DATA GOVERNANCE AND MASTER DATA MANAGEMENT CONFERENCE EUROPE

11 - 14 March 2024 | London, UK

Please score and comment on this session and speaker in the event mobile app

IRMUK

INTELLIGENT BUSINESS STRATEGIES

The Value of AI in Master Data Management

Mike Ferguson
Managing Director
Intelligent Business Strategies

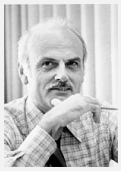
IRM EDG and MDM Conference
March 2024

Who Is Mike Ferguson? – A Leading Analyst in Data Management & Analytics

Codd & Date

Edgar F Codd

Turing Award Winner



Relational Model

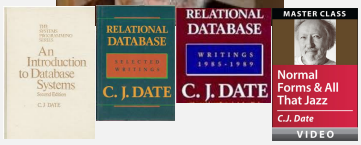
All relational DBMSs | SQL Language | Data Normalisation

Mike Ferguson

Chief Architect teradata.

1st massively parallel relational DBMS

C J Date



Founded 1992 — INTELLIGENT BUSINESS STRATEGIES

Mike Ferguson Is a Leading Industry Analyst / Consultant and Conference Chairman of Big Data LDN - Leading The Industry In Data Management and Analytics

BIG DATA LDN.

Big Data LDN is the largest data & analytics conference in Europe

- 20000 delegates
- 180 vendors
- 15 theatres
- 300+ speakers

About Intelligent Business Strategies

- UK-based independent IT analyst and consulting firm founded 1992 specialising in data management and analytics
- Mike Ferguson is co-founder of the Data and Analytics Retreat and Conference Chairman Big Data LDN
- Three main lines of business

Research

- Market research
  - 4th Industrial Revolution Survey
- D&A product research
  - Data Catalogs
  - Data Fabric
  - Data Governance
  - Master Data Management
  - Distributed DBMS
  - Analytical DBMS

Education

- Data Strategy
- How to Govern Data Across a Distributed Data Landscape
- Practical Guidelines for Implementing a Data Mesh
- Modern Data Architecture
- DW Modernisation
- DW Migration to the Cloud
- Embedded Analytics, Intelligent Apps & AI Automation
- Public classes (anyone)
- On-site classes (single client)
  - Customers, vendors, systems integrators
- On-line (public & on-sites)

Consulting

- Customers
  - D&A Strategy, Data Architecture
  - D&A Technology selection
  - D&A Reviews, Data Governance
  - Implementation Project advisory
- Vendors
  - Product strategy
  - Product positioning
  - Marketing support
    - Speaking at vendor events
    - White papers
    - Webinars
- Venture Capitalists
  - Due-diligence, Asset advisory

www.intelligentbusiness.biz

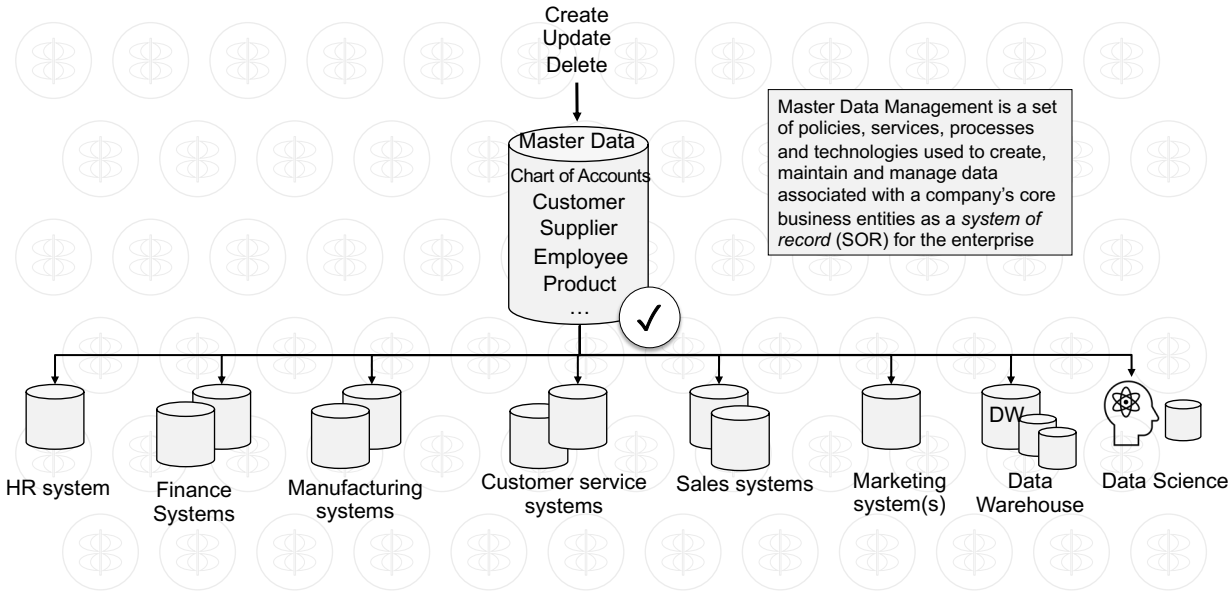
5

Topics

- What is master data management?
- How does AI help in building master data entities?
  - Automated enrichment of master data business terms in a business glossary
  - Using generative AI to build a data models for master data
  - Using AI to auto discover physical master data and map it to a business glossary
  - Automating sensitive data detection and master data metadata enrichment using AI
  - Using generative AI to create data quality rules to detect poor quality source data
  - AI-assisted data cleansing and join recommendations in data engineering
  - Using AI to automate matching to create master data
  - Using Generative AI to build pipelines to create master data products
- Generative AI to automatically create product information and marketing content from master data

6

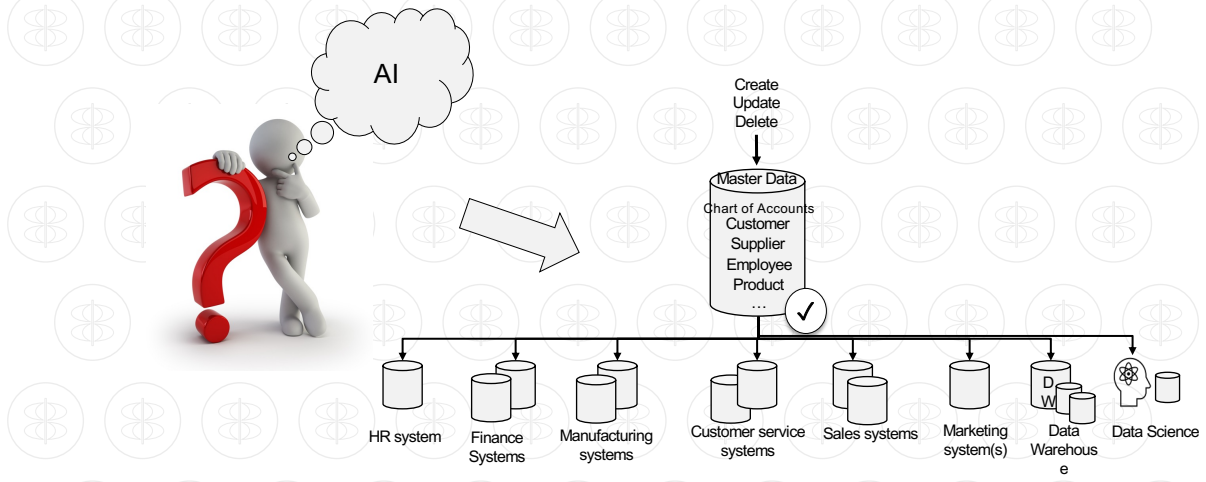
What Is Master Data Management?



Business Benefits From MDM Implementation

- 360º view of core business entities, e.g. Customer, Product
- Sharing data across the enterprise
- Improving master data quality across all applications
- Consistent, better quality data-driven decisions and analytics
- Consistent integrated customer intelligence in all front-office channels
- Reduced cost of operating
- Optimisation of operations to speed up process execution
- Better business planning
- Legislative and regulatory compliance, e.g. GDPR

How Does AI Help In Master Data Management?



There are Different Types of Artificial Intelligence

- Machine learning
  - Supervised learning
    - Predictive models
    - Prescriptive models (predictive models + rules = recommendations, alerts)
  - Unsupervised learning
    - Clustering
    - Anomaly detection
- Deep learning
  - e.g., Multi-layer neural networks for computer vision, language translation
- Reinforcement learning
  - Reward oriented self-learning – e.g. Ray Framework
- Generative AI
  - Uses embeddings and large language models

What is Generative AI?

What is Generative AI?

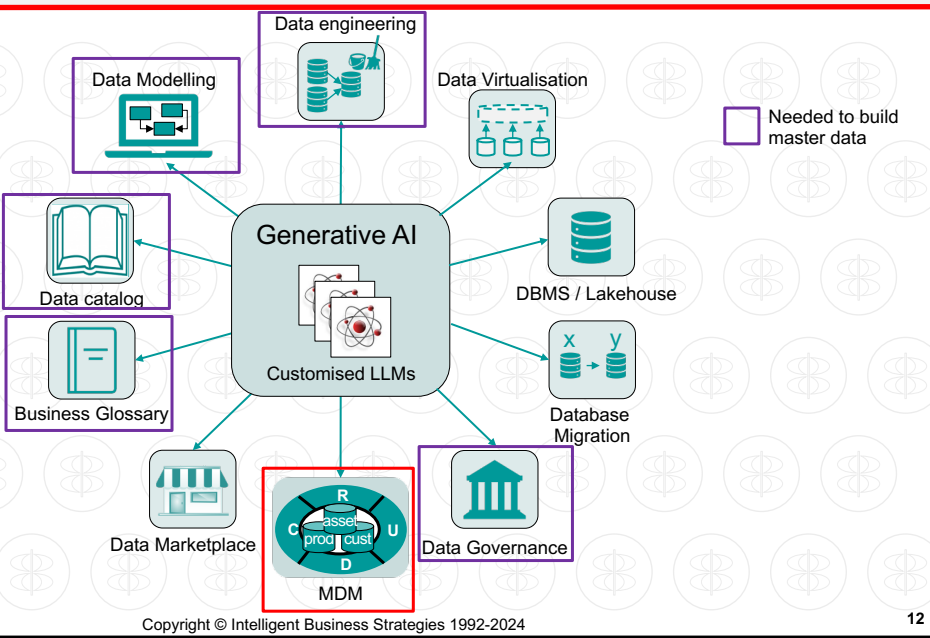
A subset of deep learning where multi-layer neural network models **generate new content** such as text, images, audio, video, code, and synthetic data in response to prompts based on what the models have learned from patterns in the content they were trained on

- Foundation Large Language Model examples:
  - AWS Titan, Databricks Dolly, Google PaLM2 & Codex,
  - Open AI Generative Pre-Trained Transformer (GPT4)
  - Anthropic Claude

- General use cases
  - Text generation
  - Virtual assistant – chat
  - Conversational search
  - Summarisation – text extraction
  - Code generation
  - Synthetic data generation
  - Image generation and classification
  - Video generation
- Benefits
  - Improve customer and employee experience
  - Productivity, automation
  - Ease of use, lower skills bar
  - Democratisation of D&A development

