Classification	Start

BPM workflow on top of REST APIs

Automated SmartData Discovery and Classification By



SmartData Fabric™





Providing domain admin user credentials

Automated SmartData Discovery and Classification By



SmartData Fabric™



Gavin

Task List

BPMN Id	Task Id	Task Description	Task Actions
100077	100085	Provide Machine/Domain/Subnet Credentials	Task running...

Provide Machine/Domain/Subnet Credentials (BPMN Id: 100077 Task Id: 100085)

Provide Machine/Domain/Subnet Credentials

Fill information to connect and initiate the data discovery

User Name:

Password

Complete Task



Selecting discovered data sources to analyze

Automated SmartData Discovery and Classification By



SmartData Fabric™

Task List

BPMN Id	Task Id	Task Description	Task Actions
92577	92592	Select DBMS	Task running...

Select DBMS (BPMN Id: 92577 Task Id: 92592)

Data Source Selection

The following data sources were discovered. Please select the data sources to initiate data discovery

List discovered devices, data sources and drivers

- win-j3s4eclj043\dev.whamtech.com
 - * tstbox\dev.whamtech.com
 - * win-771udhc70rt\dev.whamtech.com
 - * testbox-1\dev.whamtech.com
 - * asdac\dev.whamtech.com
 - * sqlserv\dev.whamtech.com
 - * DBMS
 - * SQL Server
 - * Drivers
 - * oraserv\dev.whamtech.com
 - * DBMS
 - * Oracle
 - * Drivers

Complete Task



Providing data source credentials to build raw indexes and configure an adapter

WHAMTECH

Task List

BPMN Id	Task Id	Task Description	Task Actions
92577	92613	Provide Data Source Credentials	Task running...
92577	92615	Provide Data Source Credentials	Open Task

Provide Data Source Credentials (BPMN Id: 92577 Task Id: 92613)

Data Source Information

Provide data source credentials to initiate data discovery on the selected data sources

oraserv\dev.whamtech.com@Oracle

oraserv

cbr

...

1521

creditbureau

Complete Task



Data discovery and automated data identification results with Standard Data View mappings and sensitive data information

Repository Explorer

- DBMS
 - SQL Server
 - CREDITCARD
 - sys.SYS#_VIEWS
 - CARD.ACCOUNT
 - CARD.ACTTT
 - CARD.CUSMS
 - CUFN
 - CULN
 - CUDOB
 - CUEML
 - CUG
 - CUSSN
 - CUADSTR
 - CUADC
 - CUADST
 - CUADZC
 - CUPN
 - ID
 - LONGITUDE
 - LATITUDE
 - CARD.MERCHANT
 - CREDITCARD_ASDAC
 - master
 - model
 - msdb
 - NORTHWIND
 - POLICE
 - ReportServer
 - ReportServerTempDB
 - tempdb
 - XDatatypes

Properties for column: CARD.CUSMS.CUFN

Column Properties | General Analysis | Value Distribution | Most/Least Frequent

Data Discovery Findings

This column very likely contains sensitive information

Suggested Standard Data Model Column Mappings

Very Likely | Likely | Possible

CUSTOMER_FIRST_NAME

General Properties

Property	Value
TABLE_NAME	CUSMS
COLUMN_NAME	CUFN
DATA_TYPE_NAME	VARCHAR
SUPERSHEMA_FLAGS	0
NUMERIC_PRECISION	null
INTERNAL_NAME	T3C1
COLLATION_NAME	null



Data analysis and profile used for identification and, eventually, data transforms

Repository Explorer

- DBMS
 - SQL Server
 - CREDITCARD
 - sys.SYS#_VIEWS
 - CARD.ACCOUNT
 - CARD.ACTTT
 - CARD.CUSMS
 - CUFN
 - CULN
 - CUDOB
 - CUEML
 - CUG
 - CUSSN
 - CUADSTR
 - CUADC
 - CUADST
 - CUADZC
 - CUPN
 - ID
 - LONGITUDE
 - LATITUDE
 - CARD.MERCHANT
 - CREDITCARD_ASDAC
 - master
 - model
 - msdb
 - NORTHWIND
 - POLICE
 - ReportServer
 - ReportServerTempDB
 - tempdb
 - XDataFvnes

Properties for column: CARD.CUSMS.CUFN

Column Properties | **General Analysis** | Value Distribution | Most/Least Frequent

General Analysis

Property	Value
Minimum Value	AADEN
Maximum Value	ZOEY
Null Count	0
Distinct Count	769
Row Count	1051
Blank Count	0
Duplicate Count	213
Unique Count	556



