June 8, 2021

Data Governance FAQ’s – Axon, EDC, IDQ, DPM

Steven Fleishman, Principal Consultant
Sumit Saraswat, Solution Architect
Informatica Professional Services

Frequently Asked Questions from the Field

- What does CLAIRE exactly do ?
  - Are we leveraging it or how can we leverage it better ?
  - Does it learn from human actions of curation etc ?
- I see profiling in EDC, IDQ, and DPM. Are they the same? Which one is leveraged when?
- Where does AI/ML come into the picture in the Informatica DG Solution What does it do exactly?
- How can we make the information in Axon actionable ?
- What is the approach to classify information easily and efficiently in Axon ?
  - Similarly, how can I classify information in EDC if I do not have DPM?
- How can I easily generate a 360 graphical view of related and impacted assets in Axon?
- How to segment information based on LOBs/departments in Axon ?
  - Can I have a common/shared repository of assets and a department specific one.
  - Can I have local/private change management processes for my specific group ?
- How can I record and expose data dictionaries in the tools ?
- What are the best practices of scanning and cataloging now widely adopted solutions such as the data lake on S3 or Azure ?

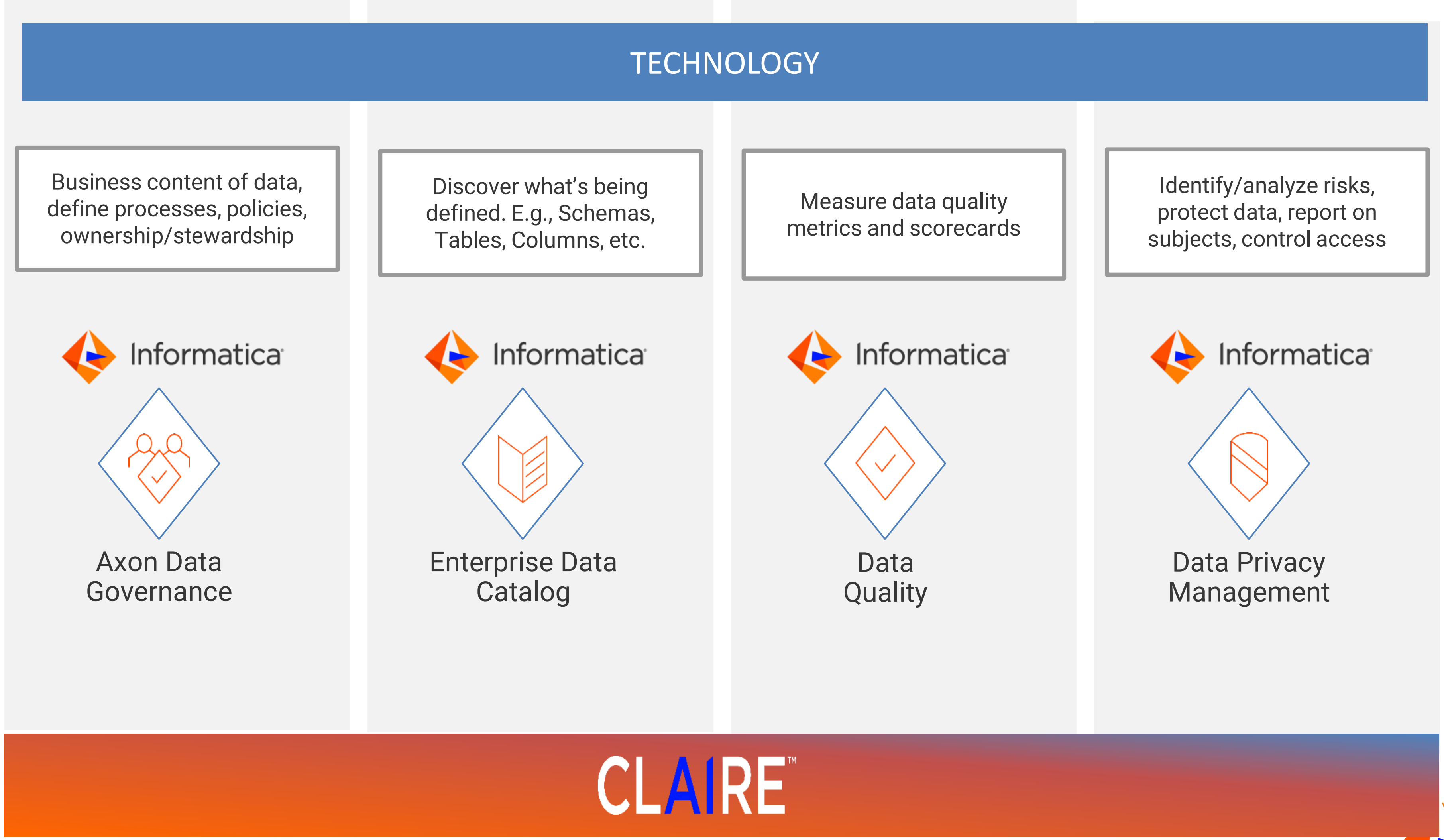
- What does CLAIRE exactly do ?
  - Are we leveraging it or how can we leverage it better ?
  - Does it learn from human actions of curation etc ?
- I see profiling in EDC, IDQ, and DPM. Are they the same? Which one is leveraged when?
- Where does AI/ML come into the picture in the Informatica DG Solution What does it do exactly?

Agenda

- Product Integration, Navigation, and Terminology
- Domains and Data Domain Types
- AI/ML - Claire
- Profiling functionality – EDC compared to IDQ
- Q&A

Product Integration, Navigation, and Terminology

Data Governance and Privacy Solutions



TECHNOLOGY

Business content of data, define processes, policies, ownership/stewardship

Informatica

Axon Data Governance

Discover what's being defined. E.g., Schemas, Tables, Columns, etc.

Informatica

Enterprise Data Catalog

Measure data quality metrics and scorecards

Informatica

Data Quality

Identify/analyze risks, protect data, report on subjects, control access

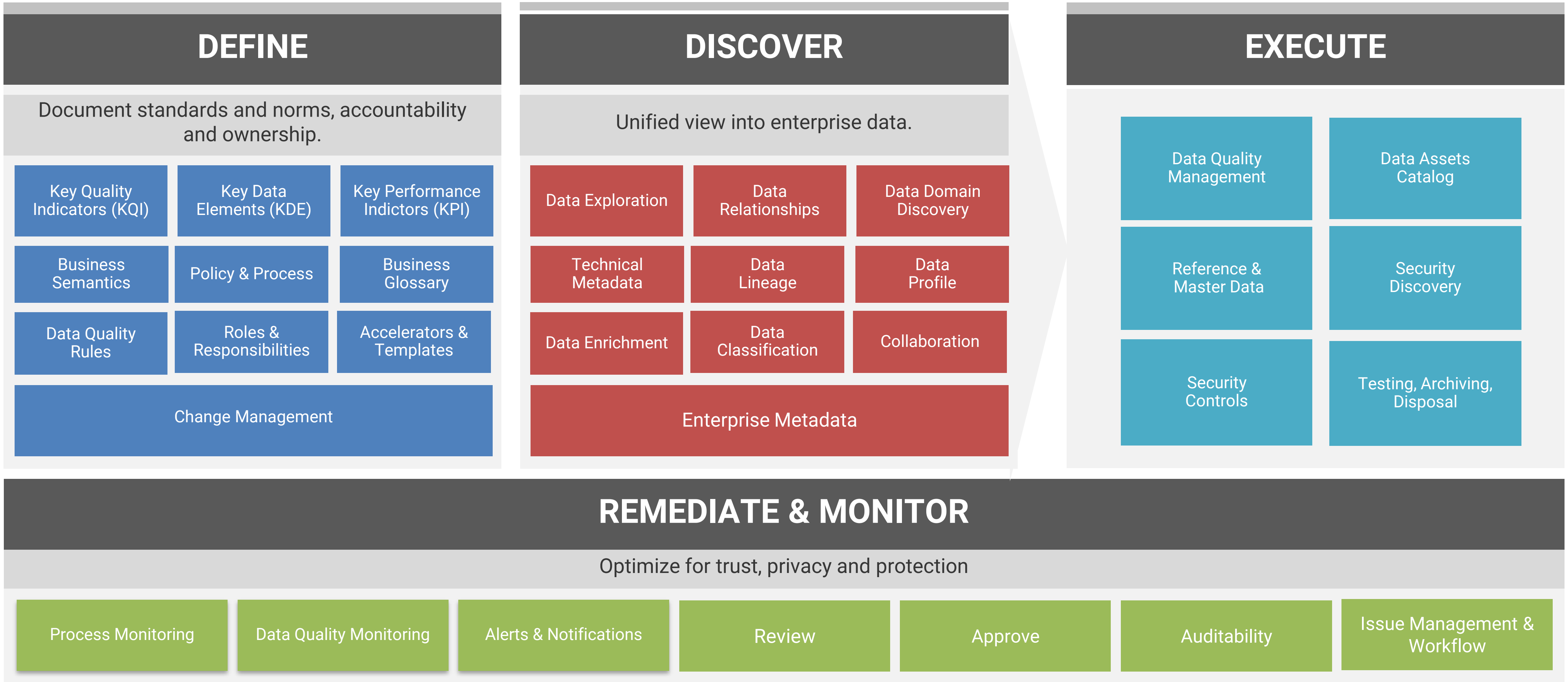
Informatica

Data Privacy Management

CLAIRE™

Informatica™

Architecture to Support Data Governance Framework



DEFINE

Document standards and norms, accountability and ownership.

- Key Quality Indicators (KQI)
- Key Data Elements (KDE)
- Key Performance Indictors (KPI)
- Business Semantics
- Policy & Process
- Business Glossary
- Data Quality Rules
- Roles & Responsibilities
- Accelerators & Templates
- Change Management

DISCOVER

Unified view into enterprise data.