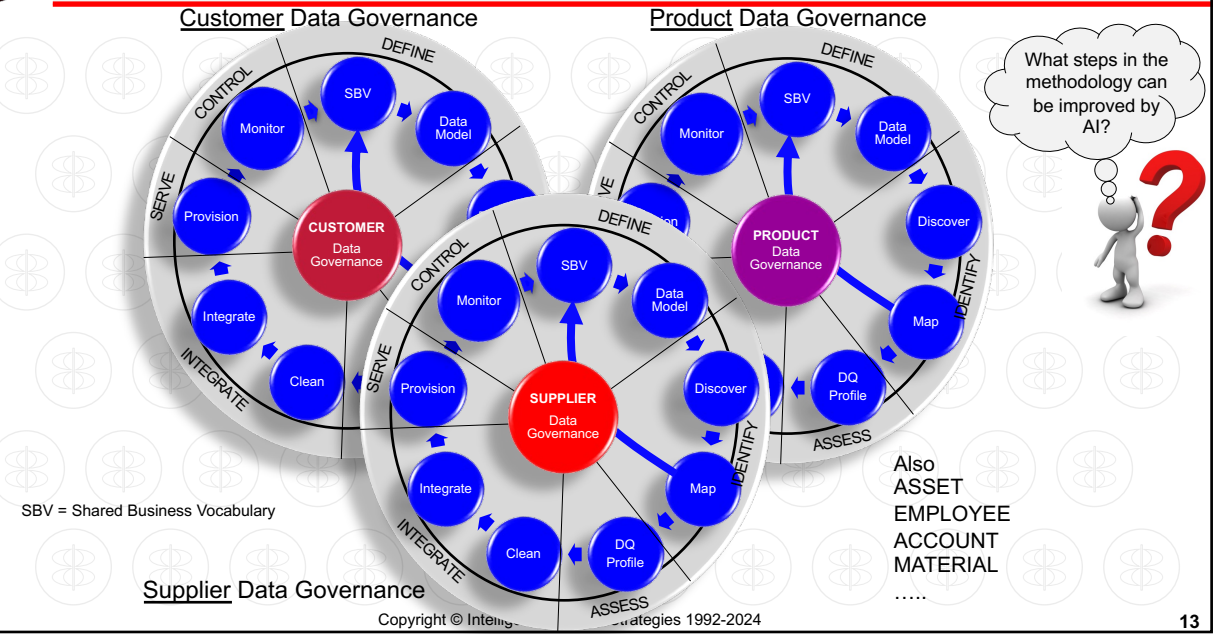
11

Generative AI Has Emerged In Almost Every Aspect of Data Management



12

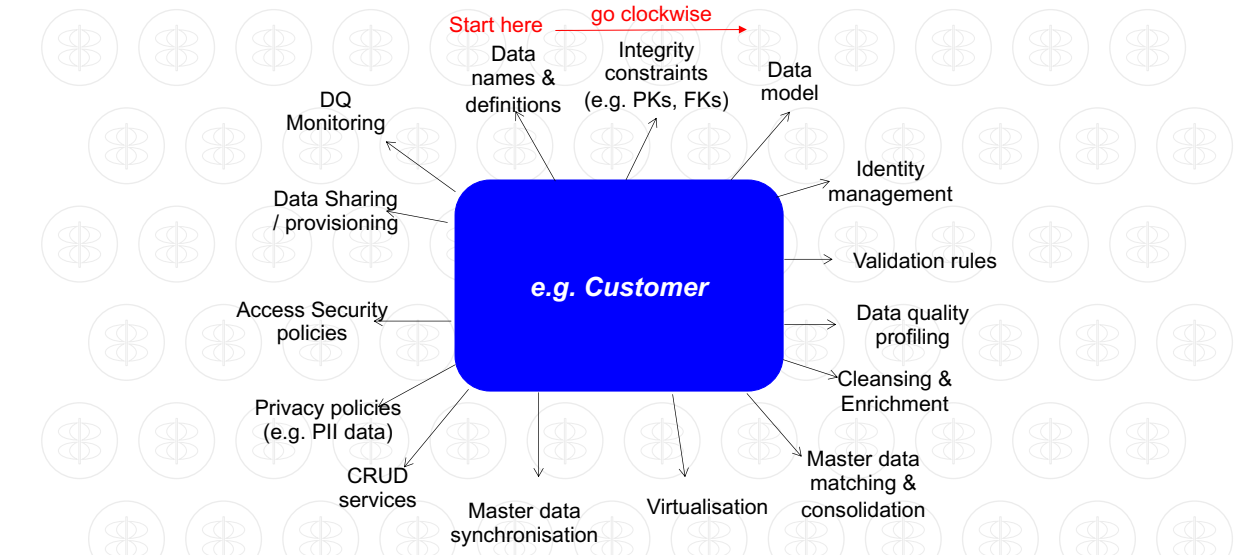
A Methodology Is Needed To Build Master Data Entities When Implementing MDM - AI Can Help Improve Speed of Implementation



Master Data Governance Policies Also Need To Be Defined

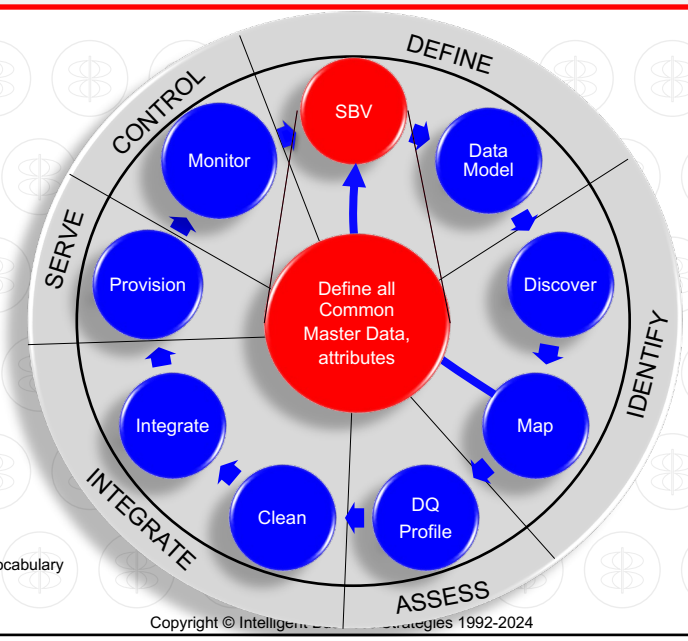
- Several types of policy need to be defined to govern master data
  - Master data integrity constraints
  - Master data quality validation policies
  - Master data cleansing rules
  - Master data integration rules
  - Master data provisioning policies, e.g. Language rendering, formats, etc.
  - Master data access control policies (CRUD security)
  - Master data privacy policies
  - Master data sharing policies
  - Master data lifecycle (expiration) policies
- These policies can be applied to:
  - Individual data items
  - All data items in a master data entity, e.g. Customer, Asset, Product
  - Both
- Policies may be restricted to limit their governance capability
  - e.g. limited to master data in a specific application or process

Implementing Master Data Management For Core Entities – What Needs To Be Considered?



15

A Shared Business Vocabulary Is The Anchor Point For Data Provisioning and Data Governance



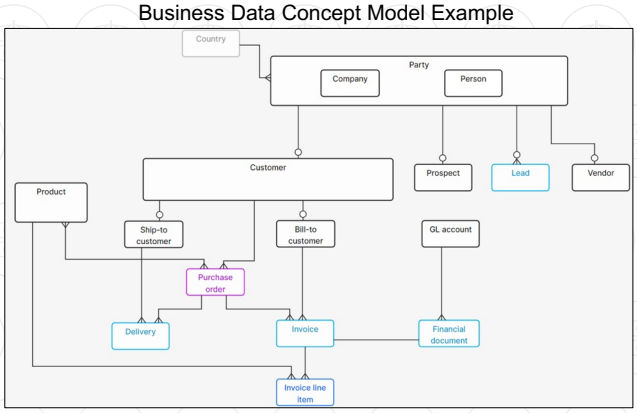
SBV = Shared Business Vocabulary

16

A Good Way To Start A Business Glossary Is To Create A Business Data Concept Model

- Identify the data concepts, properties and relationships and construct a data concept model
- It is good practice to highlight all master and transaction data concepts in your data concept model

Business Data Concept Model Example

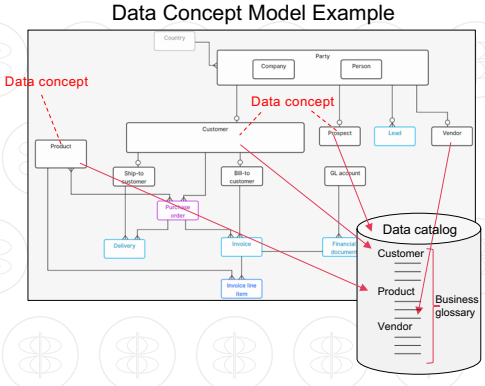


Some vendors provide complete definitions for business data entities and all their attributes to get you started quickly, e.g., Collibra uses Schema.org, Microsoft has a CDM

Steps To Creating A Common Vocabulary – From Data Concept Model To Master Data Product

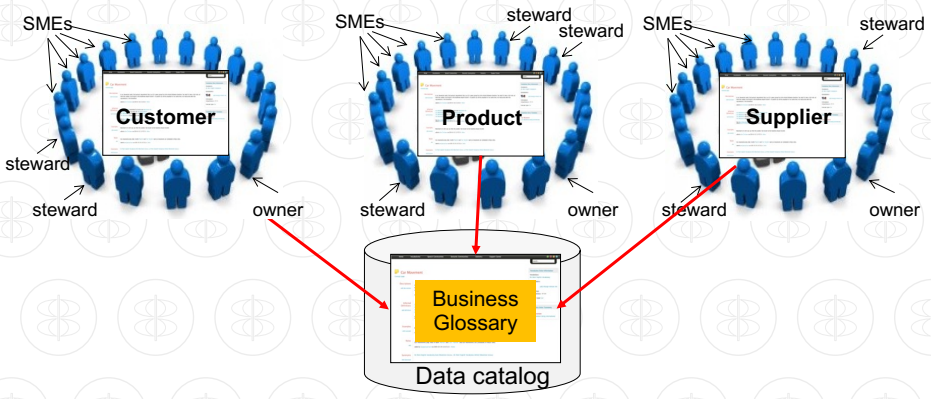
1. Identify the data concepts, properties and relationships and construct a data concept model
2. The data concepts become the 'skeleton data entities' in the common vocabulary
3. Each data concept and its attributes should be defined in the business glossary as a data entity with a data owner
   - Including master data entities
4. Use the catalog to discover the data for each master data entity in underlying data stores across the data landscape
5. Design DataOps component-based pipelines to create the master data entities (master data products) with common vocabulary data names and store them in an MDM system
6. Publish all master data entities stored in the MDM system as data products in a data marketplace

Data Concept Model Example

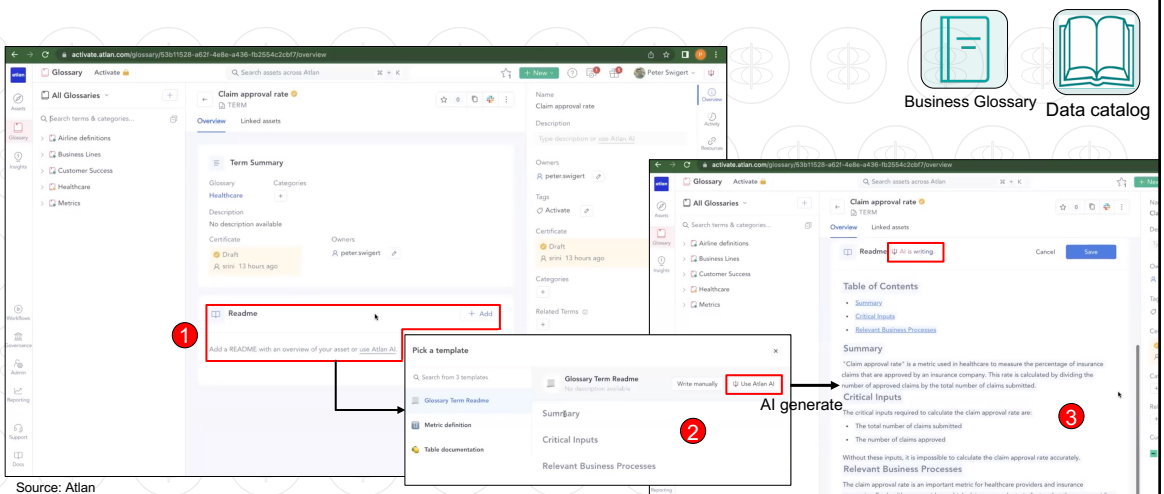


Several Communities Are Likely To Be Involved In Defining Common Data Names And Definitions For Each Master Data Entity In A Business Glossary

Communities may include a mix of IT and subject matter expert (SME) business users



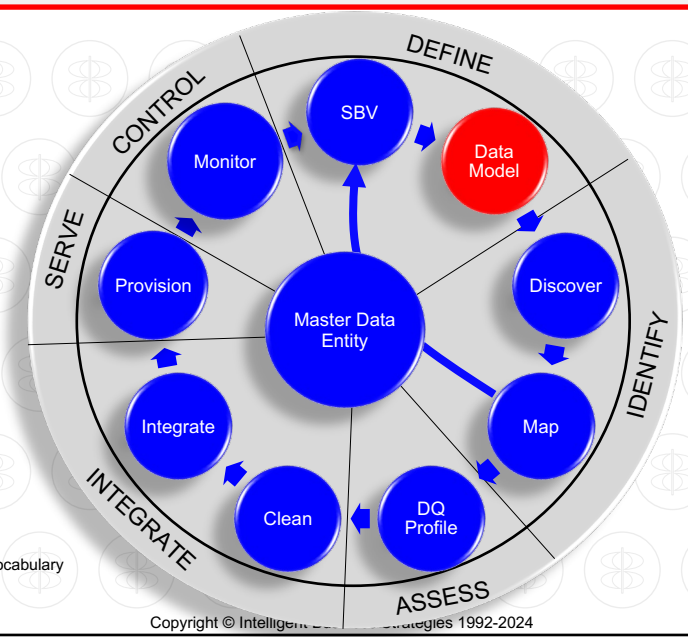
Using Generative AI For Data Catalog Metadata Enrichment – Auto-Generation Of Business Term Descriptions In A Business Glossary In Atlan