Display the value distribution for the selected column (1 of 2)



Automated SmartData Discovery and Classification

Gavin

- Process Dashboard
- Task Inbox (0)
- Process Status
- Repository Explorer**
- Export Repository

Toggle Process Diagram

Clear Data Discovery Repository

Repository Explorer

- * DBO.EMPLOYERS
- * DBO.INSURANCE_CATEGORIES
- * DBO.PATIENT_POLICIES
- * DBO.PATIENTS
- * DBO.POLICIES
- * DBO.PRACTITIONERS
- * DBO.SERVICE_DIAGNOSES
- * DBO.SERVICES
- * DBO.VOUCHERS
 - * PATIENT_ID
 - * ACCOUNT_ID
 - * REF_PRACTITIONER_ID
 - * ACTUAL_PROV_PRACTITIONER_ID
 - * VOUCHER_ID
 - * CARRIER_ID
 - * LOCATION_ID
 - * DEPARTMENT_ID
 - * PLACE_OF_SERVICE_ID
 - * UPDATE_STATUS
 - * DATE_UPDATED
 - * FEES
 - * POSTED_PAYMENTS
 - * POSTED_ADJUSTMENTS
 - * POSTED_REFUNDS
 - * POSTED_MISC_DEBITS
 - * SERVICE_DATE
 - * BILLING_DATE
- * POLICE
- * tempdb
- * test_db
- * WORKS_NEW

Properties for column: VOUCHERS.POSTED_PAYMENTS

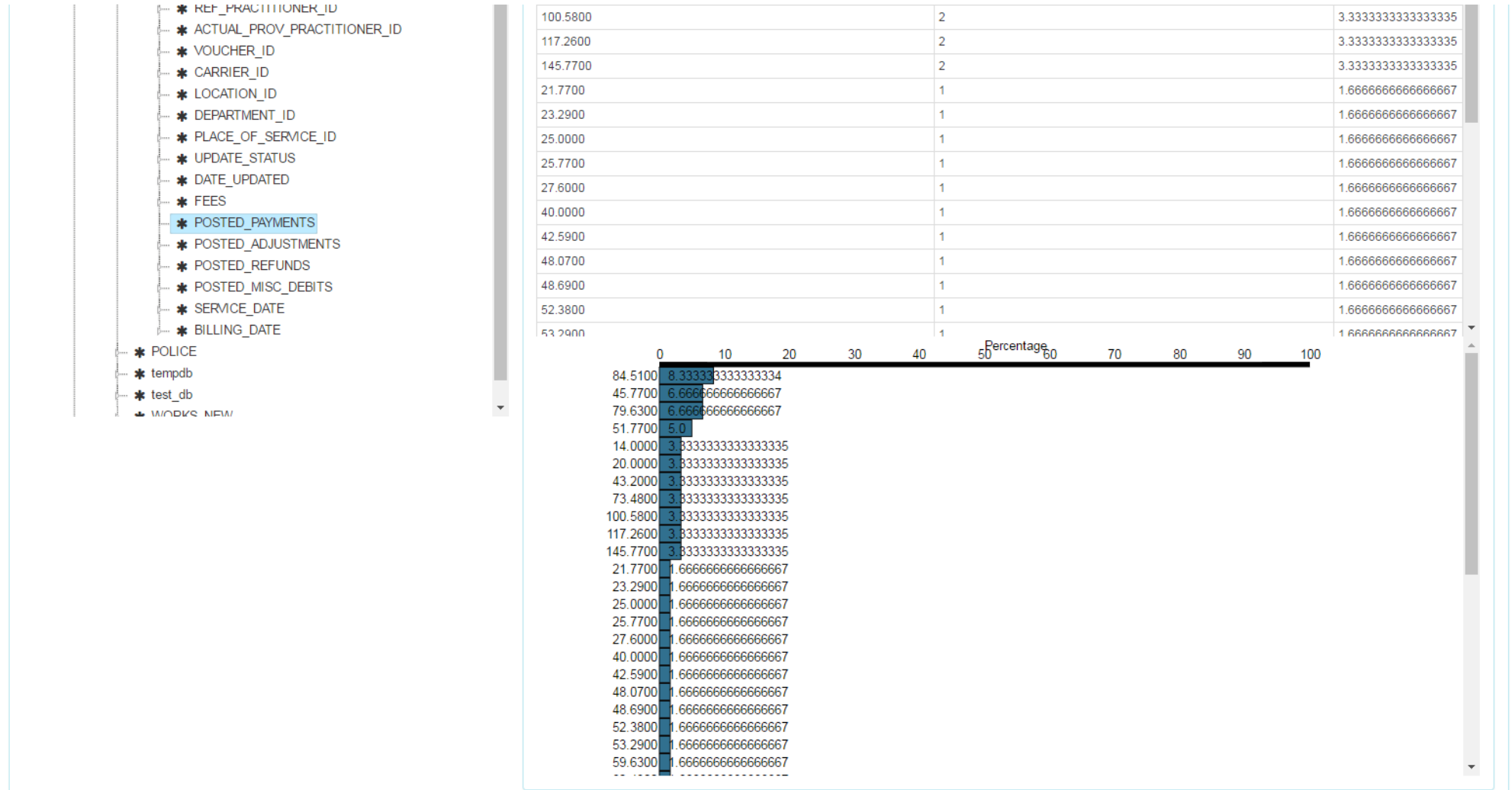
Column Properties General Analysis **Value Distribution** Most/Least Frequent Numerical Analysis