- Data Exploration
- Data Relationships
- Data Domain Discovery
- Technical Metadata
- Data Lineage
- Data Profile
- Data Enrichment
- Data Classification
- Collaboration
- Enterprise Metadata

EXECUTE

- Data Quality Management
- Data Assets Catalog
- Reference & Master Data
- Security Discovery
- Security Controls
- Testing, Archiving, Disposal

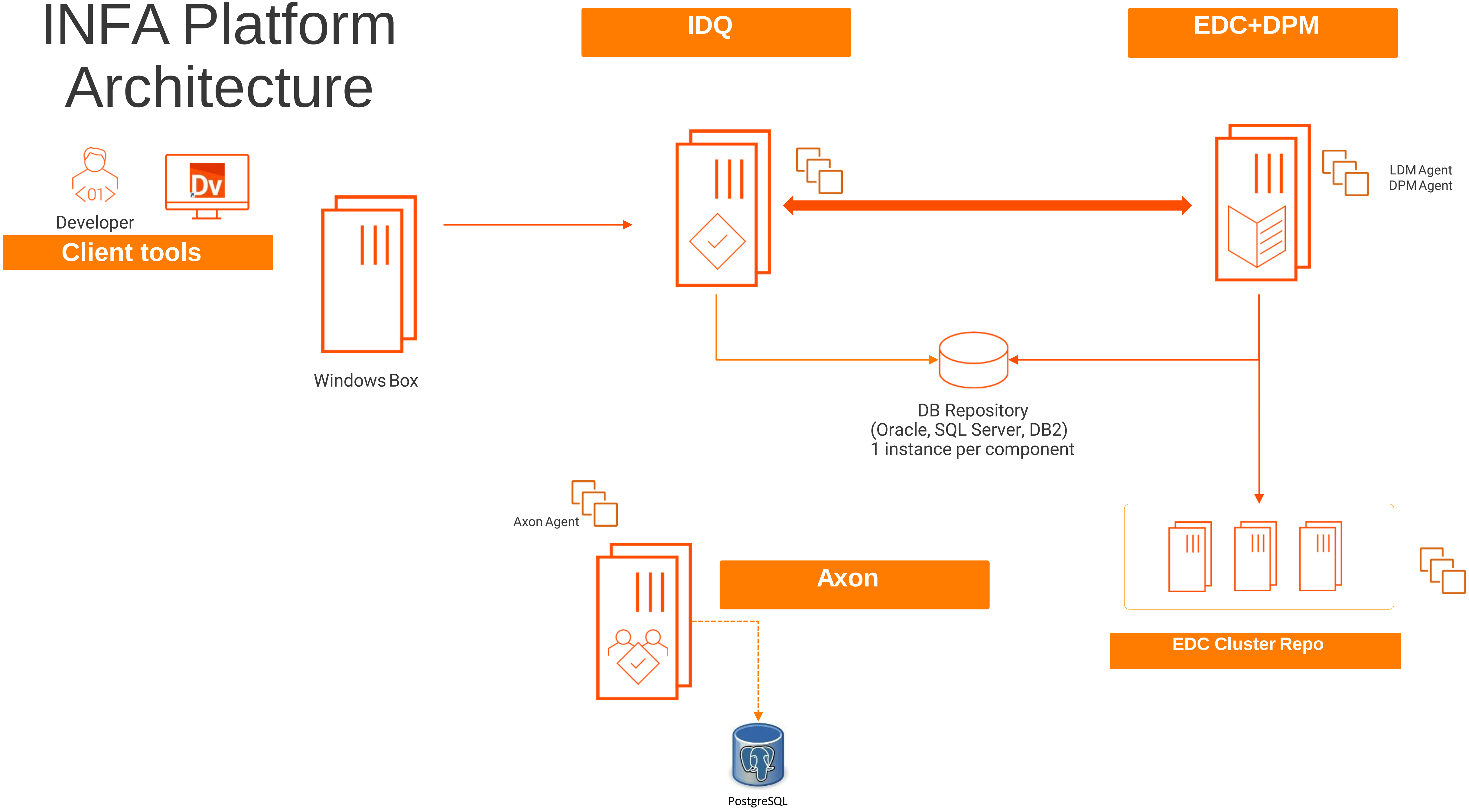
REMEDIATE & MONITOR

Optimize for trust, privacy and protection

- Process Monitoring
- Data Quality Monitoring
- Alerts & Notifications
- Review
- Approve
- Auditability
- Issue Management & Workflow

Informatica

INFA Platform Architecture



IDQ

EDC+DPM

Developer

Client tools

Windows Box

LDM Agent
DPM Agent

DB Repository
(Oracle, SQL Server, DB2)
1 instance per component

Axon Agent

Axon

EDC Cluster Repo

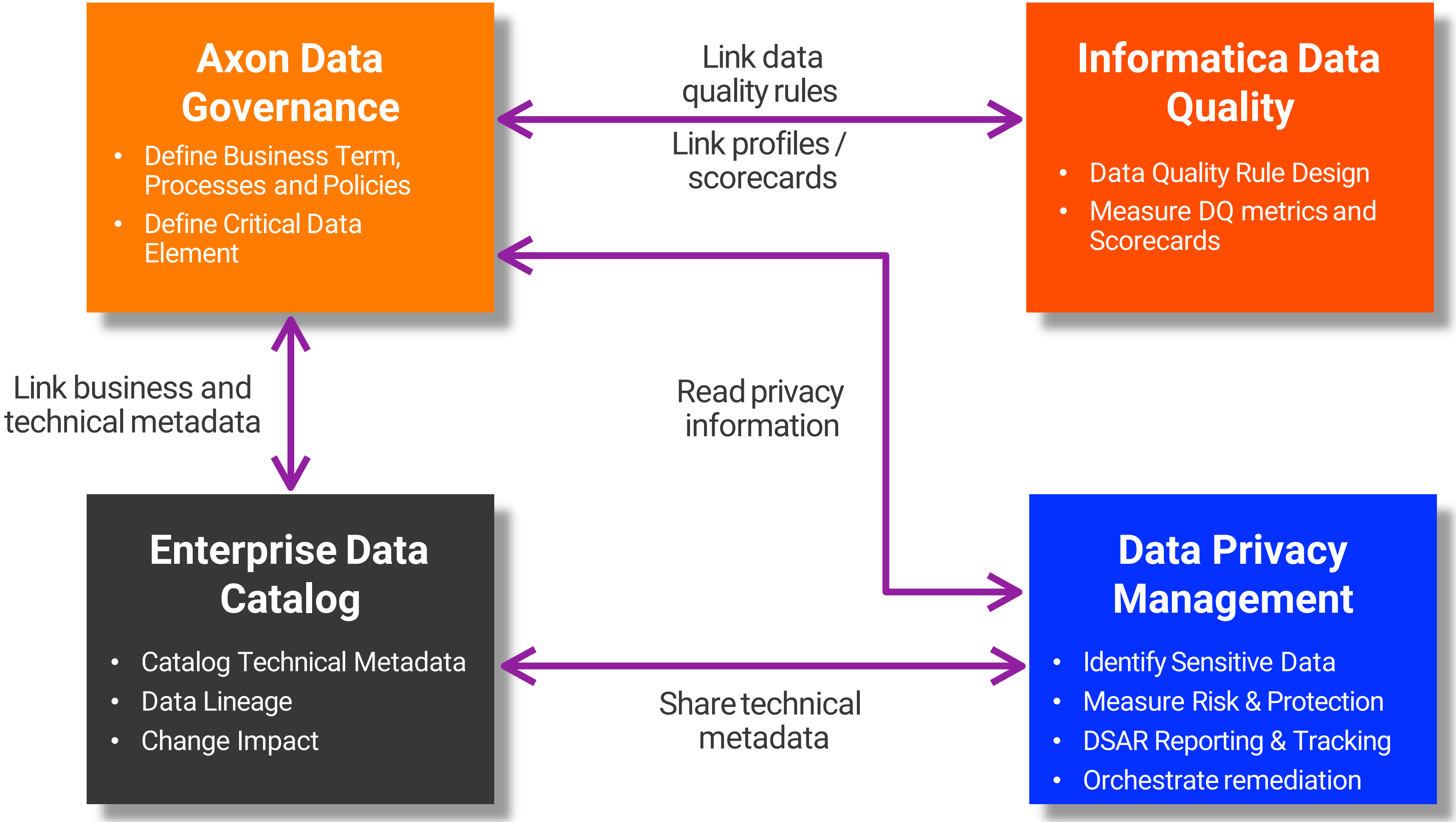
PostgreSQL

INFA Split Domain: EDC and IDQ

Recommendation and Best Practice for EDC and/or DPM and IDQ to be installed in separate Domain, here are pointers:

- Flexibility applying patches, fixes, upgrades for respective product
- IDQ is higher volume (longer running jobs-less jobs-more operational driven)
- EDC is Metadata (more jobs-less operational driven)
- IDQ licensing is based on number of cores in the machine, whereas EDC licensing is based on number of Resources
- Profiling: Context of Profiling in EDC is for Data Domain Discovery, Similarity Discovery, Unique Key Inference, CLAIRE on larger set of data, however context of Profiling on IDQ is to perform checks on Data Quality Rules, Scorecards focused on key sets of data.