Source: Atlan

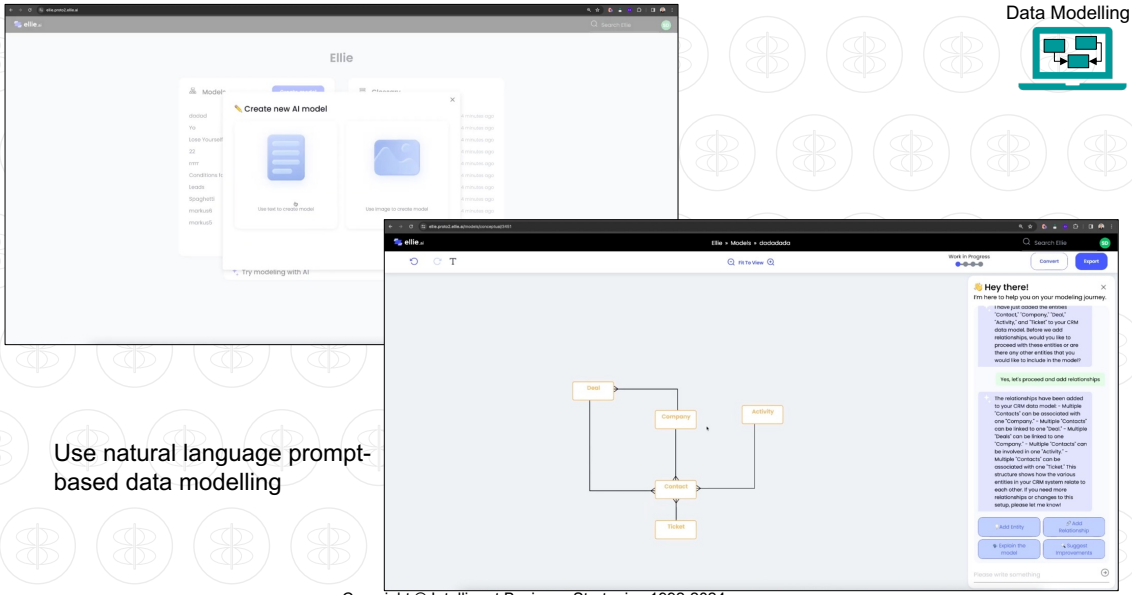
Add a Readme, pick from a template and auto generate the Readme text using a generative LLM

Master Data Modelling



SBV = Shared Business Vocabulary

Data Modelling Using Generative AI – Ellie.ai Is a Popular Data Modelling Tool That Can Use Generative AI to Help Build Data Models

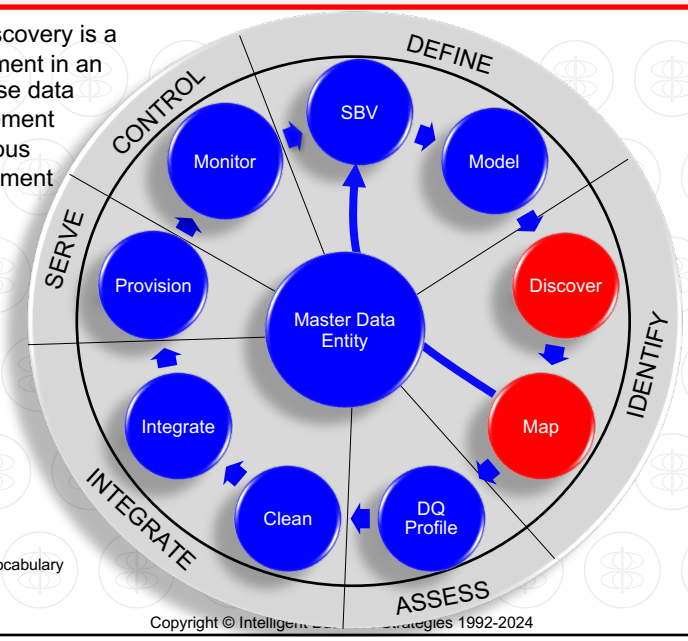


Data Modelling

Use natural language prompt-based data modelling

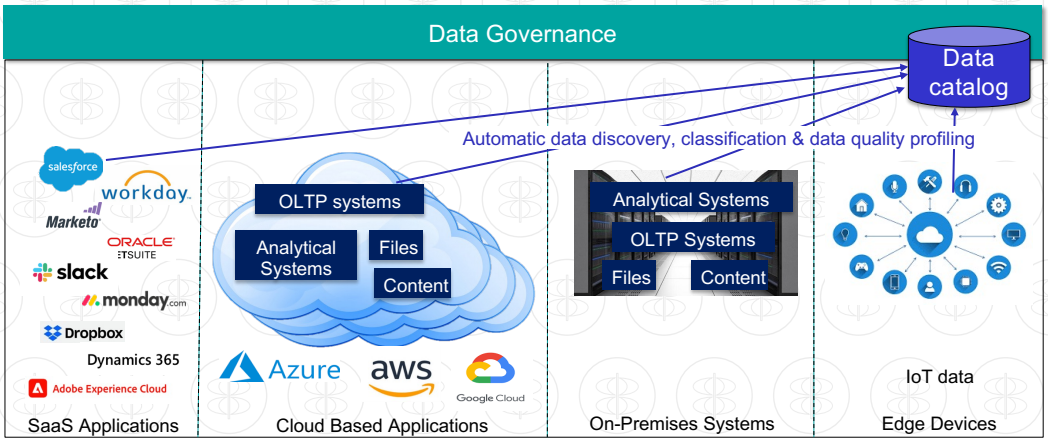
Data Discovery Of On-Premise And Cloud Data Using a Data Catalog

Data discovery is a key element in an enterprise data management continuous improvement process



SBV = Shared Business Vocabulary

The Role Of The Data Catalog – Automated Data Discovery, AI-Classification, Data Quality Profiling And Cataloguing of Source Data

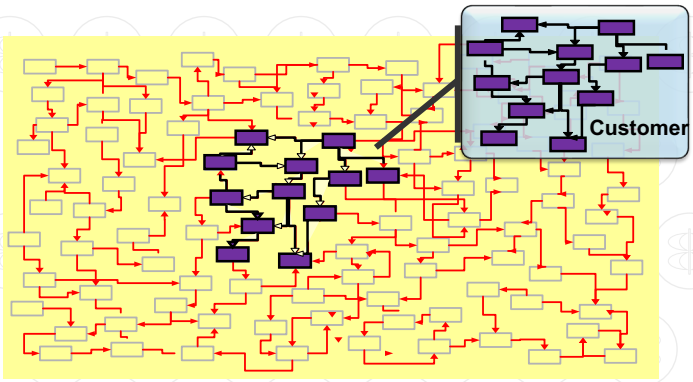


We need a data catalog to automatically discovery what data is available, its quality, sensitivity and where it is across the landscape

Data Catalog Products – A Very Crowded Market

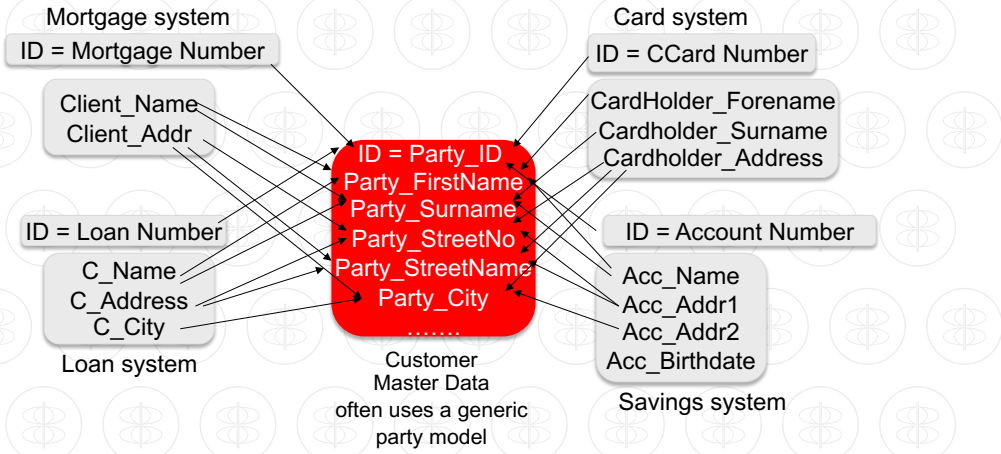
- Alation
- Alex Solutions
- Alteryx Connect
- Amazon Glue Catalog
- Apache Atlas (open source)
- Atlan
- Ataccama ONE Data Catalog
- BigID Data Catalog
- Boomi Atomsphere Data Platform Catalog
- Cambridge Semantic Anzo Catalog
- Cloudera Data Platform SDX Catalog
- Collibra Catalog
- Databricks Unity Catalog
- data.world
- Denodo Catalog
- Google Cloud Data Catalog
- Hitachi Vantara Lumada Data Catalog
- IBM Watson Knowledge Catalog
- Informatica Enterprise Data Catalog
- Microsoft Purview
- Oracle Cloud Infrastructure Data Catalog
- Qlik
  - Enterprise Data Catalog
  - Talend Data Catalog
- Quest (formerly erwin) Data Catalog
- Rocket Software (formerly ASG) Intelligent Data Catalog
- SAP Datasphere Catalog (cloud)
- SAS Information Catalog
- Salesforce Tableau Catalog
- TIBCO Cloud Metadata Catalog
- Top Quadrant TopBraid EDG Data Catalog
- Truist Zaloni Arena Data Catalog

Human Assisted AI Automated Data Discovery Seeks To Determine What Data You Have And Find Complete Data Objects Across Heterogeneous Systems

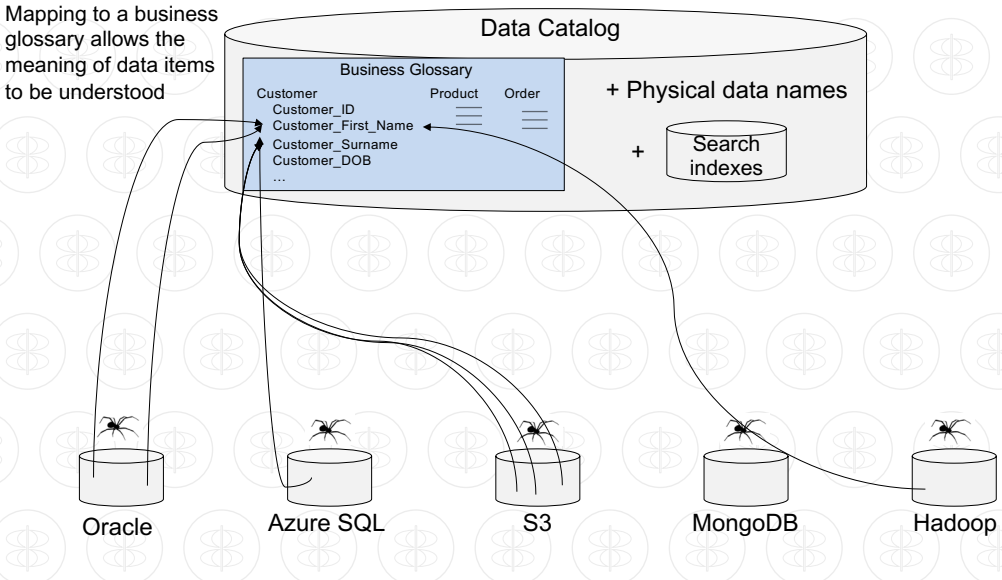


- Automated grouping of tables across multiple systems into business entities

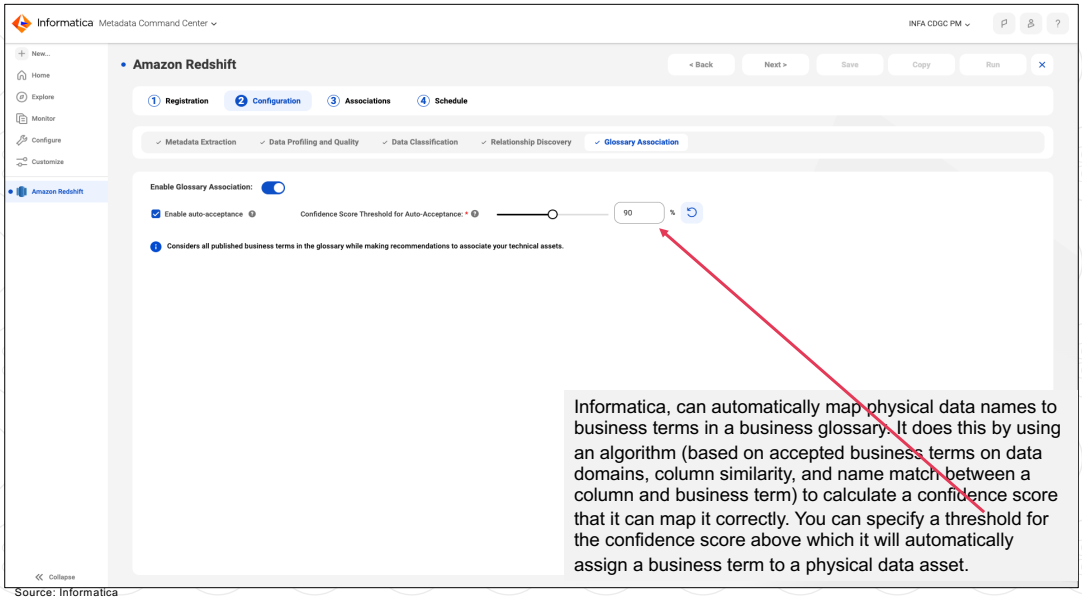
To Get A Single View Of Master Data You Need a Master Data Common Business Vocabulary with Global IDs AND Use AI to Help Map Sources to Business Terms



