# Automated SmartData Discovery and Classification (ASDAC) Demo

# January 2020

# ASDAC Demo Challenges and Solution

## Customer Challenges

| |
|---|
| Incomplete understanding of all data sources and data in the organization |
| Concerns for locations and associated risk of personal data for PCI, data theft risk, GDPR, etc. |
| Once discovered, need a means to secure personal data, but still allow it to be used for operations |

## SmartData Fabric® Solution

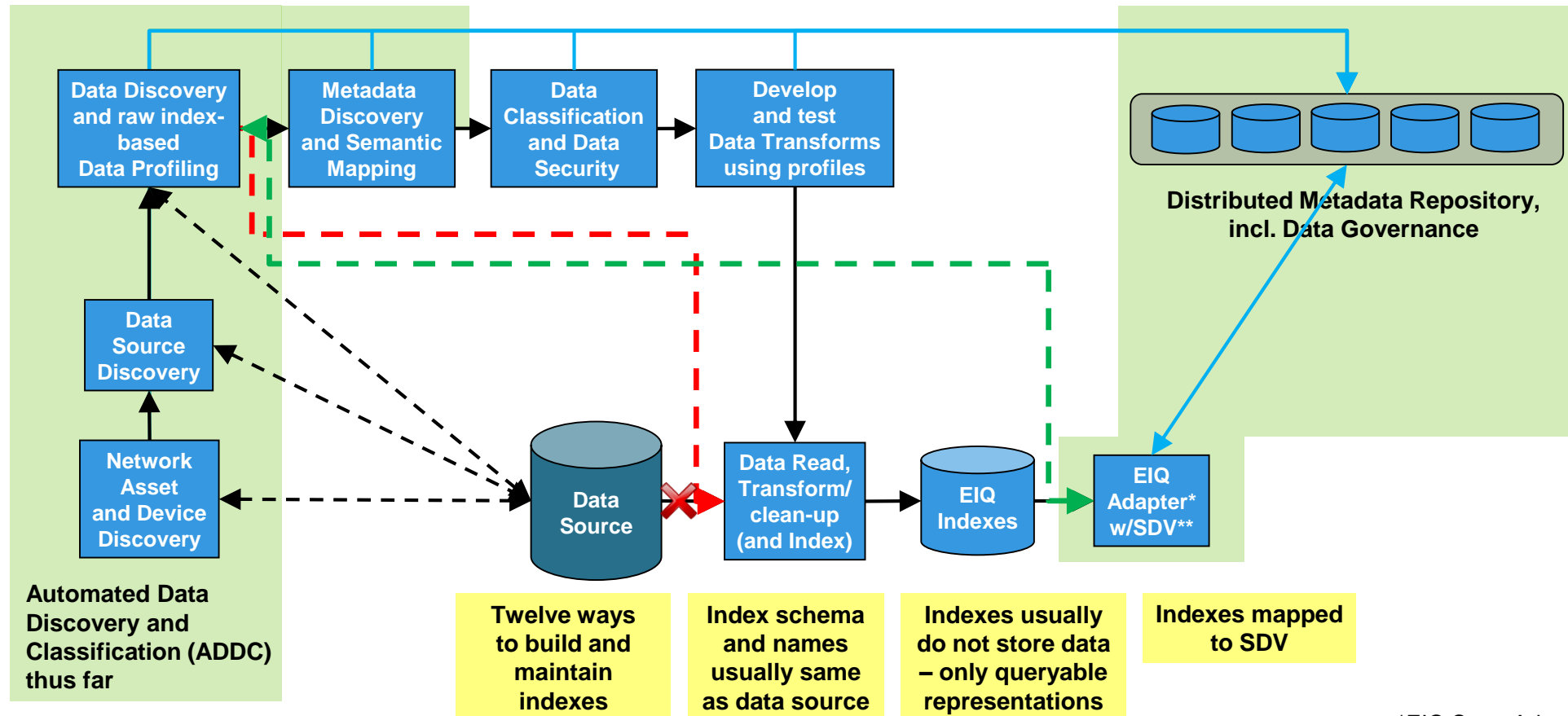| |
|---|
| Use popular network discovery tool, Open-AudIT (open source version, for now) |
| Use Open-AudIT REST API and SmartData Fabric® REST APIs to build workflows using BPM tool, Activiti (open source version, for now) |
| Discover devices and data sources |
| Build raw indexes on data sources with appropriate credentials – do not store data or pointers (for now) |
| Use data profiles generated from raw indexes to obtain metadata, semantically identify and classify data, and develop and test data transforms |
| Map indexes to a Standard Data View (SDV) |
| Handover to next stage production indexing |