Cross Product View



Axon Data Governance

- Define Business Term, Processes and Policies
- Define Critical Data Element

Informatica Data Quality

- Data Quality Rule Design
- Measure DQ metrics and Scorecards

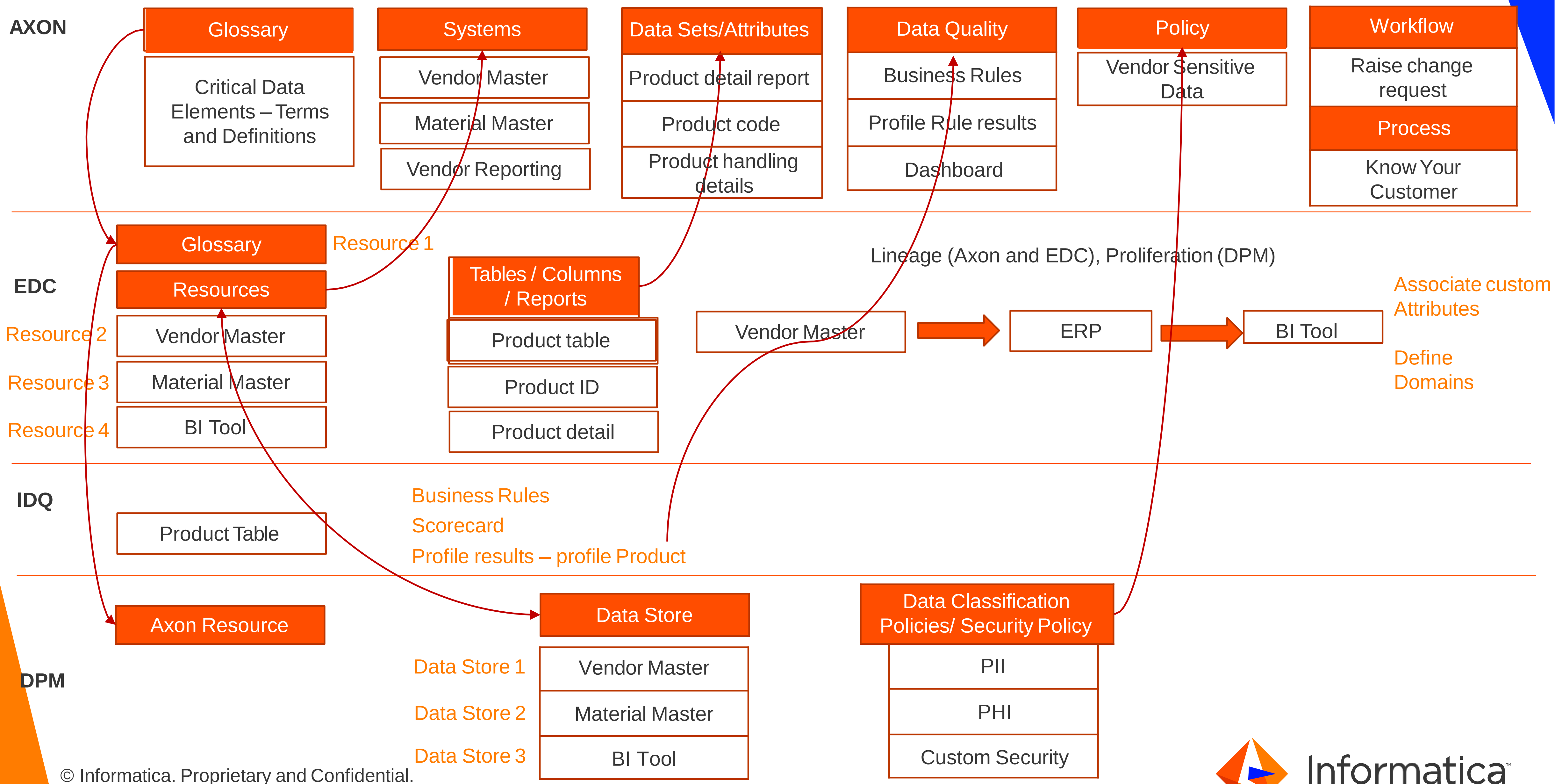
Enterprise Data Catalog

- Catalog Technical Metadata
- Data Lineage
- Change Impact

Data Privacy Management

- Identify Sensitive Data
- Measure Risk & Protection
- DSAR Reporting & Tracking
- Orchestrate remediation

Data Governance Application Relationships



AXON

Glossary

Critical Data Elements – Terms and Definitions

Systems

Vendor Master

Material Master

Vendor Reporting

Data Sets/Attributes

Product detail report

Product code

Product handling details

Data Quality

Business Rules

Profile Rule results

Dashboard

Policy

Vendor Sensitive Data

Workflow

Raise change request

Process

Know Your Customer

EDC

Glossary

Resource 1

Resources

Resource 2

Vendor Master

Resource 3

Material Master

Resource 4

BI Tool

Tables / Columns / Reports

Product table

Product ID

Product detail

Lineage (Axon and EDC), Proliferation (DPM)

Vendor Master

ERP

BI Tool

Associate custom Attributes

Define Domains

IDQ

Product Table

Business Rules

Scorecard

Profile results – profile Product

DPM

Axon Resource

Data Store

Data Store 1

Vendor Master

Data Store 2

Material Master

Data Store 3

BI Tool

Data Classification Policies/ Security Policy

PII

PHI

Custom Security

Informatica

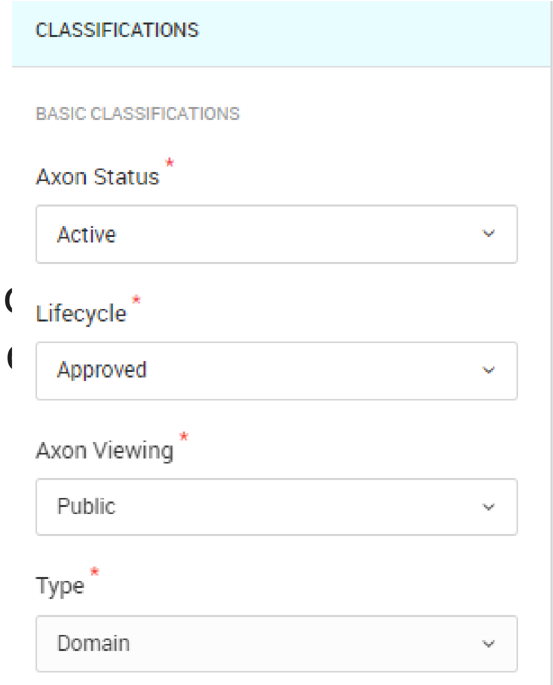
Data Domains and Domain Types

“Domain” Usage in Data Governance and Privacy

- Informatica Domain
  - A collection of nodes and services that define the Informatica platform. You group nodes and services in a domain based on administration ownership

- Axon Domain
  - A glossary type, that’s a way of classifying data
  - Describes a broad category of data concepts, for example, customer domain or transaction data domain
  - Specific to Axon and can be modified

- Data Domain
  - Predefined or user-defined Model repository (
  - Based on the semantics of column data or a (



CLASSIFICATIONS

BASIC CLASSIFICATIONS

Axon Status *: Active

Lifecycle *: Approved

Axon Viewing *: Public

Type *: Domain

Informatica

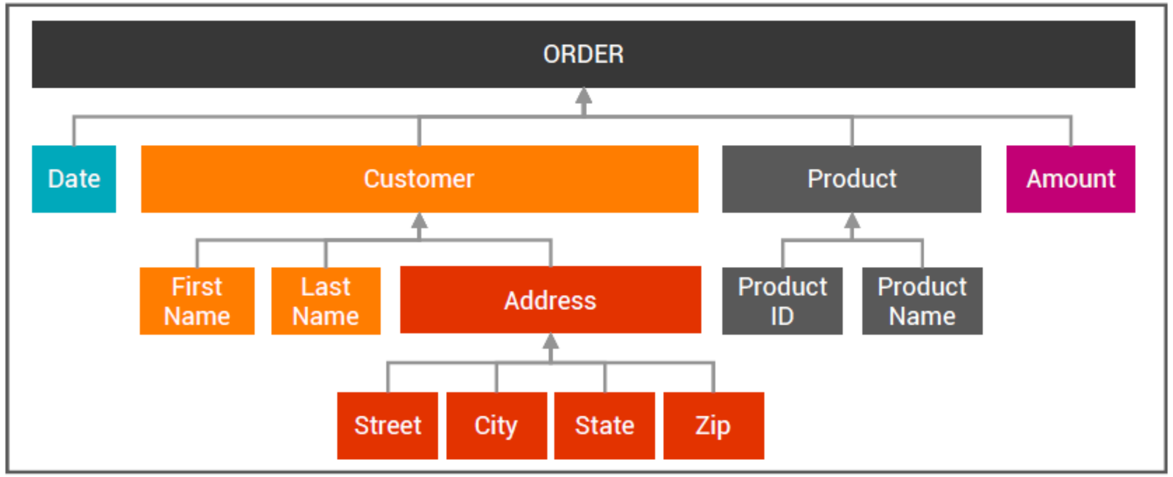
Types of Data Domains

- Rule-Based
  - Run against Metadata, Data or Both
  - 125+ predefined data domains
  - Regex - pattern
    - credit card, SSN, phone number
  - Reference – finite, non-overlapping
    - ISO country code, currency codes
  - Mapplet – Leverage Informatica Developer and Analyst for complex rules

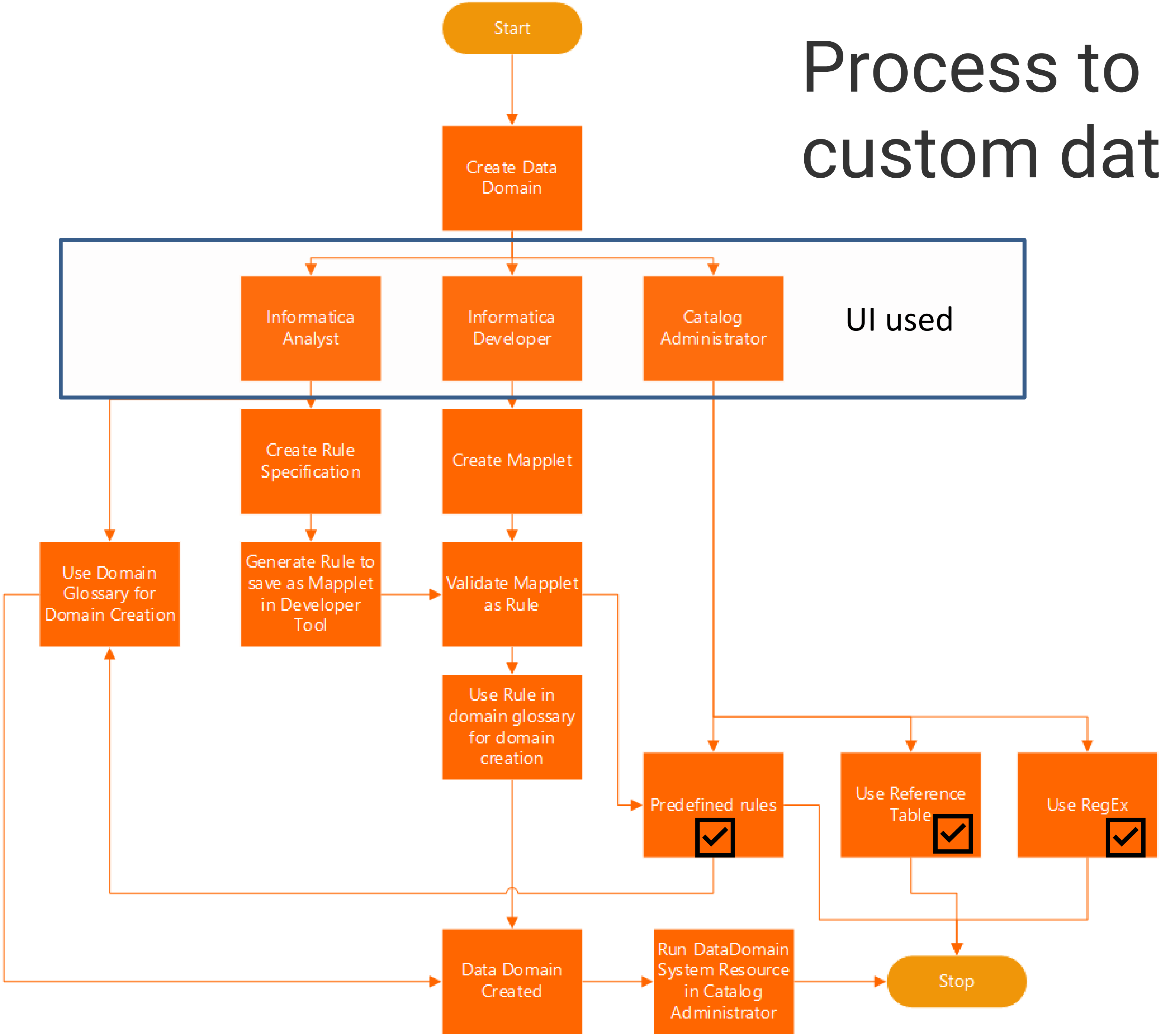
Types of Data Domains Continued

- Smart – Specific to EDC
  - Example based data domain
  - Data tagging and propagation
- Composite Data Domain
- Data Domain Group