Automated Data Discovery, **AND Mapping To Common Business Terms In The Business Glossary** Using Data Catalog Crawlers, Rules And Machine Learning



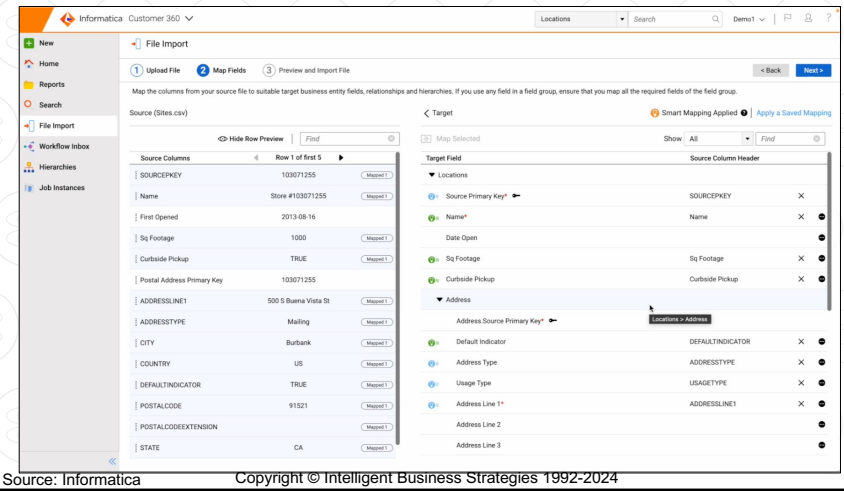
Enabling Glossary Business Term Association In Informatica IDMC Data Governance and Catalog To Map Physical Data Names To Business Terms in a Business Glossary



Source: Informatica

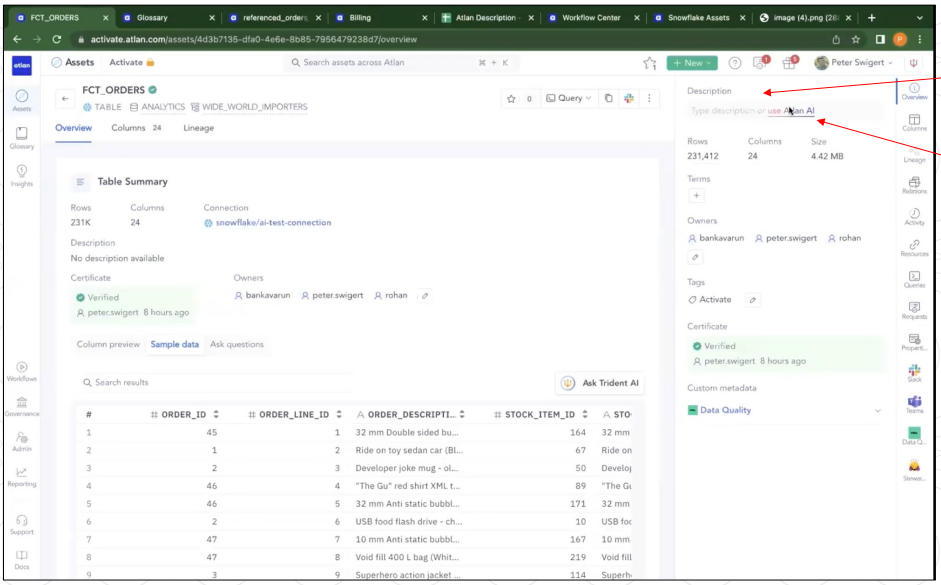
Informatica MDM AI-Driven Smart Field Mapping – Based on Name Similarity

- Exact Field Name mapping - Matches fields of the same name.
- Smart Field Mapping - Matches fields with similar names
  - E.g. If an incoming field `Cust_Name` and a target field `Customer_Name`, it automatically links the `Cust_Name` field with the `Customer_Name` field



Source: Informatica

Generative AI In Data Catalogs
– Automated Metadata Enrichment Using Atlan AI To Generate A Description Of An Asset



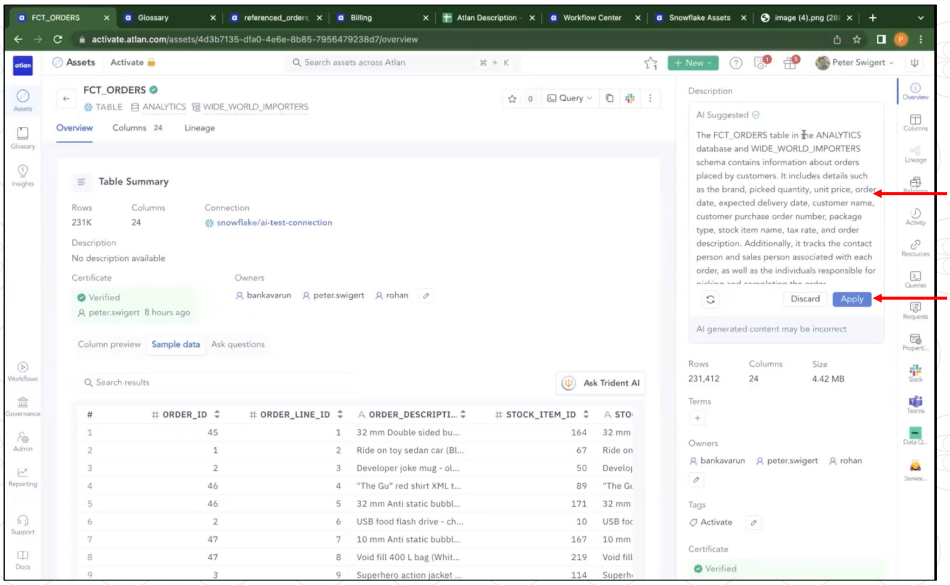
FCT_ORDERS table currently has no description

Request Atlan AI to generate one

Data catalog

Source: Atlan

Atlan Generated Contextual Description
– E.g. The Description Includes The Database The Table Is In, The Schema



Data catalog

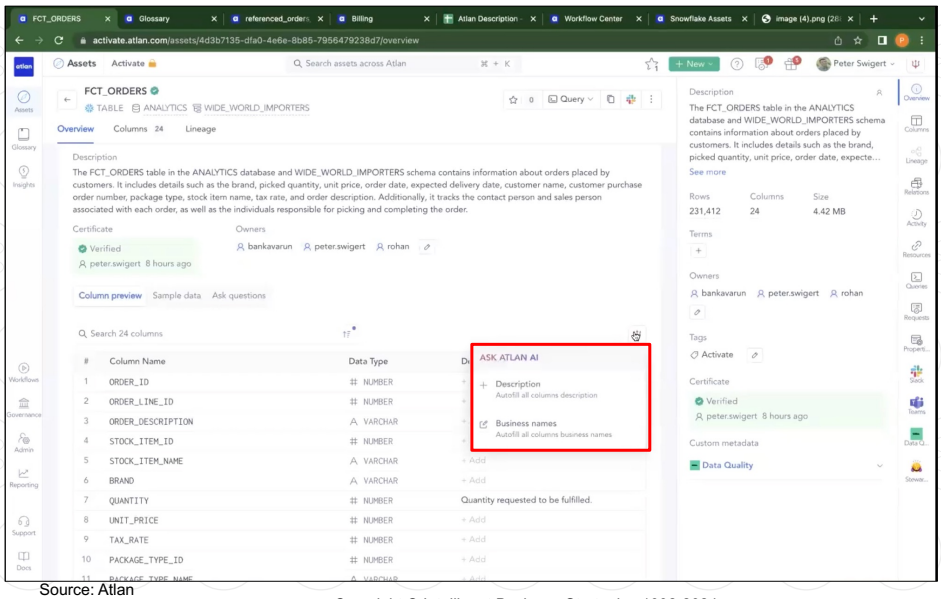
Generated description

User can accept this by clicking “Apply”

The user can also edit the description before accepting it

Source: Atlan

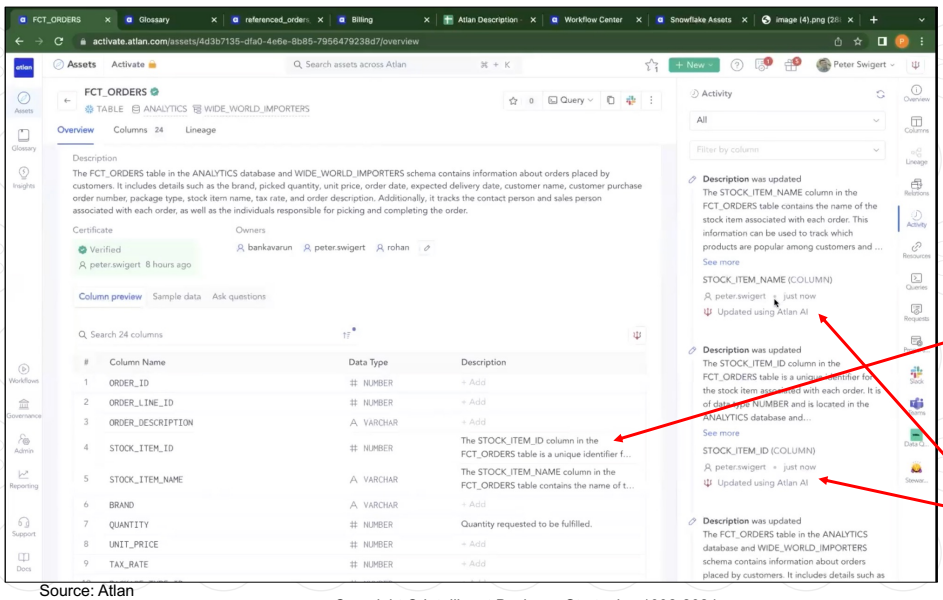
Generative AI And Data Catalogs - Automated Metadata Enrichment Using Atlan AI To Auto-Generate Column Descriptions And Business Data Names



Data catalog

Source: Atlan

Atlan AI Activity Shows A Log Of The Column Description Changes Made By A User Using Atlan AI