# Initial EIQ Adapter configuration, index build and data view mapping

Data Discovery and raw index-based Data Profiling

Metadata Discovery and Semantic Mapping

Data Classification and Data Security

Develop and test Data Transforms using profiles

Distributed Metadata Repository, incl. Data Governance

Data Source Discovery

Network Asset and Device Discovery

**Automated Data Discovery and Classification (ADDC) thus far**

Data Source

Data Read, Transform/ clean-up (and Index)

EIQ Indexes

EIQ Adapter* w/SDV**

**Twelve ways to build and maintain indexes**

**Index schema and names usually same as data source**

**Indexes usually do not store data – only queryable representations**
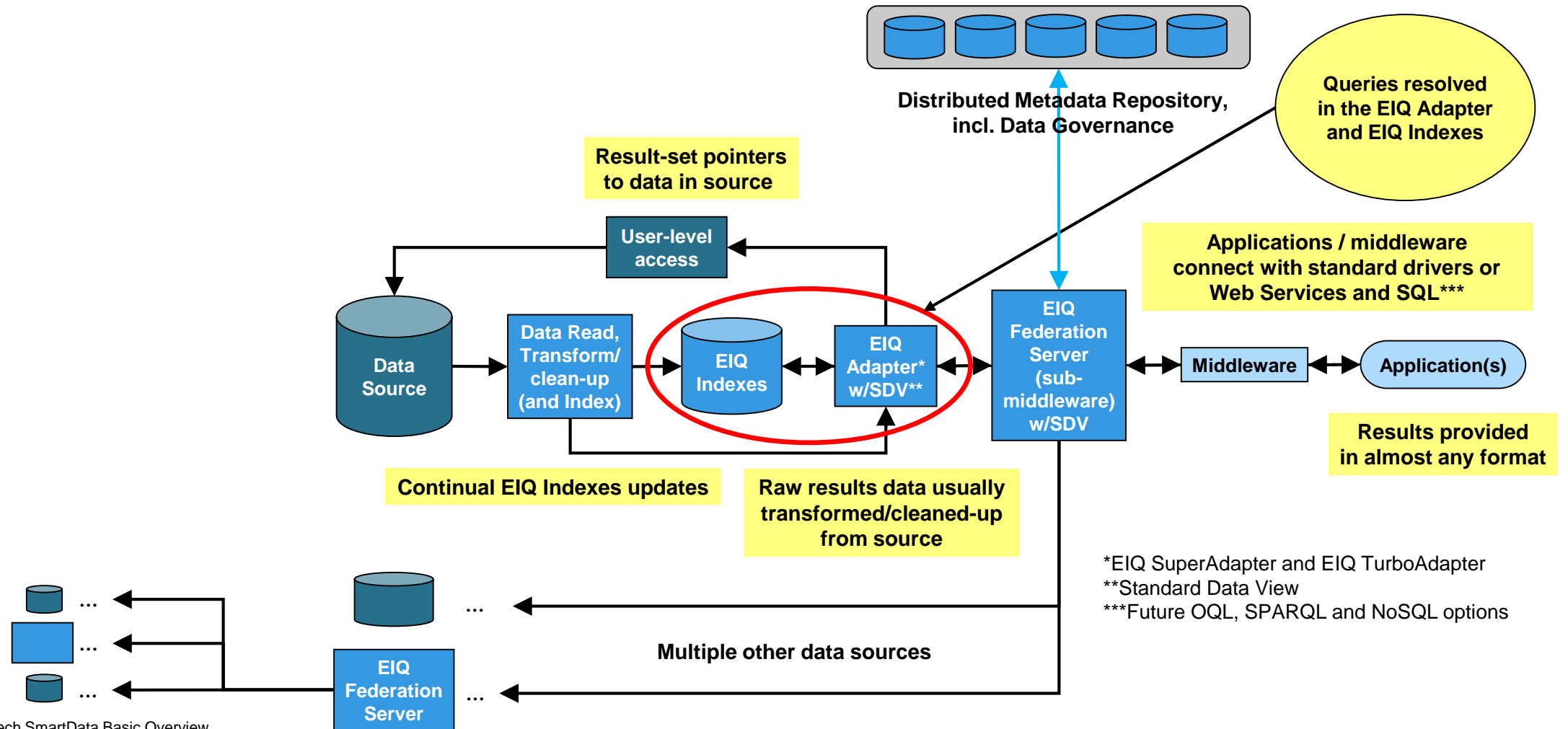
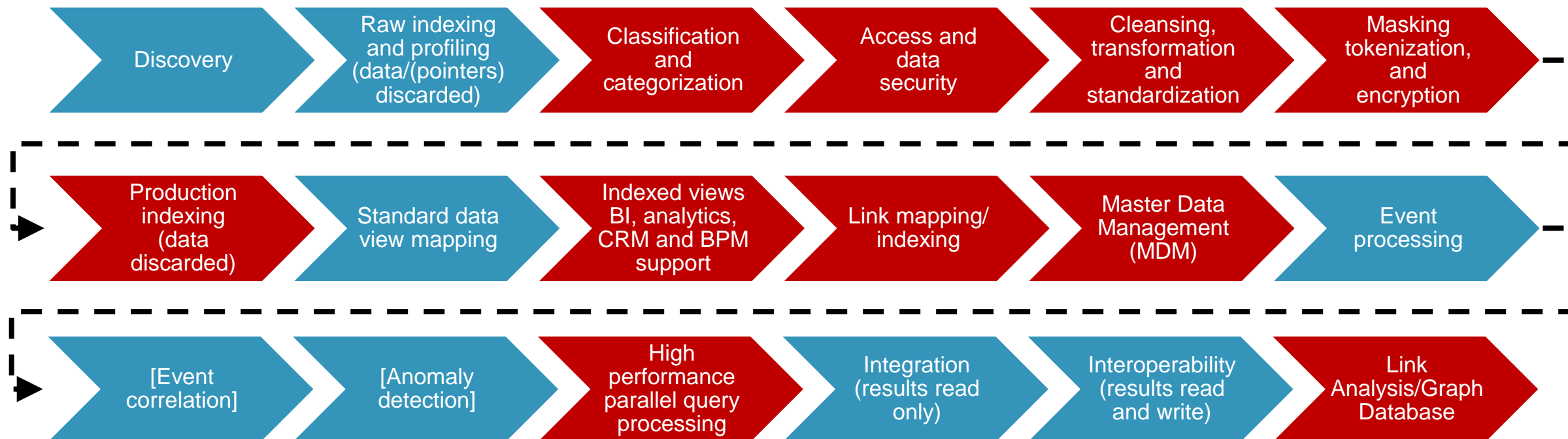**Indexes mapped to SDV**