- Collection of data domains or other composite data domains linked using rules
- Enables you to search for the required details of an entity across multiple schemas defined for the database



Process to create custom data domains



Start
Create Data Domain
Informatica Analyst
Informatica Developer
Catalog Administrator
UI used
Create Rule Specification
Create Mapplet
Use Domain Glossary for Domain Creation
Generate Rule to save as Mapplet in Developer Tool
Validate Mapplet as Rule
Use Rule in domain glossary for domain creation
Predefined rules ☑
Use Reference Table ☑
Use RegEx ☑
Data Domain Created
Run DataDomain System Resource in Catalog Administrator
Stop

Informatica™

Out of the Box Data Domains

- The following data domains may create large number of false positives; Use with caution
  - Age
  - Salary
  - Weight
  - Height
  - Alphanumeric_specialCharacters
  - Date_allFormats
  - Admission_dates
  - JobPosition
  - Binary Value
  - Admission_date

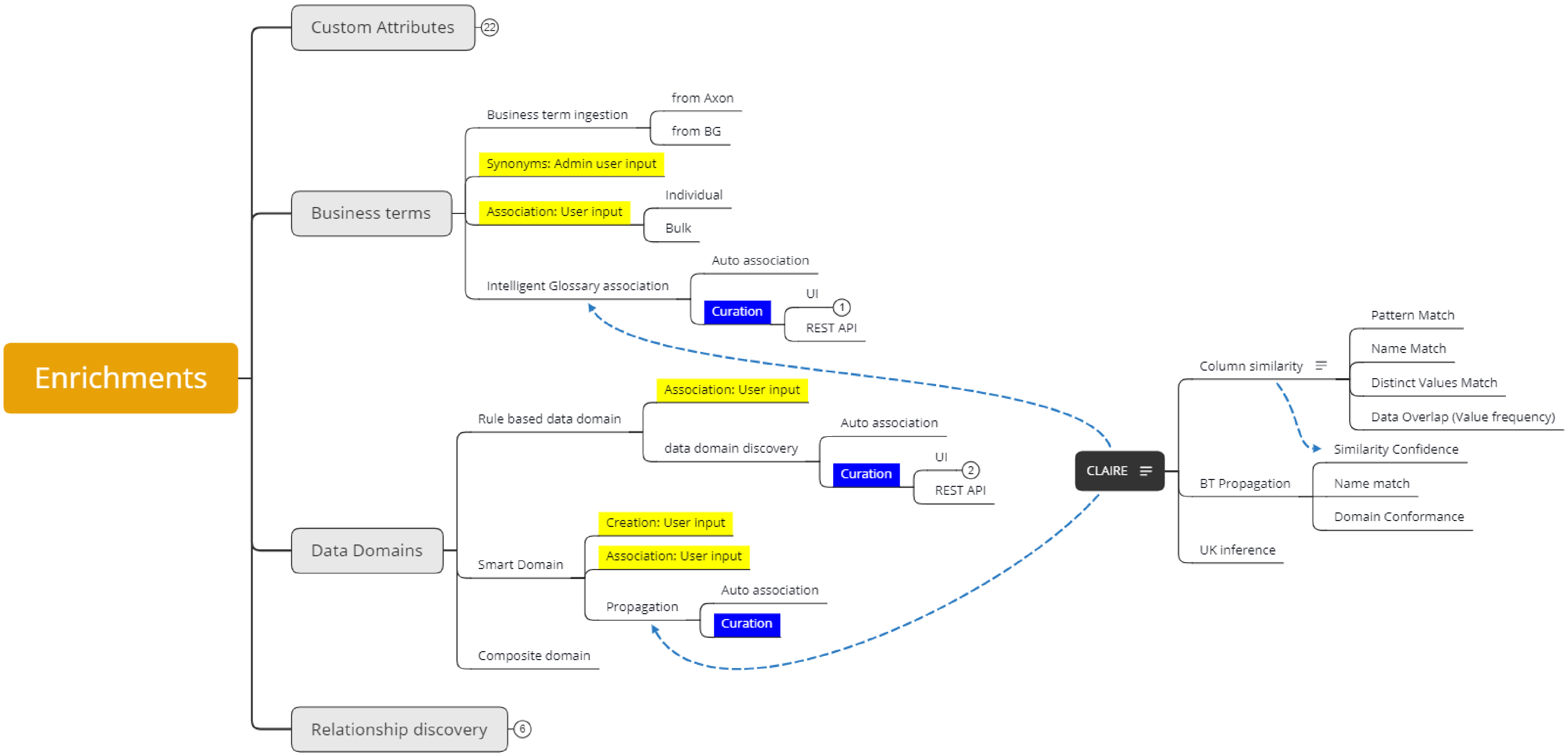
- Avoid using “All” data domains

- Make a copy of the original data domain before modifying

AI / ML - Claire

CLAIRE

- CLAIRE stands for Cloud-Scale AI powered Real-Time Engine.
- Identifies all capabilities in Informatica products and services that use artificial intelligence (AI) and machine-learning techniques on enterprise-wide data and metadata to significantly boosts the productivity and experience of users of our technology.
- The only real way to discover velocity and diversity of data manage this complexity is to increase automation and to significantly improve the productivity and effectiveness of the data management staff.
- This is where artificial intelligence and machine learning come in.



Enrichments

- Custom Attributes 22
- Business terms
  - Business term ingestion
    - from Axon
    - from BG
  - Synonyms: Admin user input
  - Association: User input
    - Individual
    - Bulk
  - Intelligent Glossary association
    - Auto association
    - Curation
      - UI 1
      - REST API
- Data Domains
  - Rule based data domain
    - Association: User input
    - data domain discovery
      - Auto association
      - Curation
        - UI 2
        - REST API
  - Smart Domain
    - Creation: User input
    - Association: User input
    - Propagation
      - Auto association
      - Curation
  - Composite domain
- Relationship discovery 6

CLAIRE

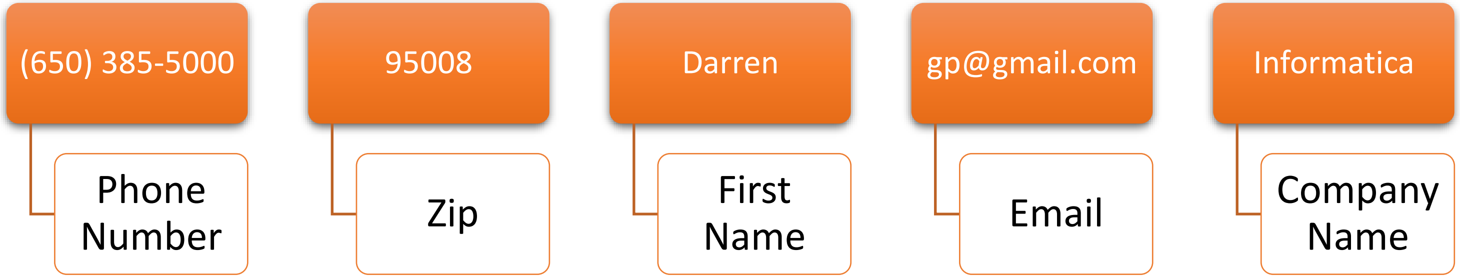
- Column similarity
  - Pattern Match
  - Name Match
  - Distinct Values Match
  - Data Overlap (Value frequency)
- BT Propagation
  - Similarity Confidence
  - Name match
  - Domain Conformance
- UK inference

Smart Data Domains

Process of discovering semantic meaning of data in the data sources

Smart domains

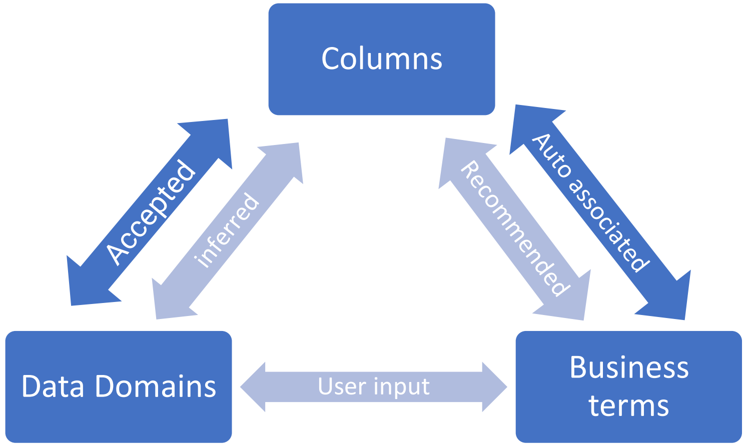
- Act as tags
- Learn by example and propagated by looking at column similarity.
- Exist as an object in the catalog and can be enriched as well.
- Requires access to the data



Column similarity

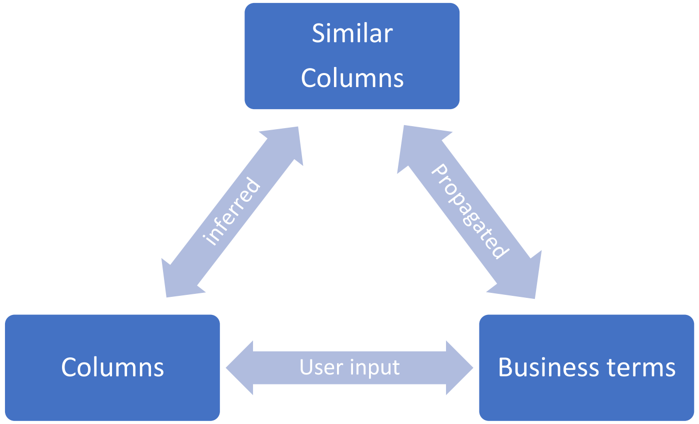
- Identify clusters of columns that contain similar data within and across data sources.
- **Use:**
- Identifying data
- Detecting duplicates
- Combining individual data fields into business entities
- Propagating tags across data sets
- Recommending data sets to users