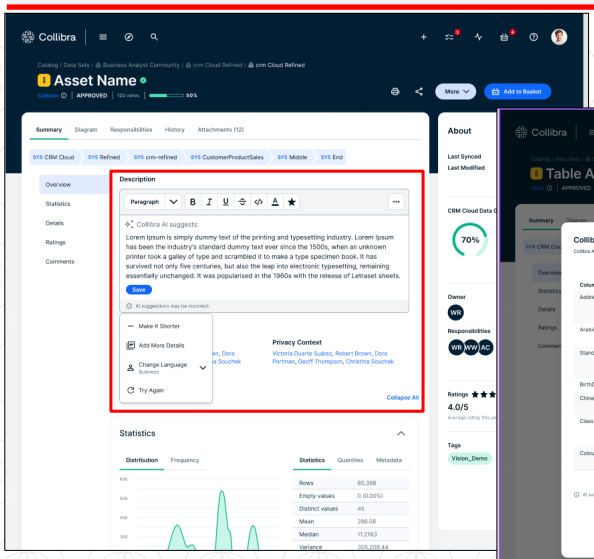
Data catalog

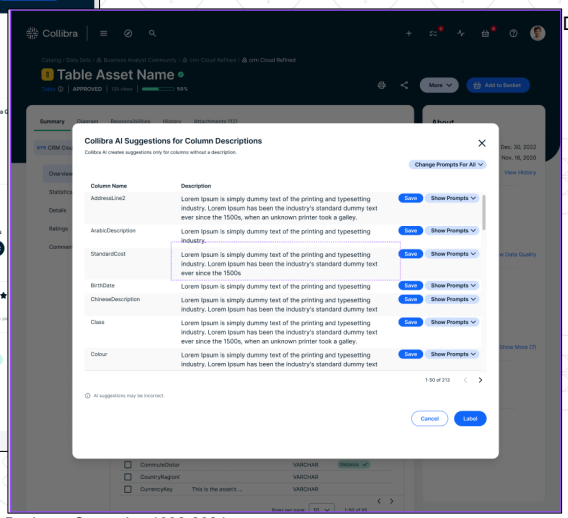
Accepted AI-generated column descriptions are immediately visible

The activity log shows they were updated by a user using Atlan AI

Source: Atlan

Generative AI In Data Catalogs - Automated Metadata Enrichment Using AI-Generated Column Descriptions In Collibra

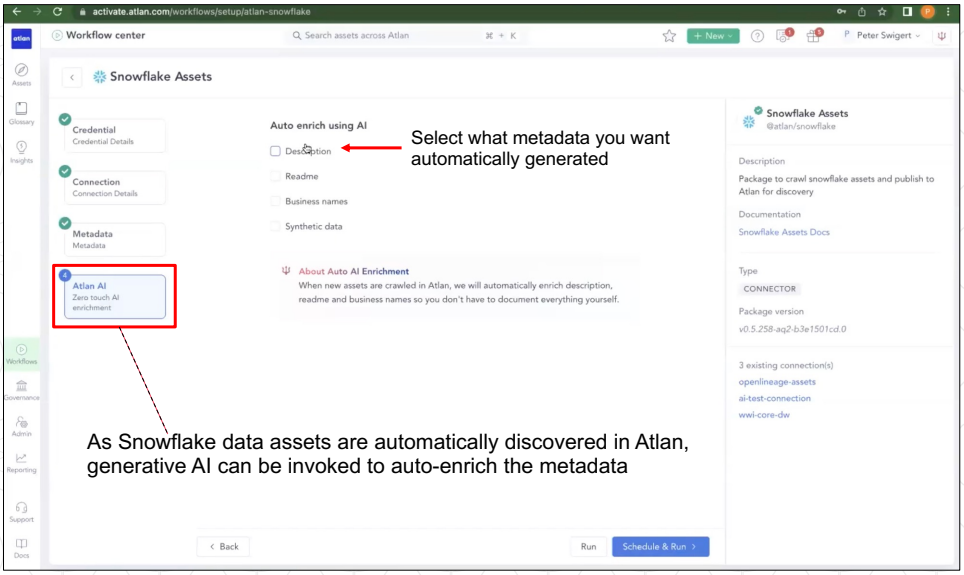




Data catalog

Source: Collibra

Generative AI In Data Catalogs – Automatic AI-Generated Metadata Enrichment At Scale During Source Data Discovery In Atlan



Data catalog

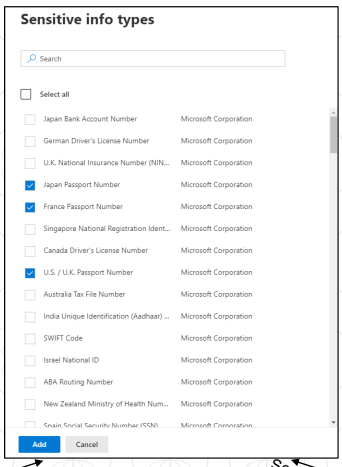
Source: Atlan

Several Vendors Support Predefined Classifiers To Classify Sensitive Data Types – E.g. 200+ Pre-Defined ML Model Classifiers Used In Microsoft Purview

Examples

- People's names
- Phone numbers
- Email addresses
- Postal codes
- Passport numbers
- Driving license numbers
- Social security numbers
- Credit card numbers
- Bank account numbers

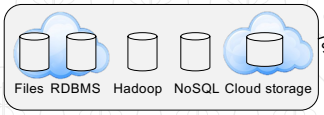
….



These are examples of very specific sensitive data classifications

You can also add your own sensitive data types

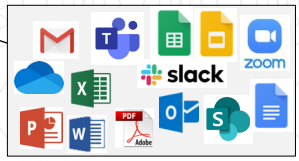
Data



Scan & classify

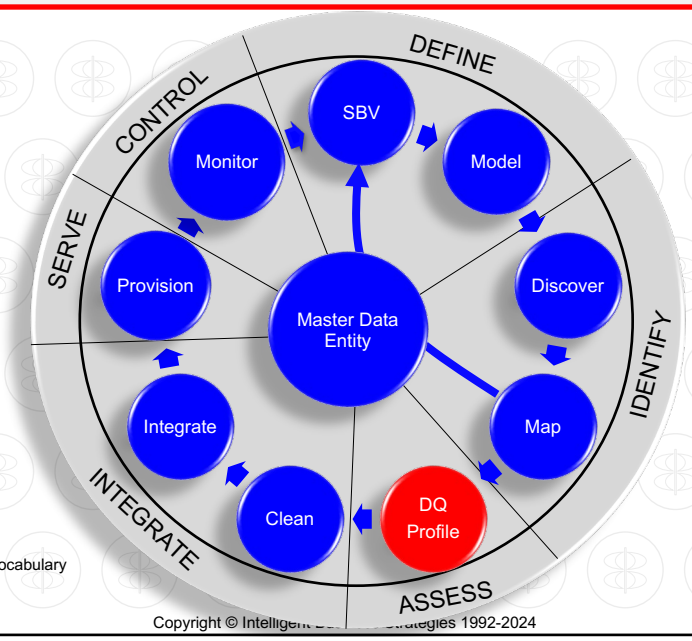
Scan & classify

Content



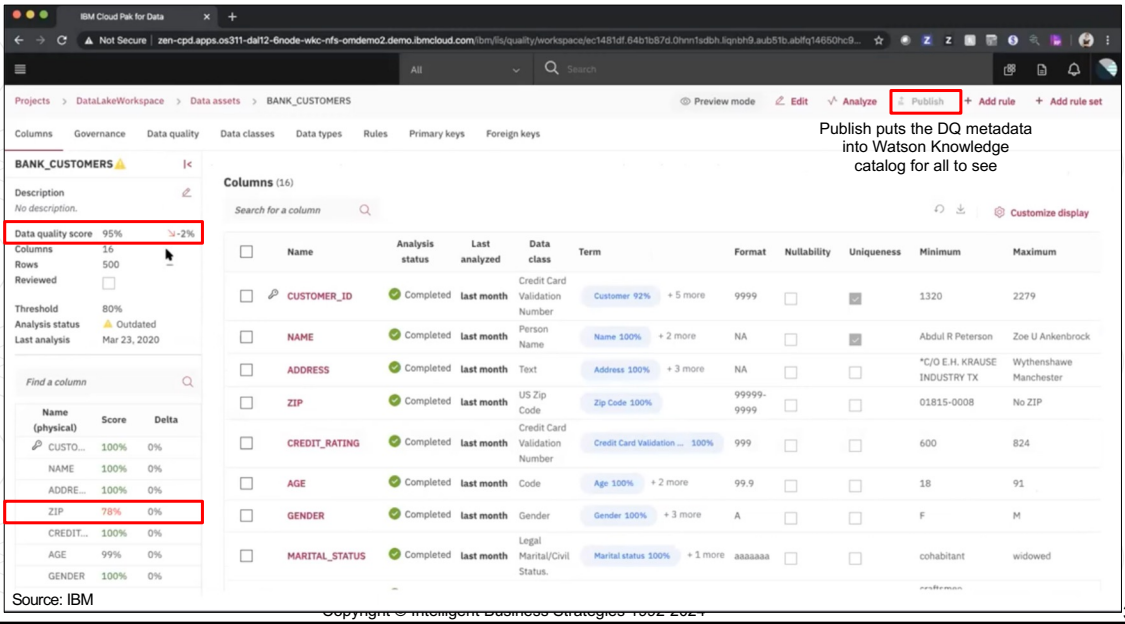
You need to identify where this sensitive data is in all data stores and content

Data Profiling On-Premise And Cloud Master Data Sources



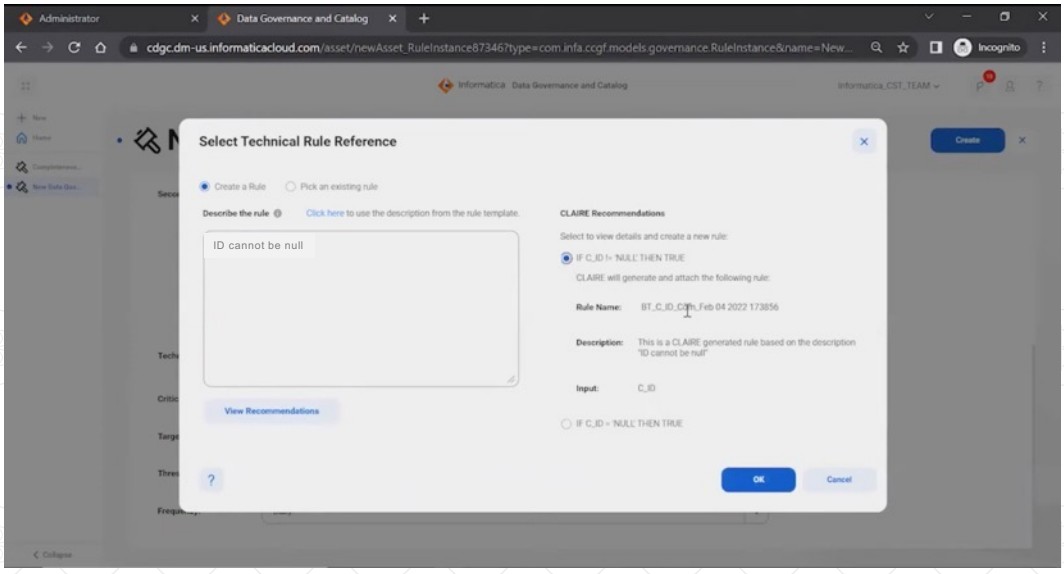
SBV = Shared Business Vocabulary

Automated Data Quality Profiling – IBM Cloud Pak For Data & IBM Knowledge Catalog



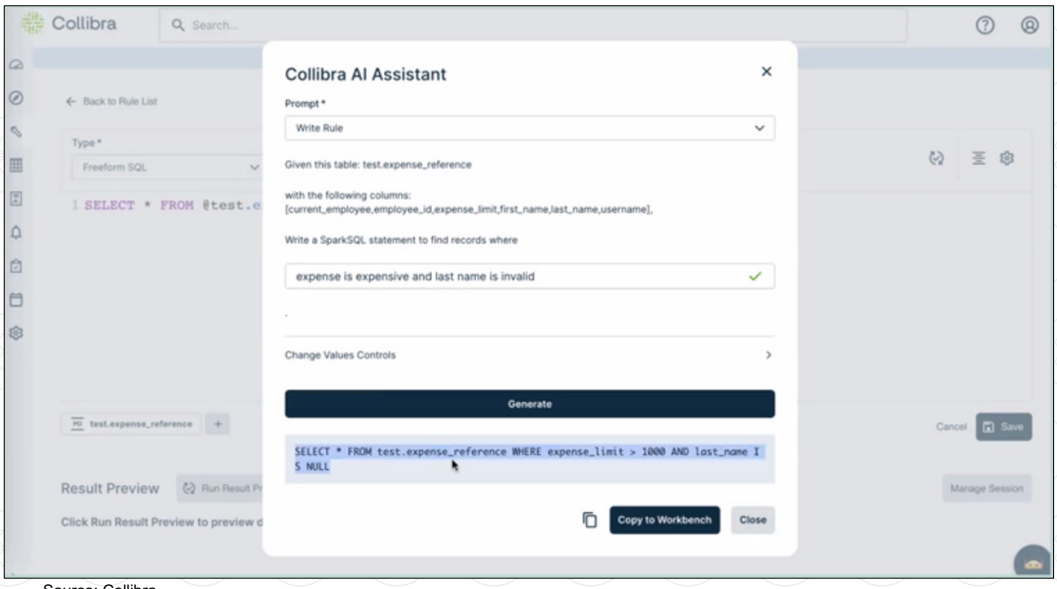
Source: IBM

Informatica Allows Data Quality Rules To Be Defined Using Natural Language And Its CLAIRE AI Engine Will Recommend A Data Quality Rule