*EIQ SuperAdapter and EIQ TurboAdapter
**Standard Data View

**Alternate use of raw indexes to initially build EIQ Indexes**
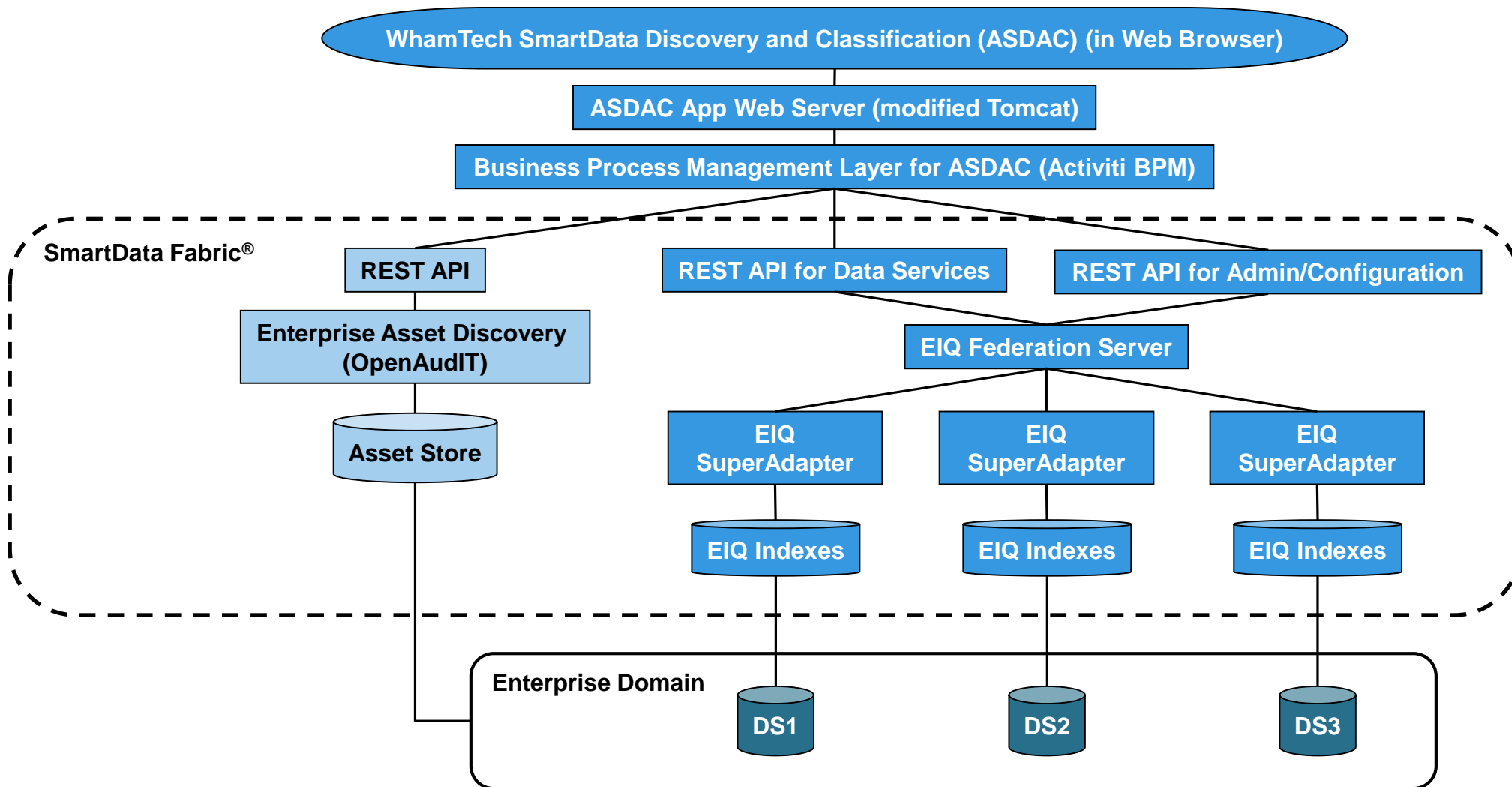
# EIQ Adapter index update, query and results retrieval



**Distributed Metadata Repository, incl. Data Governance**

**Queries resolved in the EIQ Adapter and EIQ Indexes**

**Result-set pointers to data in source**

**User-level access**

**Applications / middleware connect with standard drivers or Web Services and SQL\*\*\***

**Data Source**

**Data Read, Transform/ clean-up (and Index)**

**EIQ Indexes**

**EIQ Adapter\* w/SDV\*\***

**EIQ Federation Server (sub-middleware) w/SDV**

**Middleware**

**Application(s)**

**Results provided in almost any format**

**Continual EIQ Indexes updates**

**Raw results data usually transformed/cleaned-up from source**

\*EIQ SuperAdapter and EIQ TurboAdapter
\*\*Standard Data View
\*\*\*Future OQL, SPARQL and NoSQL options

**Multiple other data sources**

**EIQ Federation Server**

Source: #WhamTech SmartData Basic Overview

# SmartData Fabric® for data federation/abstraction

Discovery → Raw indexing and profiling (data/(pointers) discarded) → Classification and categorization → Access and data security → Cleansing, transformation and standardization → Masking tokenization, and encryption

Production indexing (data discarded) → Standard data view mapping → Indexed views BI, analytics, CRM and BPM support → Link mapping/ indexing → Master Data Management (MDM) → Event processing

[Event correlation] → [Anomaly detection] → High performance parallel query processing → Integration (results read only) → Interoperability (results read and write) → Link Analysis/Graph Database

WhamTech key differentiators addressed upfront

Source: #WhamTech SmartData Basic Overview

Revision 2.8

Copyright 2020 WhamTech, Inc.