Business term association through propagation



- When data domains are inferred against specific columns, the associated glossary terms are recommended for those columns.
- When data domains are accepted, associated glossary terms are also associated to the columns

- System propagates business glossary terms to similar column
- Similarity based on name match, unique value match and data match is used for business glossary propagation

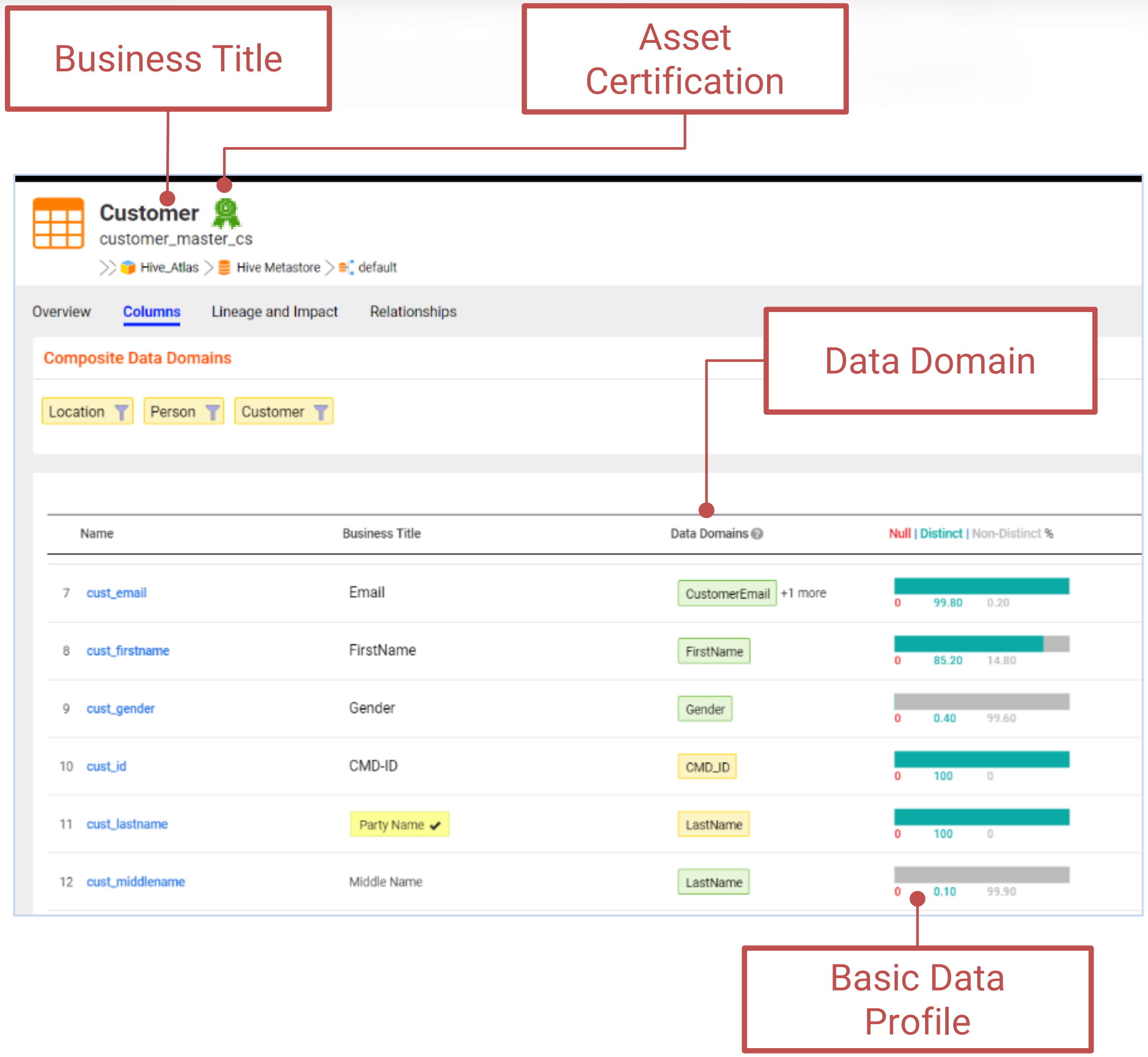


Business term association through Claire Match

- Match English phrases with technical names using sequence alignment
  - **Sequence Alignment / Delete-only Edit Distance :** The business term names that align well with asset names are sought. This approach can capture obvious abbreviations of business terms.
    - HEALTH PROGRAM CONSULTATION (Business Term Title)
    - H- - LTH P - - G - - M C- NS - LT- T- -N (Asset name)
  - **Synonym dictionary :** If available, user provides a dictionary of commonly used synonyms/abbreviations in technical asset names within the organization. This dictionary is used to improve glossary matching
- Additionally, prefix ignore options for discarding common technical prefixes(like TBL, VW etc for better matches)

How does profiling differ between IDQ and EDC?

EDC – Broad Profiling Results – Table View



Business Title

Asset Certification

Data Domain

Basic Data Profile

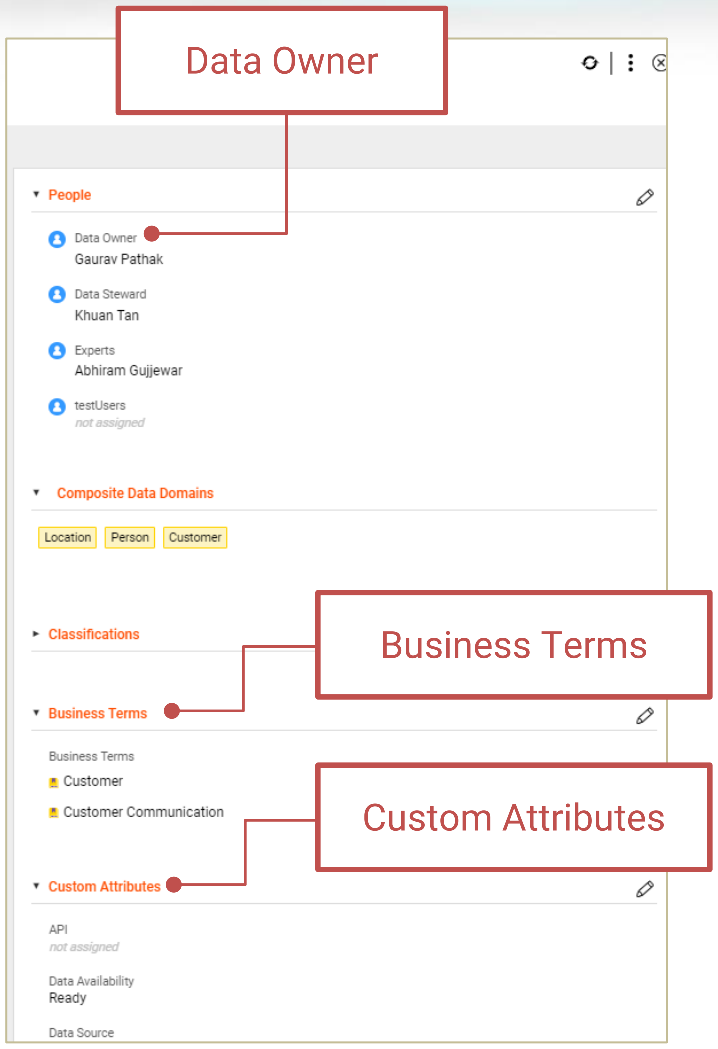
Customer
customer_master_cs
Hive_Atlas > Hive Metastore > default

Overview | Columns | Lineage and Impact | Relationships

Composite Data Domains

Location | Person | Customer

	Name	Business Title	Data Domains	Null | Distinct | Non-Distinct %
7	cust_email	Email	CustomerEmail +1 more	0 99.80 0.20
8	cust_firstname	FirstName	FirstName	0 85.20 14.80
9	cust_gender	Gender	Gender	0 0.40 99.60
10	cust_id	CMD-ID	CMD_ID	0 100 0
11	cust_lastname	Party Name ✓	LastName	0 100 0
12	cust_middlename	Middle Name	LastName	0 0.10 99.90