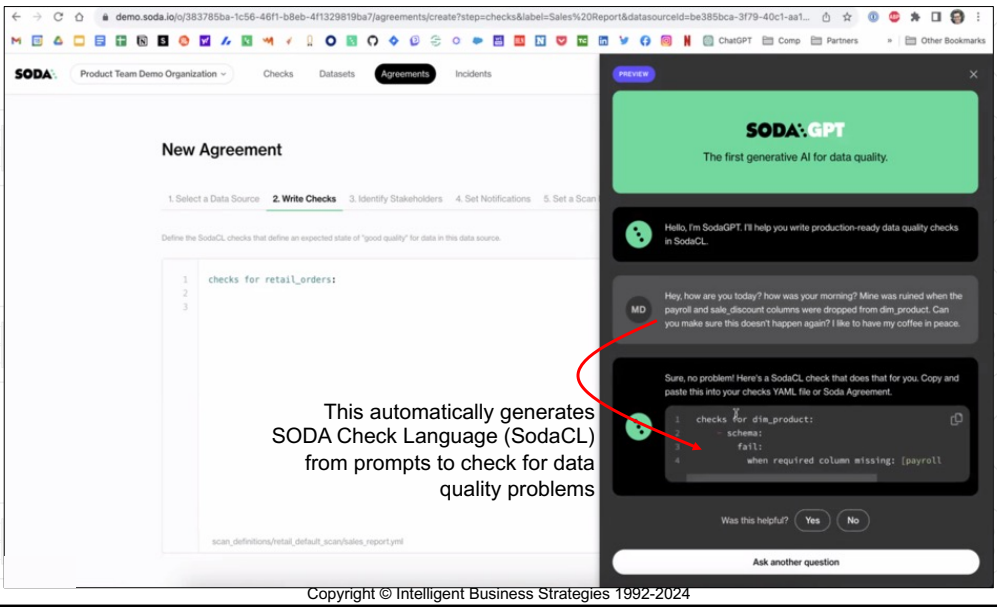
Source: Informatica

Generative AI in Data Governance – Defining Data Quality Data Validation Rules Using Collibra AI Assistant



Source: Collibra

Generative AI In Data Governance – SodaGPT Lets Business Analysts Create Data Quality Checks In Natural Language That Can Be Shared With Data Engineers To Fix Data

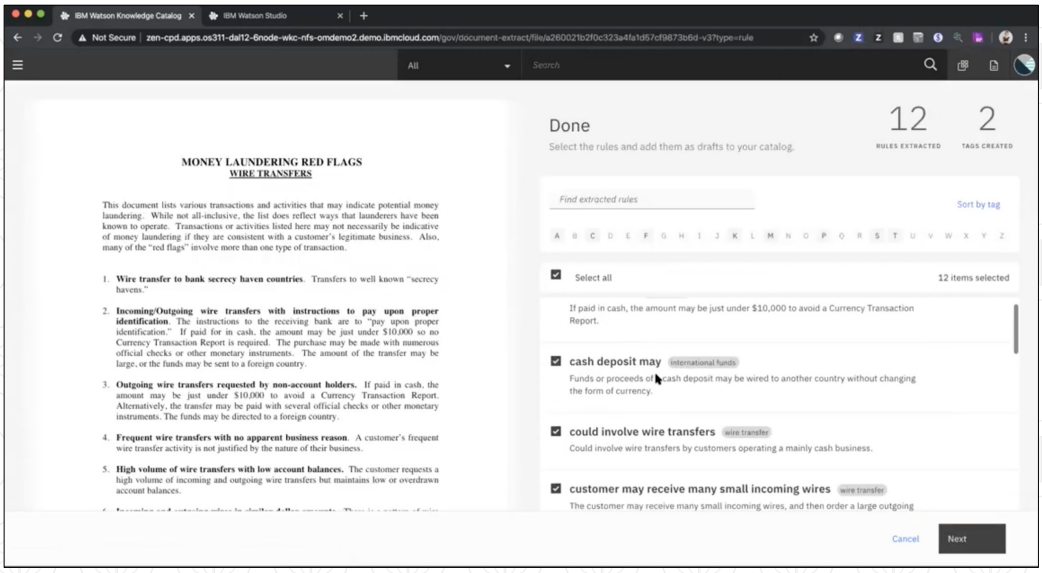


This automatically generates SODA Check Language (SodaCL) from prompts to check for data quality problems

Data Governance

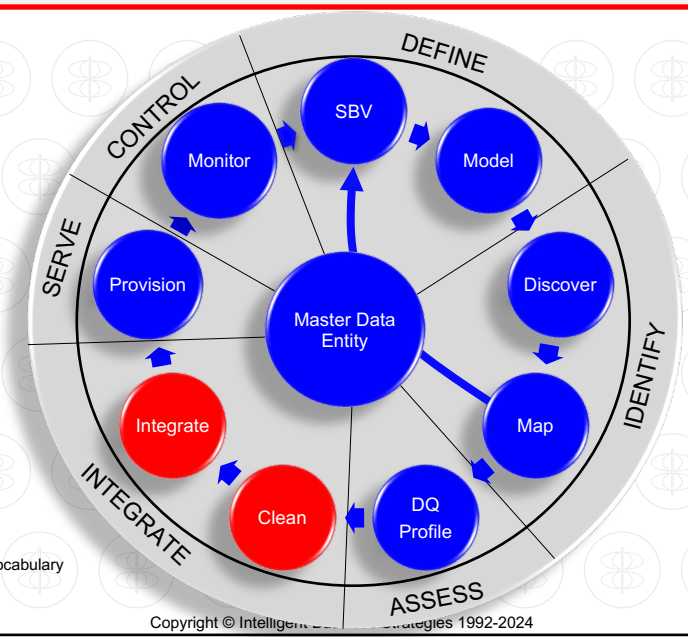
SodaCL is a low code abstract language built on SQL and Spark

Generative AI In Data Governance – It Is Also Possible To Automatically Extract And Generate Data Governance Rules From Documents, E.g., IBM Knowledge Catalog



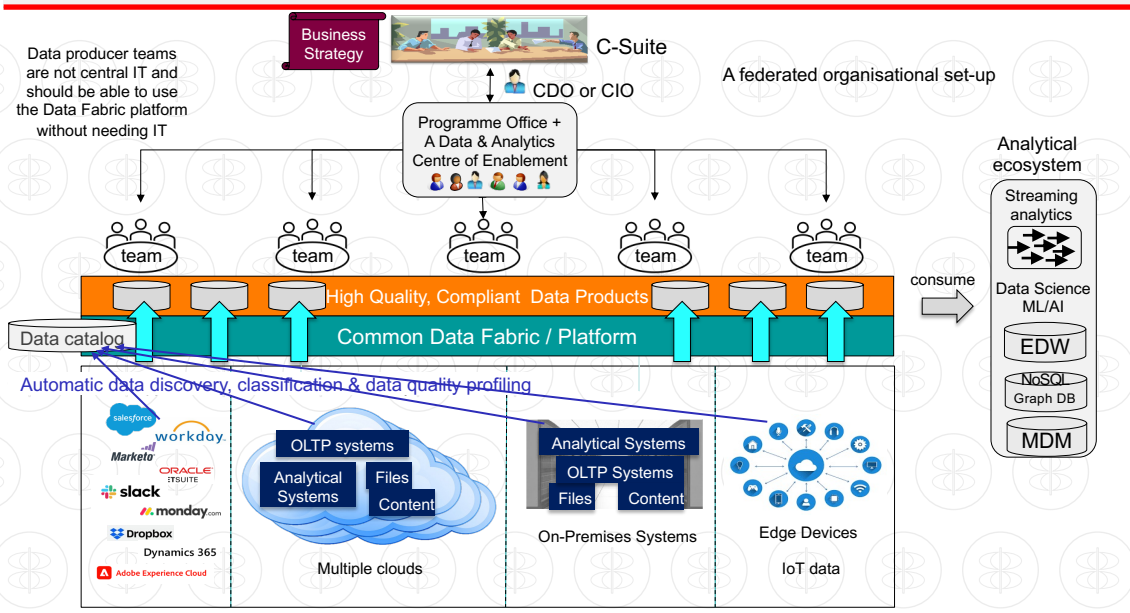
Data Governance

Where Are We? – Data Cleansing And Integration



SBV = Shared Business Vocabulary

Federated Operating Model – Business Strategy Aligned Teams Producing Data Products and ML Models Using Common Data Fabric and a Data Catalog



Integrated Master Data Can Be Considered As Master Data Products In A Data Mesh

- Customer
- Product
- Employee
- Supplier
- Asset
- Location
- Financial Chart of Accounts

These are master data products

The problem is they are cross domain

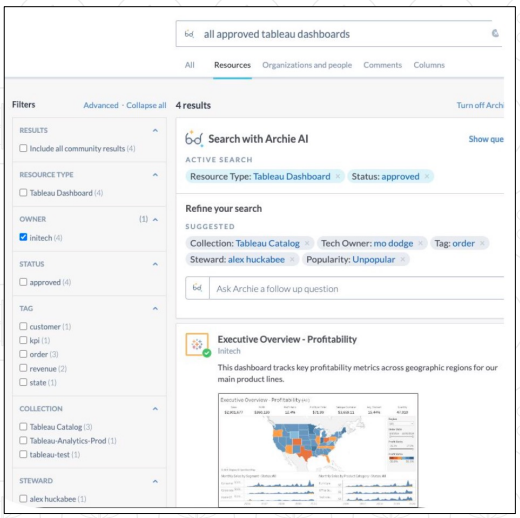
They are needed in a Data Mesh

1. Master data needs to be owned
2. Master data pipelines need to be developed
3. Master data products created, stored in an MDM system and published in a data marketplace

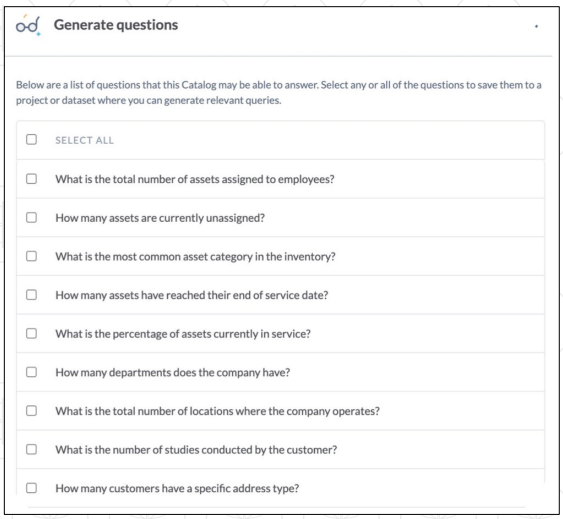
The same data fabric technology can be used to create master data products and all other data products

Prompt-Based Search and AI-Assisted Question Generation in Data Catalogs - Product Example: Data.world