Value distribution

Value	Frequency	Percentage
84.5100	5	8.333333333333334
45.7700	4	6.666666666666667
79.6300	4	6.666666666666667
51.7700	3	5.0
14.0000	2	3.3333333333333335
20.0000	2	3.3333333333333335
43.2000	2	3.3333333333333335
73.4800	2	3.3333333333333335
100.5800	2	3.3333333333333335
117.2600	2	3.3333333333333335
145.7700	2	3.3333333333333335
21.7700	1	1.6666666666666667
23.2900	1	1.6666666666666667
25.0000	1	1.6666666666666667
25.7700	1	1.6666666666666667
27.6000	1	1.6666666666666667
40.0000	1	1.6666666666666667



Display the value distribution for the selected column (2 of 2)





Display the most and least frequent values for the selected column



Automated SmartData Discovery and Classification

🔗 👤 Gavin

- 🏠 Process Dashboard
- 📧 Task Inbox (0)
- 📊 Process Status
- 🔍 Repository Explorer**
- 📄 Export Repository

Toggle Process Diagram

Clear Data Discovery Repository

Repository Explorer

- * DBO.EMPLOYERS
- * DBO.INSURANCE_CATEGORIES
- * DBO.PATIENT_POLICIES
- * DBO.PATIENTS
- * DBO.POLICIES
- * DBO.PRACTITIONERS
- * DBO.SERVICE_DIAGNOSES
- * DBO.SERVICES
- * DBO.VOUCHERS
 - * PATIENT_ID
 - * ACCOUNT_ID
 - * REF_PRACTITIONER_ID
 - * ACTUAL_PROV_PRACTITIONER_ID
 - * VOUCHER_ID
 - * CARRIER_ID
 - * LOCATION_ID
 - * DEPARTMENT_ID
 - * PLACE_OF_SERVICE_ID
 - * UPDATE_STATUS
 - * DATE_UPDATED
 - * FEES
 - * POSTED_PAYMENTS**
 - * POSTED_ADJUSTMENTS
 - * POSTED_REFUNDS
 - * POSTED_MISC_DEBITS
 - * SERVICE_DATE
 - * BILLING_DATE
- * POLICE
- * tempdb
- * test_db

Properties for column: VOUCHERS.POSTED_PAYMENTS

Column Properties General Analysis Value Distribution **Most/Least Frequent** Numerical Analysis

Most Frequent Values

POSTED_PAYMENTS	cnt
84.5100	5
45.7700	4
79.6300	4
51.7700	3
14.0000	2
20.0000	2
43.2000	2
73.4800	2
100.5800	2
117.2600	2

Least Frequent Values

POSTED_PAYMENTS	cnt
21.7700	1
23.2900	1
25.0000	1
25.7700	1
27.6000	1



Display the numerical analysis/metadata for the selected column



Automated SmartData Discovery and Classification

Gavin

- Process Dashboard
- Task Inbox (0)
- Process Status
- Repository Explorer**
- Export Repository

Toggle Process Diagram

Clear Data Discovery Repository

Repository Explorer

- * DBO.EMPLOYERS
- * DBO.INSURANCE_CATEGORIES
- * DBO.PATIENT_POLICIES
- * DBO.PATIENTS
- * DBO.POLICIES
- * DBO.PRACTITIONERS
- * DBO.SERVICE_DIAGNOSES
- * DBO.SERVICES
- * DBO.VOUCHERS
 - * PATIENT_ID
 - * ACCOUNT_ID
 - * REF_PRACTITIONER_ID
 - * ACTUAL_PROV_PRACTITIONER_ID
 - * VOUCHER_ID
 - * CARRIER_ID
 - * LOCATION_ID
 - * DEPARTMENT_ID
 - * PLACE_OF_SERVICE_ID
 - * UPDATE_STATUS
 - * DATE_UPDATED
 - * FEES
 - * **POSTED_PAYMENTS**
 - * POSTED_ADJUSTMENTS
 - * POSTED_REFUNDS
 - * POSTED_MISC_DEBITS
 - * SERVICE_DATE
 - * BILLING_DATE
- * POLICE
- * tempdb
- * test_db