Data Owner

Business Terms

Custom Attributes

People

Data Owner
Gaurav Pathak

Data Steward
Khuan Tan

Experts
Abhiram Gujjewar

testUsers
not assigned

Composite Data Domains

Location | Person | Customer

Classifications

Business Terms

Business Terms
Customer
Customer Communication

Custom Attributes

API
not assigned

Data Availability
Ready

Data Source

EDC – Broad Profiling Results – Column View

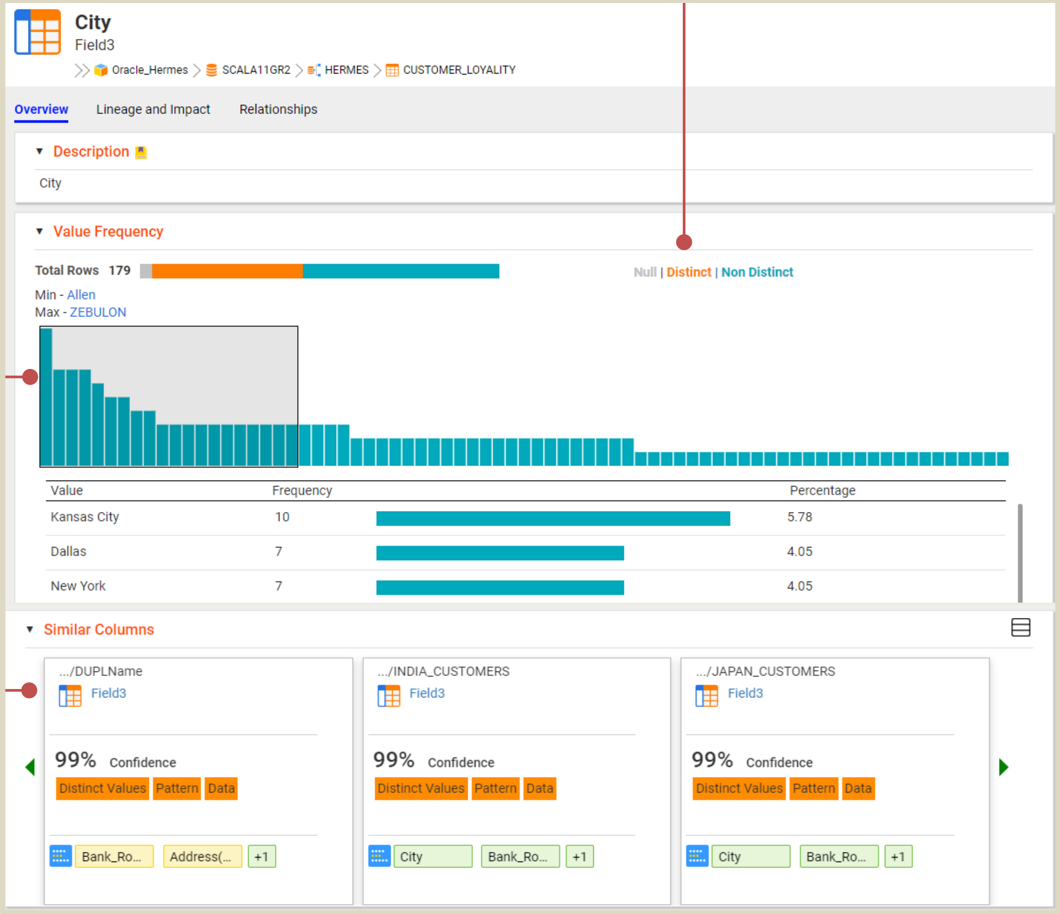
Basic Data Profile

Pattern Distribution

Data Domain

Value Frequencies

Column Similarity



City
Field3
>> Oracle_Hermes > SCALA11GR2 > HERMES > CUSTOMER_LOYALITY

Overview | Lineage and Impact | Relationships

▾ Description

City

▾ Value Frequency

Total Rows 179

Null | Distinct | Non Distinct

Min - Allen
Max - ZEBULON

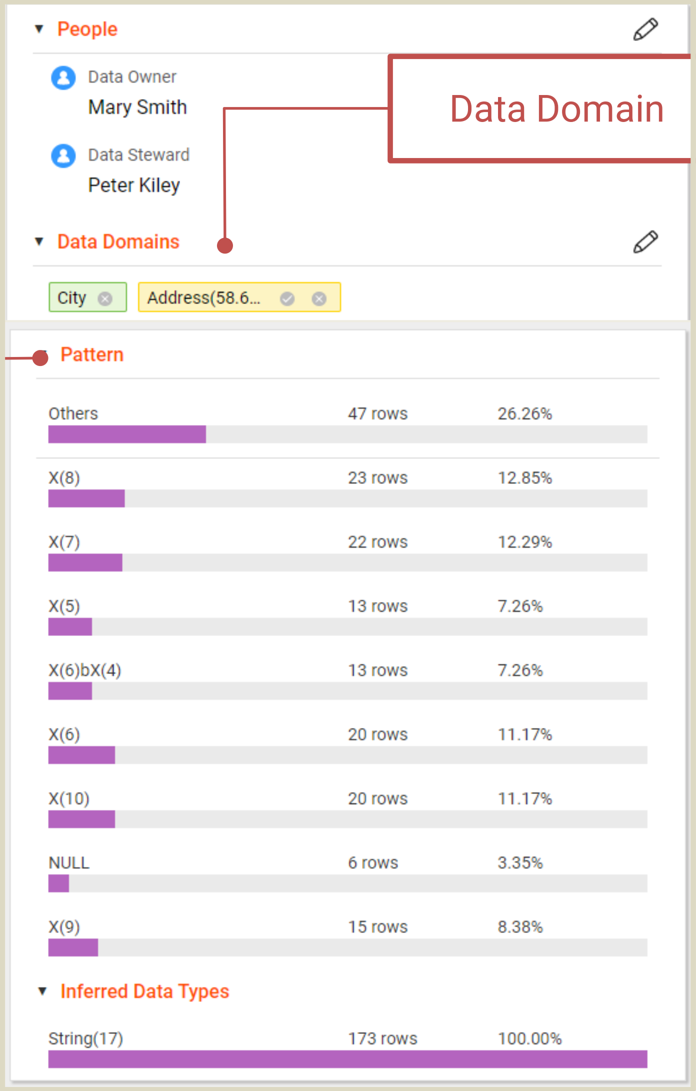
Value	Frequency	Percentage
Kansas City	10	5.78
Dallas	7	4.05
New York	7	4.05

▾ Similar Columns

.../DUPLName
Field3
99% Confidence
Distinct Values | Pattern | Data
Bank_Ro... | Address(... | +1

.../INDIA_CUSTOMERS
Field3
99% Confidence
Distinct Values | Pattern | Data
City | Bank_Ro... | +1

.../JAPAN_CUSTOMERS
Field3
99% Confidence
Distinct Values | Pattern | Data
City | Bank_Ro... | +1



▾ People

Data Owner
Mary Smith

Data Steward
Peter Kiley

▾ Data Domains

City | Address(58.6...

Pattern

Others	47 rows	26.26%
X(8)	23 rows	12.85%
X(7)	22 rows	12.29%
X(5)	13 rows	7.26%
X(6)bX(4)	13 rows	7.26%
X(6)	20 rows	11.17%
X(10)	20 rows	11.17%
NULL	6 rows	3.35%
X(9)	15 rows	8.38%

▾ Inferred Data Types

String(17)	173 rows	100.00%

Informatica™

IDQ – Broad **and** Deep Enterprise-Grade Data Management Solution

IDQ

Discovery, search & profiling

Role-based capabilities
Enable business users to build and test logical business rules without relying on IT

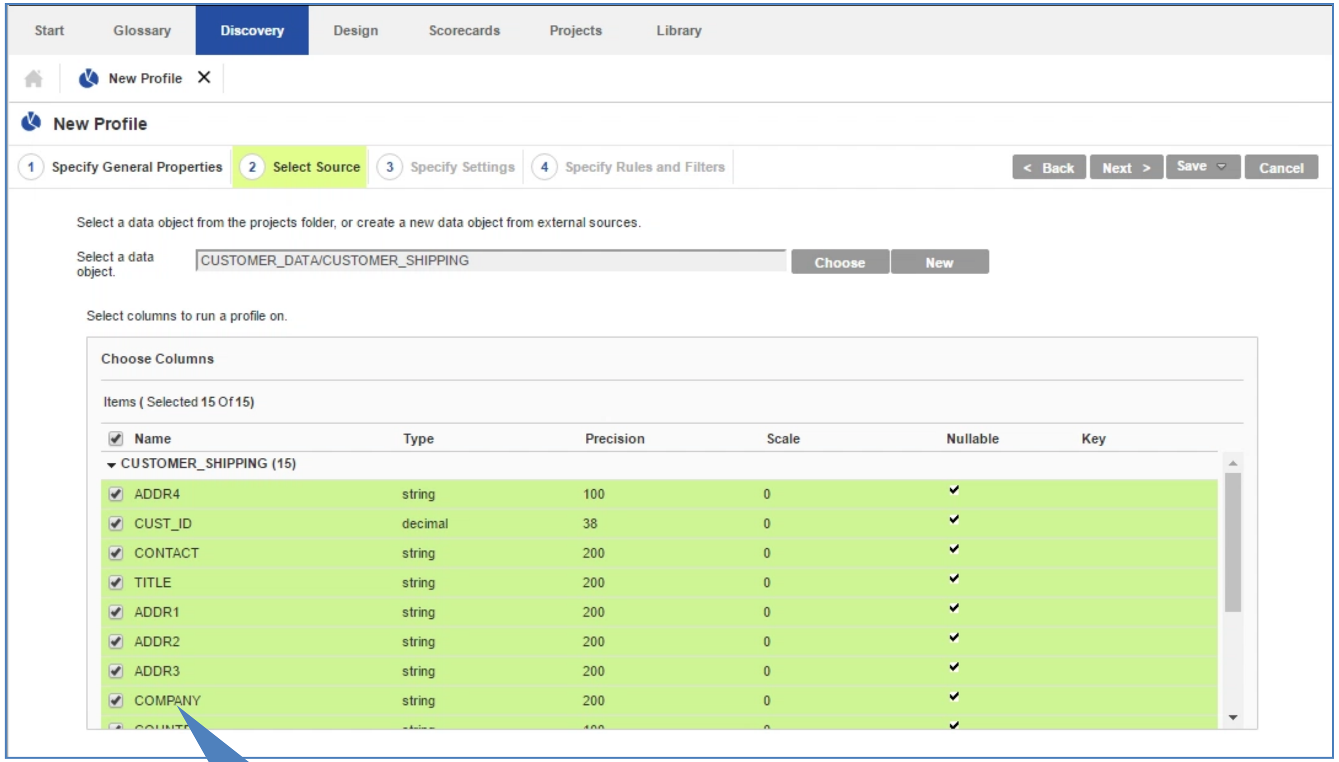
Rich set of transformations
Manage and transform data with data standardization, validation, enrichment, de-duplication, and consolidation capabilities.

Reusable rules & accelerators
Apply pre-built business rules and accelerators and reuse common data quality rules to save time and resources.

Exception management
Allow business users to review, correct, and approve exceptions throughout the automated process.

Select only columns to be profiled

IDQ



Select a data object from the projects folder, or create a new data object from external sources.

Select a data object. CUSTOMER_DATA/CUSTOMER_SHIPPING

Select columns to run a profile on.