AI-Assisted Search



Automated AI-Assisted Question Generation

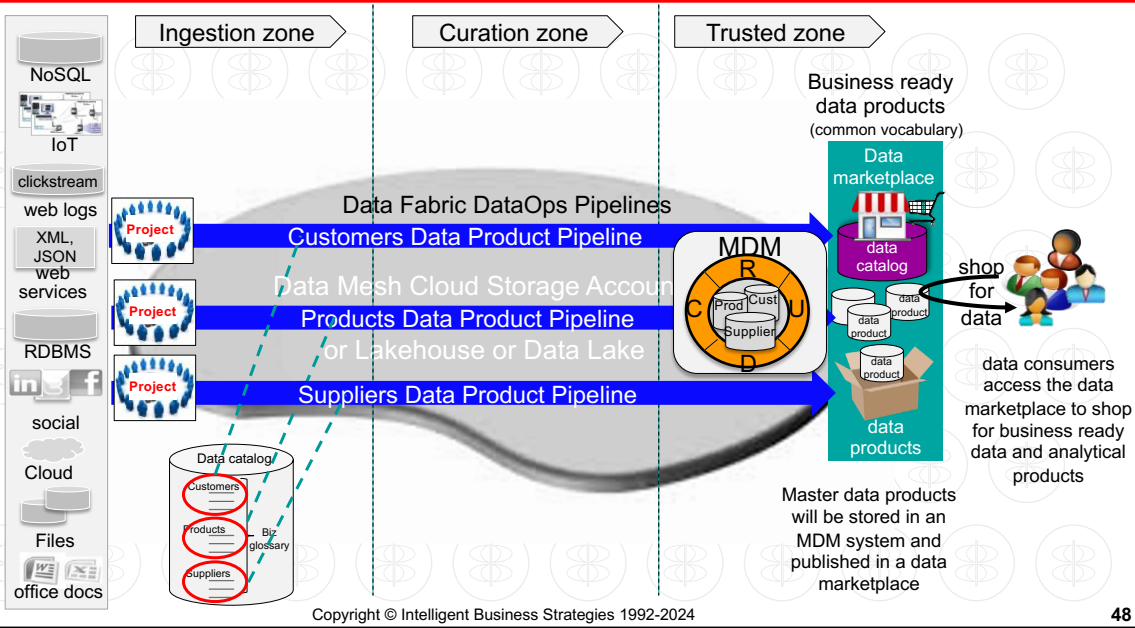


Source: data.world

data.world now has generative AI bots called Eurika bots

47

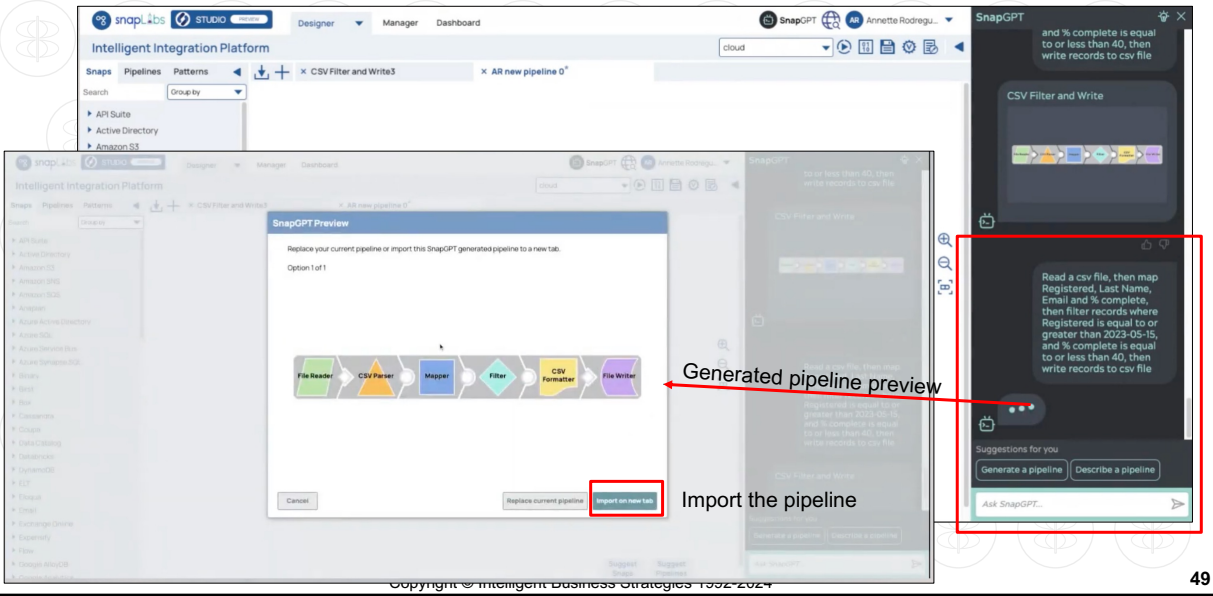
Use Data Fabric To Create Data Products To Publish In A Data Marketplace – Create A Pipeline That Produces a Master Data Product For Master Data Each Entity Your Glossary



48

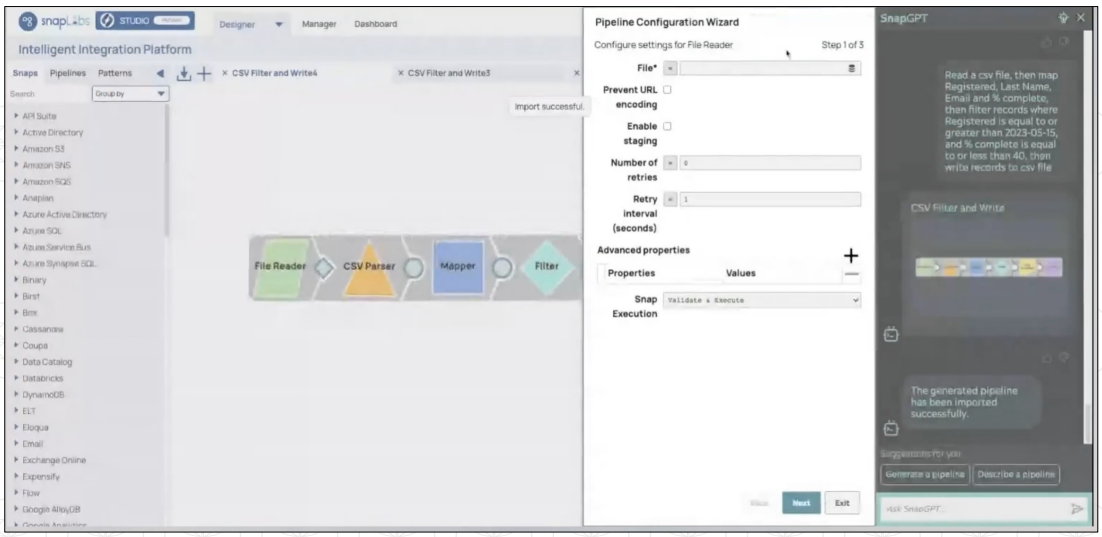
Generative AI In Data Engineering – SnapLogic SnapGPT Prompt Based Data Engineering

Data engineering



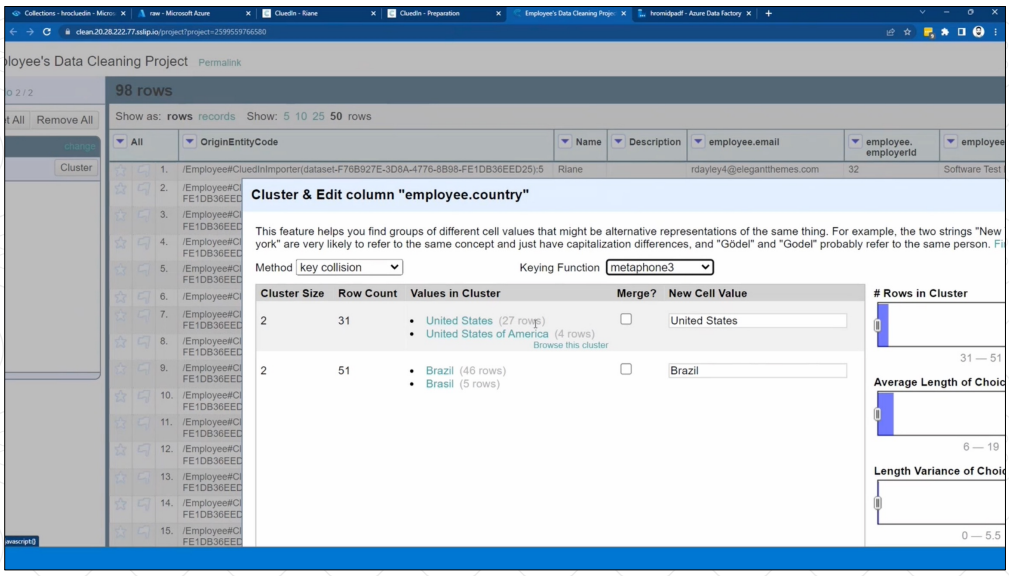
Generative AI Prompt Based Data Engineering – SnapLogic SnapGPT Pipeline Configuration Wizard

Data engineering



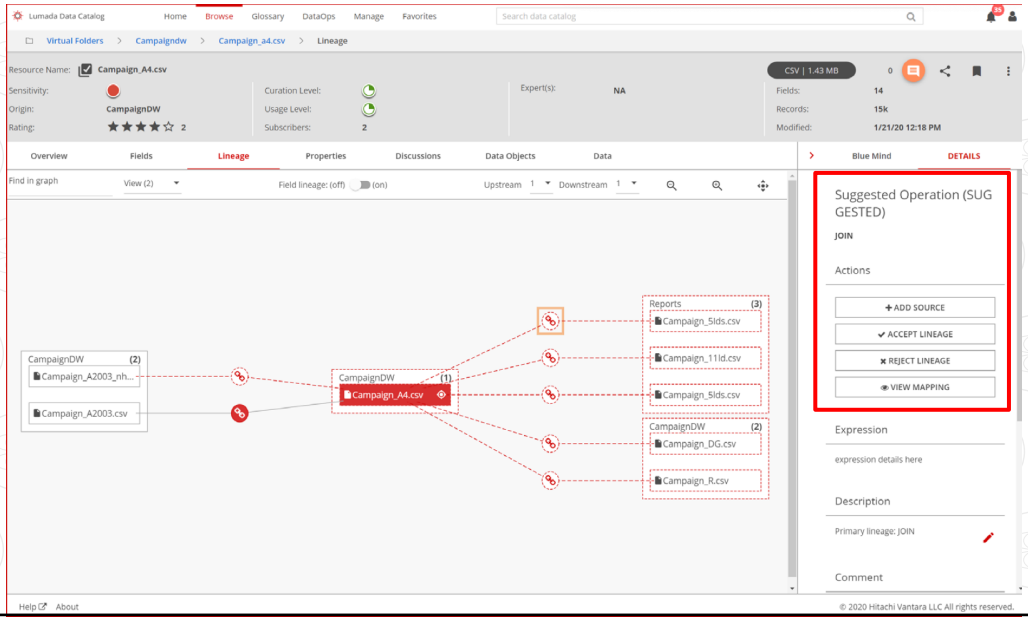
Source: SnapLogic

Data Cleansing Is Important Before Attempting A Match – Product Example: CluedIn Cluster And Edit Uses ML To Find Similar Data To Standardise

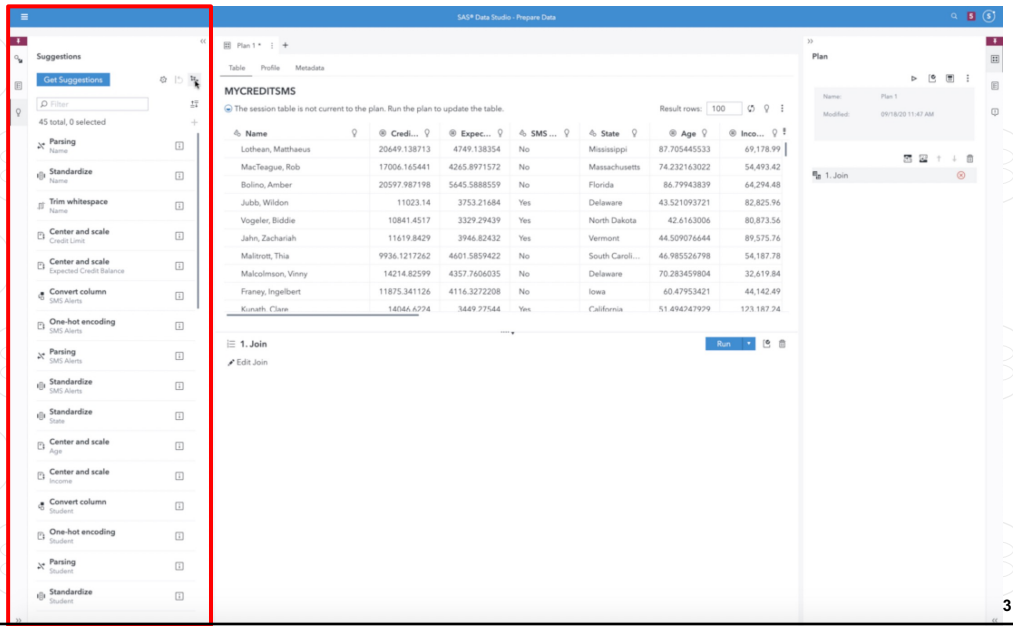


Source: CluedIn

Automatic Inferred Lineage Discovery Using Graph Analytics – Hitachi Vitara Lumada Data Catalog Suggested (Recommended) Joins