5

# ASDAC architecture

WhamTech SmartData Discovery and Classification (ASDAC) (in Web Browser)

ASDAC App Web Server (modified Tomcat)

Business Process Management Layer for ASDAC (Activiti BPM)

**SmartData Fabric®**

REST API

REST API for Data Services

REST API for Admin/Configuration

Enterprise Asset Discovery (OpenAudIT)

EIQ Federation Server

Asset Store

EIQ SuperAdapter

EIQ SuperAdapter

EIQ SuperAdapter

EIQ Indexes

EIQ Indexes

EIQ Indexes

**Enterprise Domain**

DS1

DS2

DS3

# BPM workflow on top of REST APIs

# Screenshots of ASDAC in action

Launch the ASDAC app in a Web browser

# BPM workflow on top of REST APIs
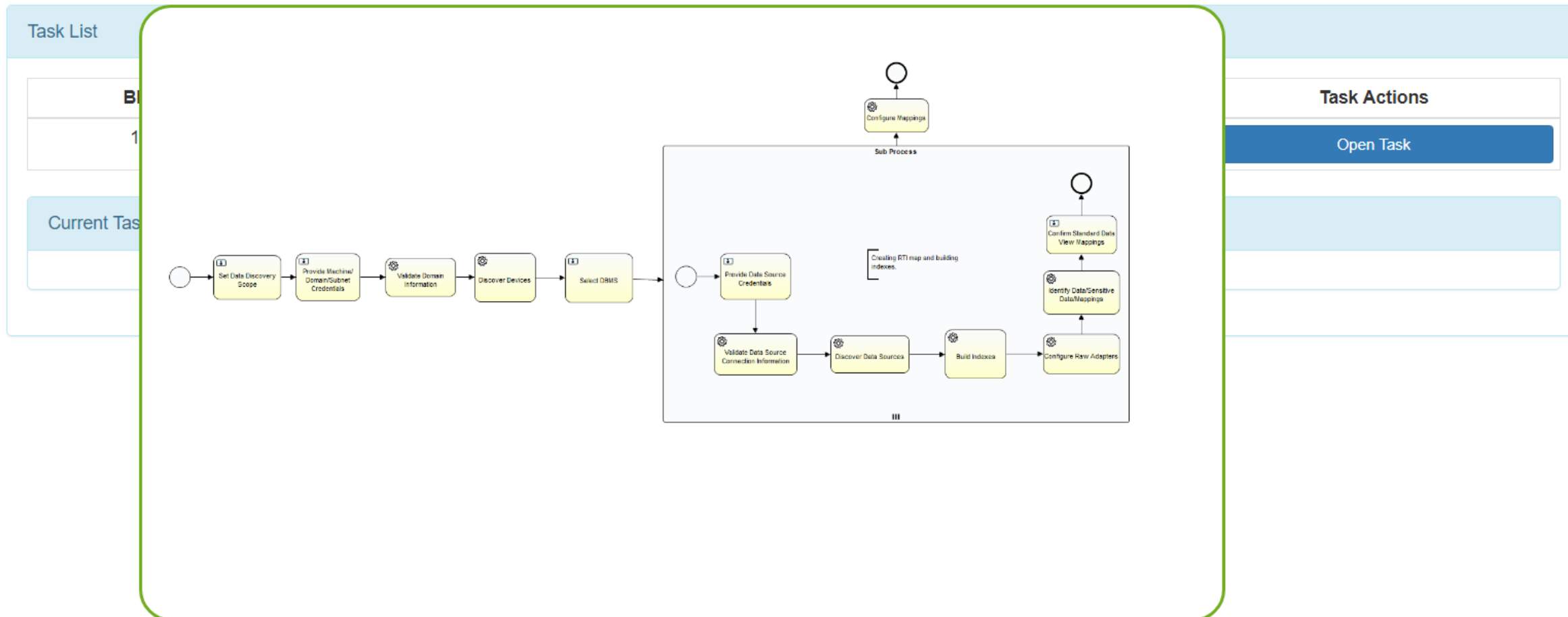
# Providing domain admin user credentials

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™**

Gavin

## 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-j9sqechju4s\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 |
| SUPERSCHEMA_FLAGS | 0 |
| NUMERIC_PRECISION | null |
| INTERNAL_NAME | T3C1 |
| COLLATION_NAME | null |