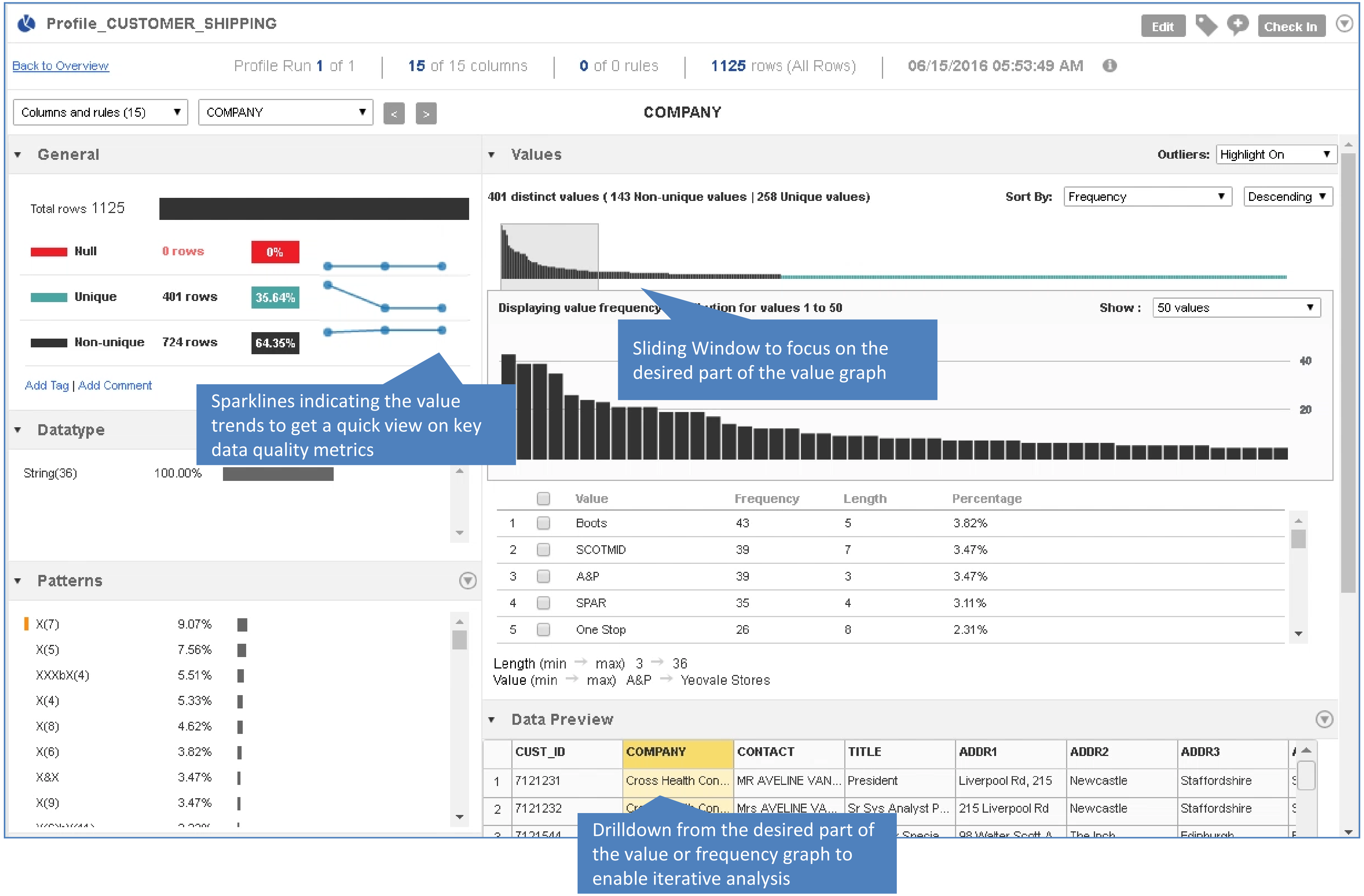
Choose Columns

Items (Selected 15 Of 15)

☑ Name	Type	Precision	Scale	Nullable	Key
CUSTOMER_SHIPPING (15)					
☑ ADDR4	string	100	0	✓	
☑ CUST_ID	decimal	38	0	✓	
☑ CONTACT	string	200	0	✓	
☑ TITLE	string	200	0	✓	
☑ ADDR1	string	200	0	✓	
☑ ADDR2	string	200	0	✓	
☑ ADDR3	string	200	0	✓	
☑ COMPANY	string	200	0	✓	

Select the columns you want to profile on.

IDQ



Profile_CUSTOMER_SHIPPING

Edit | Check In

Back to Overview | Profile Run 1 of 1 | 15 of 15 columns | 0 of 0 rules | 1125 rows (All Rows) | 06/15/2016 05:53:49 AM

Columns and rules (15) | COMPANY

COMPANY

General

Total rows 1125

Null	0 rows	0%
Unique	401 rows	35.64%
Non-unique	724 rows	64.35%

Add Tag | Add Comment

Datatype

String(36)	100.00%

Patterns

X(7)	9.07%
X(5)	7.56%
XXXbX(4)	5.51%
X(4)	5.33%
X(8)	4.62%
X(6)	3.82%
X&X	3.47%
X(9)	3.47%

Values

Outliers: Highlight On

401 distinct values (143 Non-unique values | 258 Unique values)

Sort By: Frequency | Descending

Displaying value frequency ... for values 1 to 50

Show : 50 values

	Value	Frequency	Length	Percentage
1	Boots	43	5	3.82%
2	SCOTMID	39	7	3.47%
3	A&P	39	3	3.47%
4	SPAR	35	4	3.11%
5	One Stop	26	8	2.31%

Length (min → max) 3 → 36
Value (min → max) A&P → Yeovale Stores

Data Preview

	CUST_ID	COMPANY	CONTACT	TITLE	ADDR1	ADDR2	ADDR3
1	7121231	Cross Health Con...	MR AVELINE VAN...	President	Liverpool Rd, 215	Newcastle	Staffordshire
2	7121232	Cr... Con...	Mrs AVELINE VA...	Sr Sys Analyst P...	215 Liverpool Rd	Newcastle	Staffordshire
3	7121544			... Specia...	98 Walter Scott A...	The Inch	Edinburgh

Sparklines indicating the value trends to get a quick view on key data quality metrics

Sliding Window to focus on the desired part of the value graph

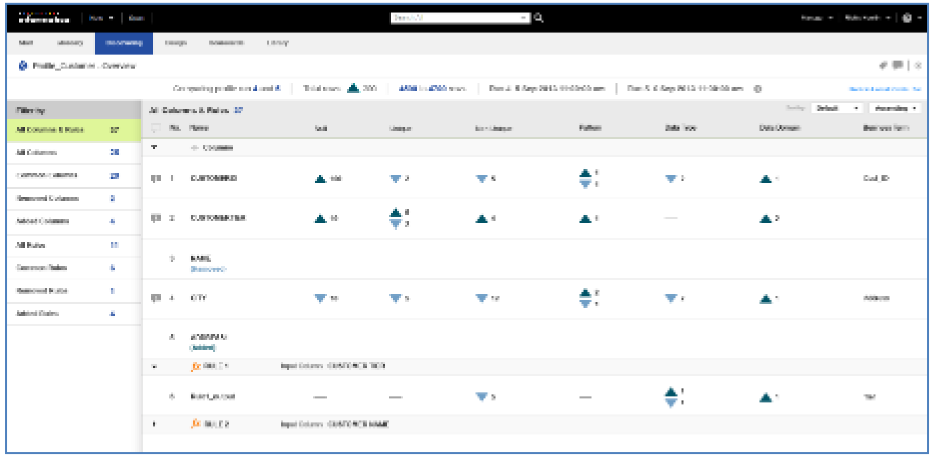
Drilldown from the desired part of the value or frequency graph to enable iterative analysis

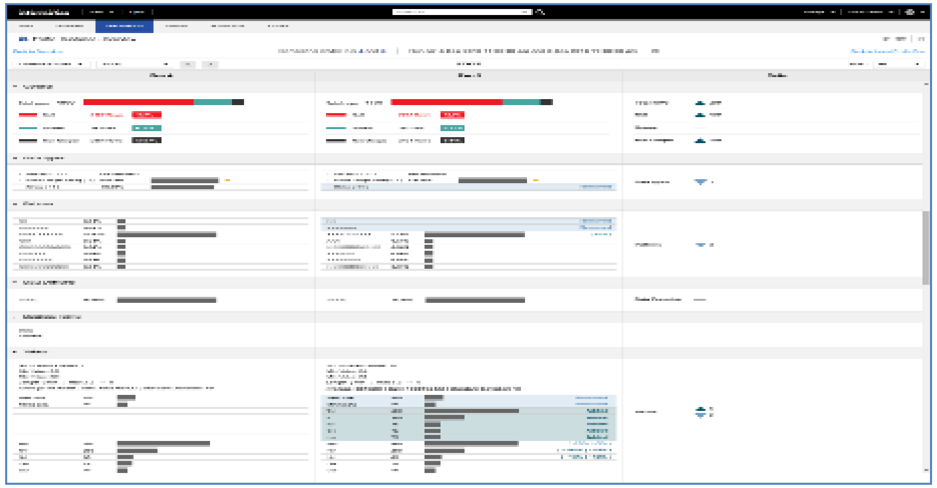
Informatica

Compare Profile Results

IDQ

- Understand data quality trends through time by comparing historical profile results
- Compare column and rule profile results between two profile runs
- Detailed comparisons include changes in datatypes, patterns, nulls and distinct counts





Informatica

Resources

1. Configure Access Axon/IDQ: Click Here
2. Configure Access Axon/EDC: Click Here
3. Configure Access Axon/DPM: Click Here
4. Axon/EDC Automatic Onboarding Workflow: Click Here
5. Automate Data Quality Rules in Axon: Click Here
6. EDC Sizing Guide: Click Here
7. Profiling Sizing Guide: Click Here
8. Integrated Monitoring for Capacity Planning/Resource Utilization: Click Here
9. Product Availability Matrix (PAM): Click Here
10. AWS Informatica Marketplace Offerings: Click Here
11. Azure Informatica Marketplace Offerings: Click Here
12. Deploying DIS on GRID: Click Here
13. Informatica Axon Data Governance Playbook: Click Here

Thank You!

Questions?

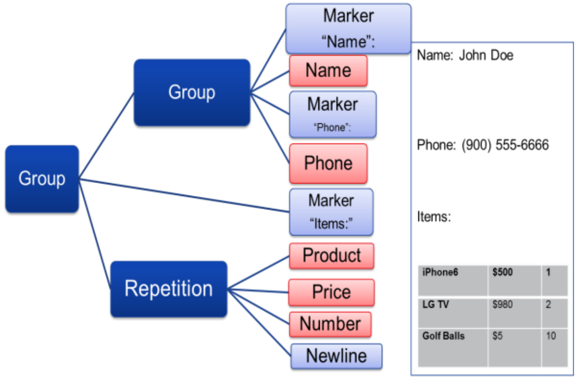
Appendix

Additional Claire Details

Decipher Data (schema extraction)

- High level analysis using A* based dynamic programming
- Genetic Algorithms to identify complex sub-structures
- Various NLP algorithms to modify model based on semantics
  - Identify text blocks that are not for parsing (comments, free text, etc)
  - Identifying patterns in the input
  - Element naming and semantics
  - Map between inputs and models