Properties for column: VOUCHERS.POSTED_PAYMENTS

Column Properties General Analysis Value Distribution Most/Least Frequent **Numerical Analysis**

Numerical Analysis

Property	Value
Average	82.3100
Standard Deviation	49.4300
Variance	2443.5500
Sum	4938.3200



Automated mappings for review and confirmation by user

BPMN Id	Task Id	Task Description	Task Actions
92577	92634	Confirm Standard Data View Mappings	Task running...

Confirm Standard Data View Mappings (BPMN Id: 92577 Task Id: 92634)

Confirm Standard Data View Mappings
The following mappings were selected by default. Please confirm the suggested standard data mappings.

List discovered devices, data sources and drivers

- * ST
- * DISP
- * MTD
- * DRN
- * CARD.CUSMS
 - ✓ CUFN
 - ✓ CULN
 - ⚠ CUDOB
 - ✓ CUEML
 - ✓ CUG
 - ⚠ CUSSN
 - ⚠ CUADSTR
 - ✓ CUADC
 - ✓ CUADST
 - ✓ CUADZC
 - ⚠ CUPN
- * ID
- * LONGITUDE
- ⚠ LATITUDE
- * CARD.MERCHANT
 - * ID
 - * MERCHANT
 - ⚠ STREET
 - ⚠ STREET2
 - * CITY
 - ✓ STATE
 - ✓ ZIPCODE
 - ✓ PHONENUMBER
 - ⚠ CATEGORY
 - * LONGITUDE
 - ⚠ LATITUDE

Complete Task

Alert the user for mappings that are not very likely or have conflicts

Confirm Standard Data View Mappings (BPMN Id: 92577 Task Id: 92634)

Confirm Standard Data View Mappings
The following mappings were selected by default. Please confirm the suggested standard data mappings.

List discovered devices, data sources and drivers

- * CUID
- * ACID
- * TRDD
- * DPST
- * MEID
- C2657AMV
- * ST
- * DISP
- * MTD
- * DRN
- * CARD.CUSMS
 - CUFN
 - CULN
 - CUDOB
 - CUEML
 - CUG
 - CUSSN
 - CUADSTR
 - CUADC
 - CUADST
 - CUADZC
 - CUPN
 - * ID
 - * LONGITUDE
 - LATITUDE
- * CARD.MERCHANT
 - * ID
 - * MERCHANT
 - STREET
 - STREET2
 - * CITY

Properties for column: CARD.CUSMS.CUADSTR

Column Properties General Analysis Value Distribution Most/Least Frequent

Data Discovery Findings

This column possibly contains sensitive information

Suggested Standard Data Model Column Mappings

Very Likely Likely Possible

*CUSTOMER_STREET

General Properties

Property	Value
TABLE_NAME	CUSMS
COLUMN_NAME	CUADSTR
DATA_TYPE_NAME	VARCHAR
SUPERSHEMA_FLAGS	0
NUMERIC_PRECISION	null
INTERNAL_NAME	T3C7
COLLATION_NAME	null
DESCRIPTION	null
KEYINFO	549
ORIGINAL_SCHEMA	CARD
COLLATION_CATALOG	null
CHARACTER_MAXIMUM_LENGTH	100
NATIVE_ROWBUFFER_OFFSET	171
COLUMN_GUID	null
DATA TYPE	129

Green flag for very likely matches and no conflicts

Confirm Standard Data View Mappings (BPMN Id: 92577 Task Id: 92634)

Confirm Standard Data View Mappings

The following mappings were selected by default. Please confirm the suggested standard data mappings.