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

WHAMTECH®

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

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 |

Percentage
0    10    20    30    40    50    60    70    80    90    100
84.5100  8.333333333333334
45.7700  6.666666666666667
79.6300  6.666666666666667
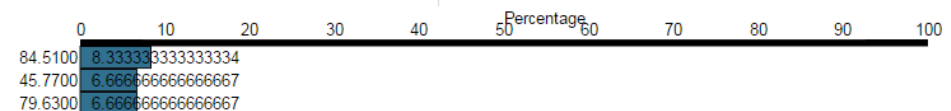
# Display the value distribution for the selected column (2 of 2)
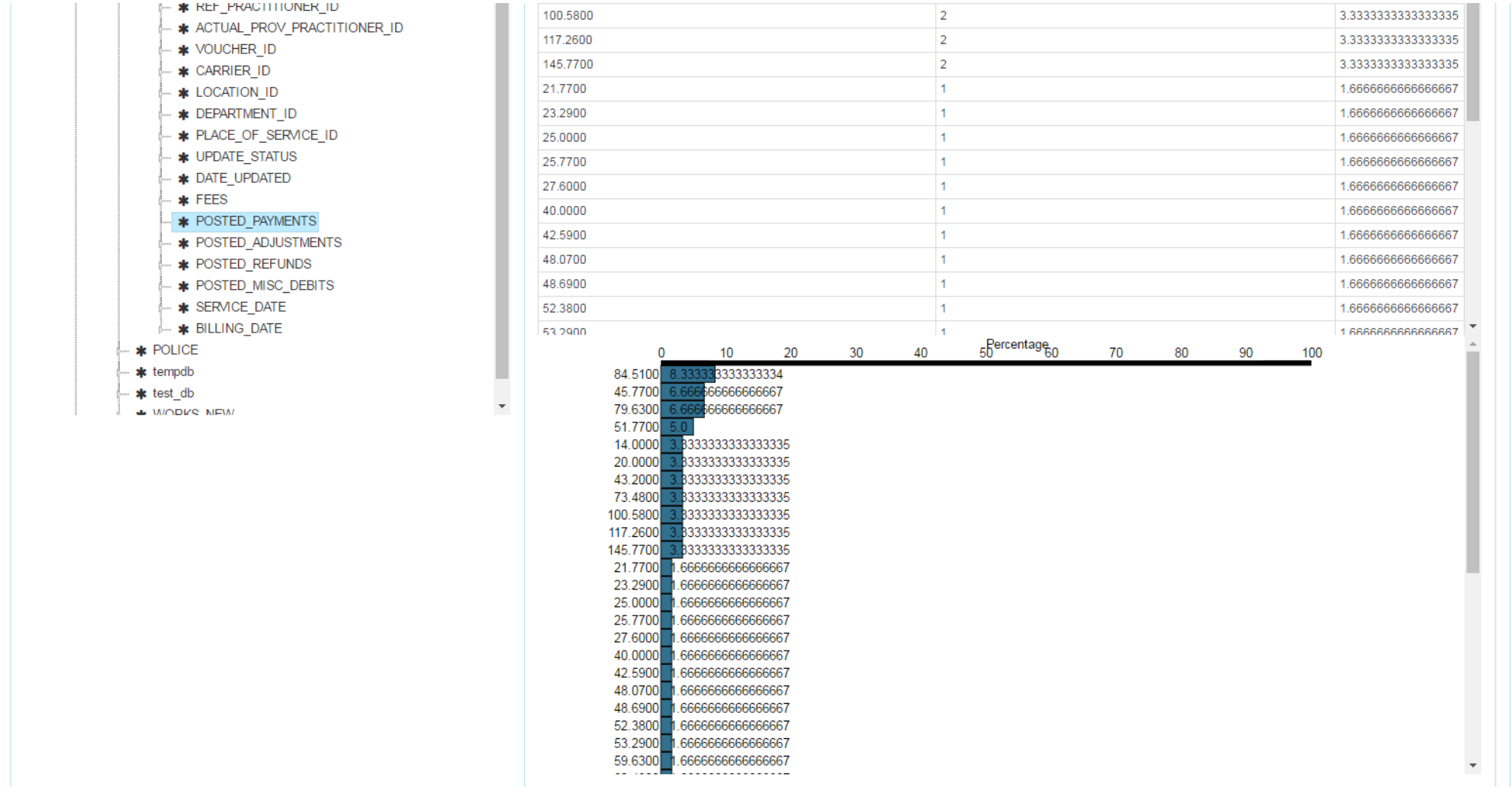
# Display the most and least frequent values for the selected column

Automated SmartData Discovery and Classification

Gavin

Repository Explorer

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

Toggle Process Diagram

Clear Data Discovery Repository

- ＊ 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 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

# Green flag for very likely matches and no conflicts

# Data mapping results summary

| Schema Na... | Table Name | Column Name | Native ... | EIQ 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