Extendible with user and vertical specific types

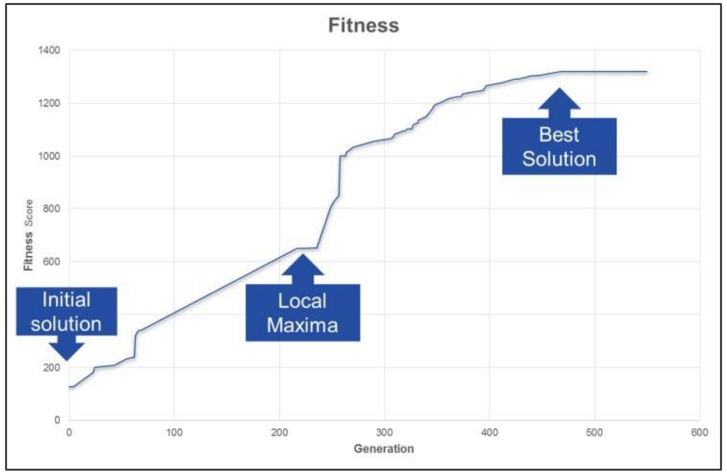


Name: John Doe

Phone: (900) 555-6666

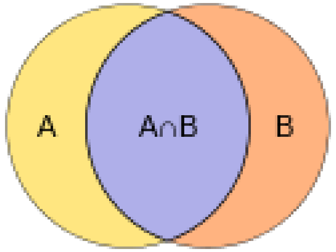
Items:

iPhone6	$500	1
LG TV	$980	2
Golf Balls	$5	10

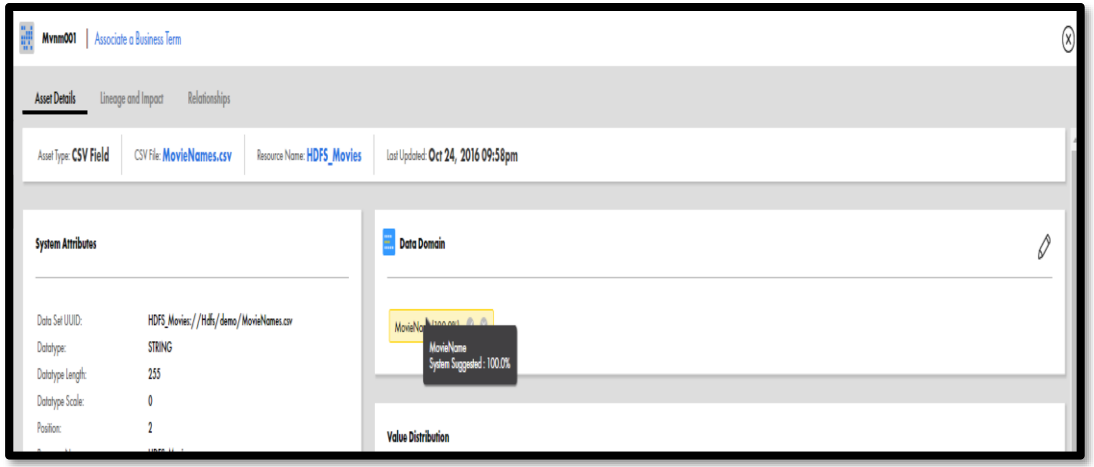


Artificial Intelligence to Cluster Data

- Column Similarity based on Data Overlap
- Large Overlap of Distinct Values:
  - Jaccard distance = $1 - \frac{|S(X) \cap S(Y)\ \ |}{|S(X) \cup S(Y)\ \ |}$
- Similar Value Frequencies for overlapping columns
  - Bray Curtis Similarity: $\frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{j=1}^{n} x_j + y_j}$
- Clustering based on Column metadata and Jaccard Coefficient and then computing Bray Curtis Similarity.



Like photo tagging
CLA*I*RE for Columns


Mvnm001 | Associate a Business Term

Asset Details | Lineage and Impact | Relationships

Asset Type: CSV Field | CSV File: MovieNames.csv | Resource Name: HDFS_Movies | Last Updated: Oct 24, 2016 09:58pm

System Attributes

Data Set UUID:	HDFS_Movies://Hdfs/demo/MovieNames.csv
Datatype:	STRING
Datatype Length:	255
Datatype Scale:	0
Position:	2

Data Domain

MovieNa... (100.0%)

MovieName
System Suggested : 100.0%

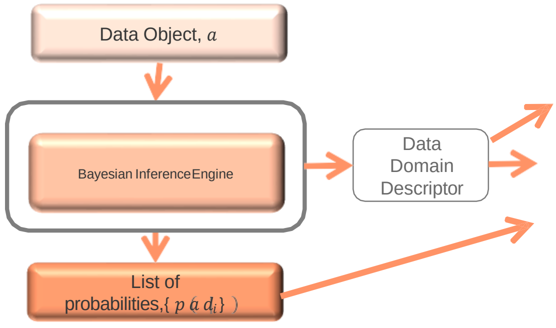
Value Distribution

Artificial Intelligence for Security Analytics

- Bayesian Inference for auto-morphism and format preserve masking
- UBA unsupervised machine learning combined with Principal component analysis to create multi-dimensional model of user activities
- BIRCH technique for unsupervised hierarchical clustering and to identify changes in user behavior
- Validation based on distance and density for outlier detection and Grubbs' test

The Grubbs' test statistic is defined as:

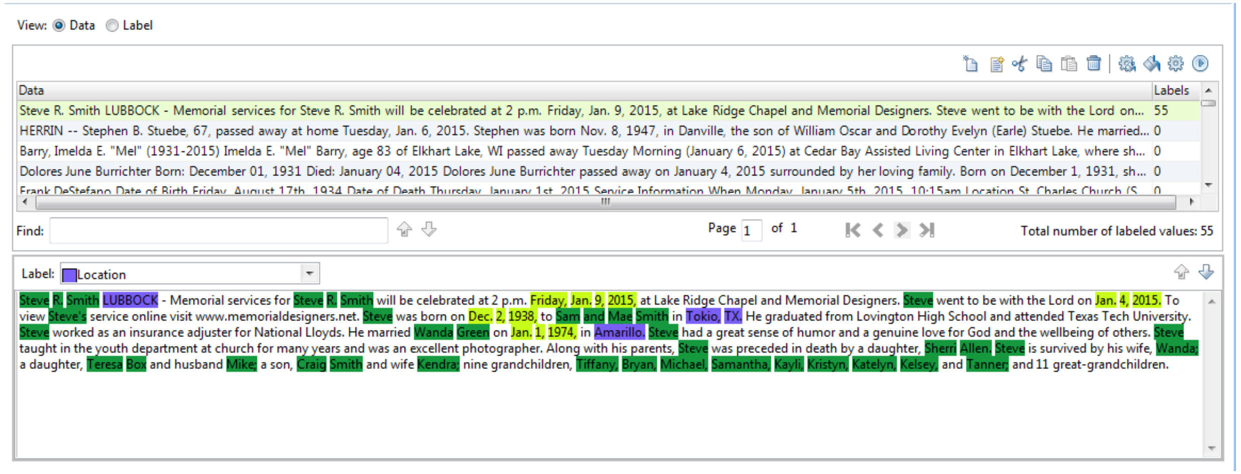
$$G = \frac{\max\limits_{i=1,...,N} \left|Y_i - \bar{Y}\right|}{s}$$


Data Object, a
Bayesian Inference Engine
Data Domain Descriptor
List of probabilities,{ $p(a \mid d_i)$ })

Alphabets	"0123456789" "ABCDEFGHJKLMNOPRQSTUVWXYZ" "234567"
Positional Map	3, 2, 2, 2, 1, 1, 1
Special Condition	if x*[0] = 3 and x*[1 - 3] are in the range between YAA and ZYZ then repeat transform

Artificial Intelligence to Extract Entities

- NLP techniques to identify and extract data entities from strings
  - Extract Product Code from product descriptions
  - Identify Organization vs. Person information
  - Extract entities from unstructured Data
- Use Classifier Transform (Mallet from UMASS) to categorize data based on a custom classification model
- Statistical algorithms identify common and uncommon elements of your data



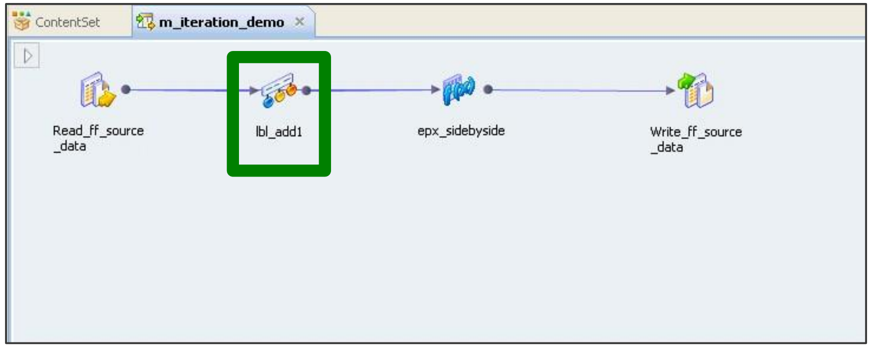
View: ● Data ○ Label

Data	Labels
Steve R. Smith LUBBOCK - Memorial services for Steve R. Smith will be celebrated at 2 p.m. Friday, Jan. 9, 2015, at Lake Ridge Chapel and Memorial Designers. Steve went to be with the Lord on...	55
HERRIN -- Stephen B. Stuebe, 67, passed away at home Tuesday, Jan. 6, 2015. Stephen was born Nov. 8, 1947, in Danville, the son of William Oscar and Dorothy Evelyn (Earle) Stuebe. He married...	0
Barry, Imelda E. "Mel" (1931-2015) Imelda E. "Mel" Barry, age 83 of Elkhart Lake, WI passed away Tuesday Morning (January 6, 2015) at Cedar Bay Assisted Living Center in Elkhart Lake, where sh...	0
Dolores June Burrichter Born: December 01, 1931 Died: January 04, 2015 Dolores June Burrichter passed away on January 4, 2015 surrounded by her loving family. Born on December 1, 1931, sh...	0
Frank DeStefano Date of Birth Friday, August 17th, 1934 Date of Death Thursday, January 1st, 2015 Service Information When Monday, January 5th, 2015, 10:15am Location St. Charles Church (S...	0

Find: Page 1 of 1 Total number of labeled values: 55

Label: ▨Location

Steve R. Smith LUBBOCK - Memorial services for Steve R. Smith will be celebrated at 2 p.m. Friday, Jan. 9, 2015, at Lake Ridge Chapel and Memorial Designers. Steve went to be with the Lord on Jan. 4, 2015. To view Steve's service online visit www.memorialdesigners.net. Steve was born on Dec. 2, 1938, to Sam and Mae Smith in Tokio, TX. He graduated from Lovington High School and attended Texas Tech University. Steve worked as an insurance adjuster for National Lloyds. He married Wanda Green on Jan. 1, 1974, in Amarillo. Steve had a great sense of humor and a genuine love for God and the wellbeing of others. Steve taught in the youth department at church for many years and was an excellent photographer. Along with his parents, Steve was preceded in death by a daughter, Sherri Allen. Steve is survived by his wife, Wanda; a daughter, Teresa Box and husband Mike; a son, Craig Smith and wife Kendra; nine grandchildren, Tiffany, Bryan, Michael, Samantha, Kayli, Kristyn, Katelyn, Kelsey, and Tanner; and 11 great-grandchildren.


ContentSet
m_iteration_demo
Read_ff_source_data
lbl_add1
epx_sidebyside
Write_ff_source_data