List discovered devices, data sources and drivers

- * CUID
- * ACID
- * TRDD
- * DPST
- * MEID
- ⚠ C2657AMV
- * ST
- * DISP
- * MTD
- * DRN
- * CARD.CUSMS
 - ✓ CUFN
 - ✓ CULN
 - ⚠ CUDOB
 - ✓ CUEML
 - ✓ CUG
 - ⚠ CUSSN
 - ⚠ CUADSTR
 - ✓ CUADC
 - ✓ CUADST
 - ✓ CUADZC
 - ⚠ CUPN
- * ID
- * LONGITUDE
- ⚠ LATITUDE
- * CARD.MERCHANT
 - * ID
 - * MERCHANT
 - ⚠ STREET
 - ⚠ STREET2
 - * CITY

Properties for column: CARD.CUSMS.CUADC

Column Properties General Analysis Value Distribution Most/Least Frequent

Data Discovery Findings

This column very likely contains sensitive information

Suggested Standard Data Model Column Mappings

Very Likely

Likely

Possible

Ⓜ CUSTOMER_CITY

General Properties

Property	Value
TABLE_NAME	CUSMS
COLUMN_NAME	CUADC
DATA_TYPE_NAME	VARCHAR
SUPERSHEMA_FLAGS	0
NUMERIC_PRECISION	null
INTERNAL_NAME	T3C8
COLLATION_NAME	null
DESCRIPTION	null
KEYINFO	549
ORIGINAL_SCHEMA	CARD
COLLATION_CATALOG	null
CHARACTER_MAXIMUM_LENGTH	45
NATIVE_ROWBUFFER_OFFSET	271
COLUMN_GUID	null
DATA_TYPE	129



Data mapping results summary

Schema Na...	Table Name	Column Name	Native ...	EQ Dat...	Metadata Column	Attribute
Card	Account	AccountType	varchar	VARCHA...	ACCOUNT_NUMBER	
Card	Account	AvailableBalance	money	NUMERI...		
Card	Account	CardNumber	varchar	VARCHA...		
Card	Account	CustomerID	int	INT		
Card	Account	ExpiryDate	datetime	TIMESTA...		
Card	Account	ID	int identity	INT		PK RID_PRIMARY
Card	Account	Issuer	varchar	VARCHA...	ISSUER	
Card	Account	Limit	real	REAL	LIMIT	
Card	ACTTT	ACID	int	INT		
Card	ACTTT	C2657AMV	money	NUMERI...	AMOUNT	
Card	ACTTT	CUID	int	INT		
Card	ACTTT	DISP	bit	CHAR(1)		
Card	ACTTT	DPST	date	DATE		
Card	ACTTT	DRN	varchar	VARCHA...		
Card	ACTTT	ID	int identity	INT		PK RID_PRIMARY
Card	ACTTT	MEID	int	INT		
Card	ACTTT	MTD	varchar	VARCHA...		
Card	ACTTT	ST	varchar	VARCHA...		
Card	ACTTT	TRDD	datetime	TIMESTA...		
Card	CUSMS	CUADC	varchar	VARCHA...	CUSTOMER_CITY	
Card	CUSMS	CUADST	varchar	VARCHA...	CUSTOMER_STATE	
Card	CUSMS	CUADSTR	varchar	VARCHA...	CUSTOMER_STREET	
Card	CUSMS	CUADZC	varchar	VARCHA...	CUSTOMER_ZIP	
Card	CUSMS	CUDOB	datetime	TIMESTA...	CUSTOMER_DOB	
Card	CUSMS	CUEML	varchar	VARCHA...	CUSTOMER_EMAIL	
Card	CUSMS	CUFN	varchar	VARCHA...	CUSTOMER_FIRST...	
Card	CUSMS	CUG	varchar	VARCHA...	CUSTOMER_GEND...	
Card	CUSMS	CULN	varchar	VARCHA...	CUSTOMER_LAST...	
Card	CUSMS	CUPN	varchar	VARCHA...	CUSTOMER_PHONE	
Card	CUSMS	CUSSN	varchar	VARCHA...	CUSTOMER_SSN	
Card	CUSMS	ID	int identity	INT		PK RID_PRIMARY
Card	CUSMS	Latitude	float	FLOAT		
Card	CUSMS	Longitude	float	FLOAT		
Card	Merchant	Category	varchar	VARCHA...		
Card	Merchant	City	varchar	VARCHA...		
Card	Merchant	ID	int identity	INT		PK RID_PRIMARY
Card	Merchant	Latitude	float	FLOAT		
Card	Merchant	Longitude	float	FLOAT		
Card	Merchant	Merchant	varchar	VARCHA...		
Card	Merchant	PhoneNumber	varchar	VARCHA...		
Card	Merchant	State	varchar	VARCHA...		
Card	Merchant	Street	varchar	VARCHA...		
Card	Merchant	Street2	varchar	VARCHA...		
Card	Merchant	ZipCode	varchar	VARCHA...		



The End