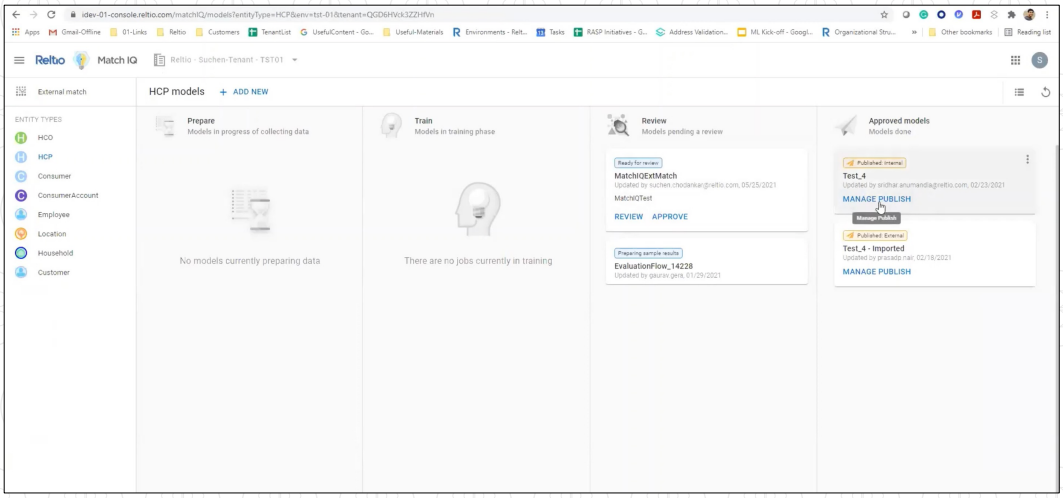
Guided Data Cleansing and Transformation
- SAS Data Studio Offers Suggestions (Recommendations)



3

Reltio ML-Powered MatchIQ
– Automate ML Driven Matching and Merge Based on Scoring

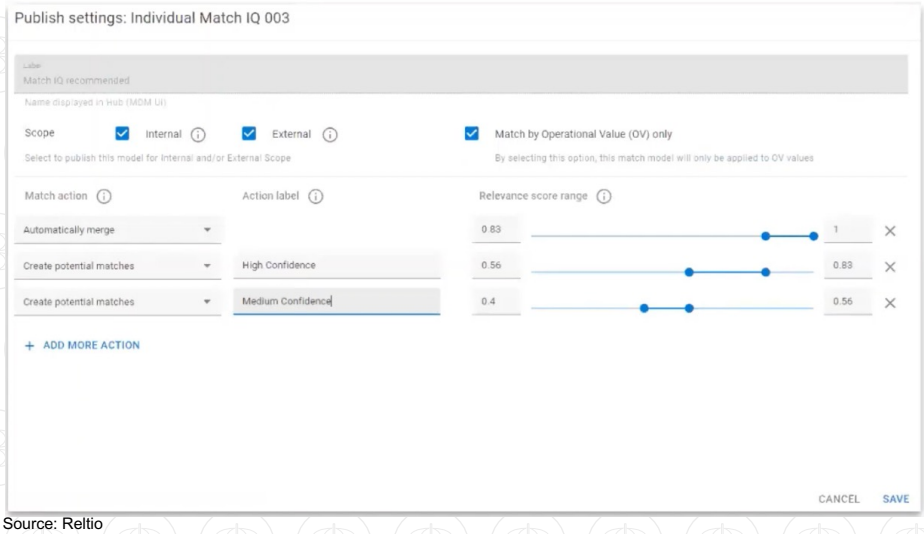
Prepare data, train and review an ML model, review results to determine accuracy, approve models



Source: Reltio

54

Reltio ML-Powered MatchIQ - Specify ML-score Thresholds to Trigger Automated Merge Or Other Matching Actions On MatchIQ Pairs

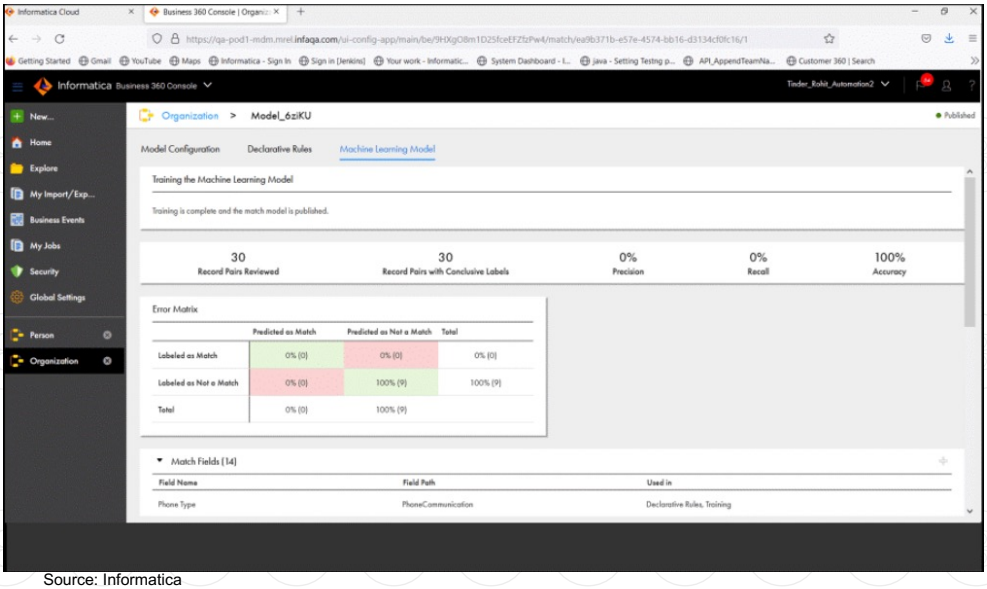


Source: Reltio

Review recommended matches in the potential match screen

55

Informatica MDM – Train ML Models for AI-Driven Match and Merge



Source: Informatica

56

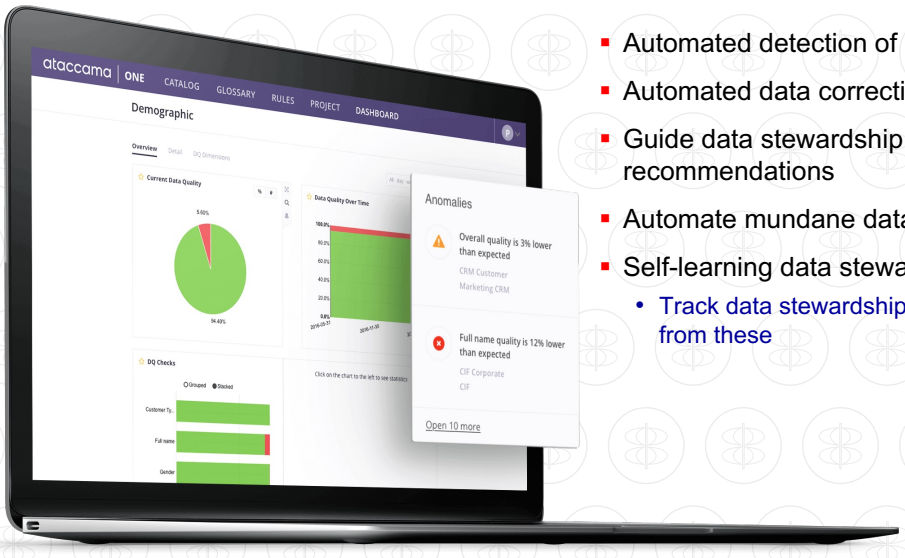
Batch Oriented Master Data Matching And Consolidation

- Matching
  - Matching is the process of comparing two records for points of similarity
  - If sufficient points of similarity are found to indicate that the two records are probably duplicates of each other, then those records are flagged for consolidation
  - The columns used for comparison purposes are called *match columns*
  - Each match column is based on one or more columns from the base object
  - Match columns are combined into *match rules* to determine the conditions under which two records are considered to be similar enough to consolidate
- Consolidation
  - After duplicate records have been identified in the match process, the consolidate process merges duplicate records into a single record

MDM System	Master ID	First Name	MN	Last Name	Address	City	State	Zip
	M-0001	Abel	Noel	Willan	161 Washington Ave.	Buffalo	NY	14263
Sales	SFA_ID	First Name	MN	Last Name	Address	City	State	Zip
	12345	Abel		Willan	161 Washington Ave.	Buffalo	NY	14263
Accounts	Cust_ID	First Name	MN	Last Name	Address	City	State	Zip
	502068	Abel	Noel	Willan	161 Washington Ave.	Buffalo	NY	14263
Marketing	Target_ID	First Name	MN	Last Name	Address	City	State	Zip
	willan05	Abel	N	Willan	Elm & Carlston Streets	Buffalo	NY	14263

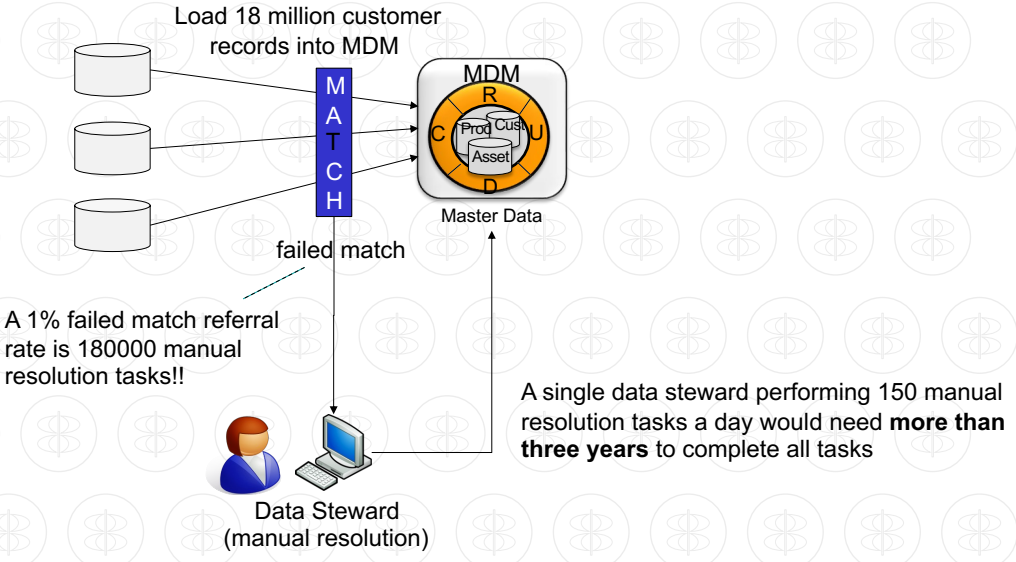
Source: Informatica

AI In Data Governance - AI Assisted Data Stewardship

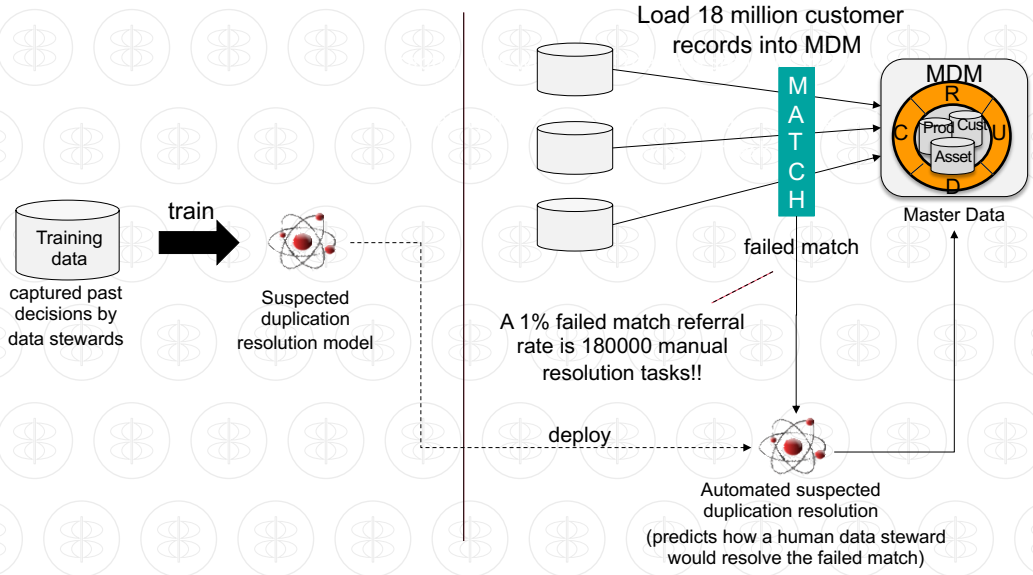


- Automated detection of poor quality data
- Automated data correction
- Guide data stewardship decisions using recommendations
- Automate mundane data stewardship tasks
- Self-learning data steward
  - Track data stewardship decisions and learn from these

Why Is Assisted Data Stewardship Needed In MDM? - Matching Anomaly Resolution

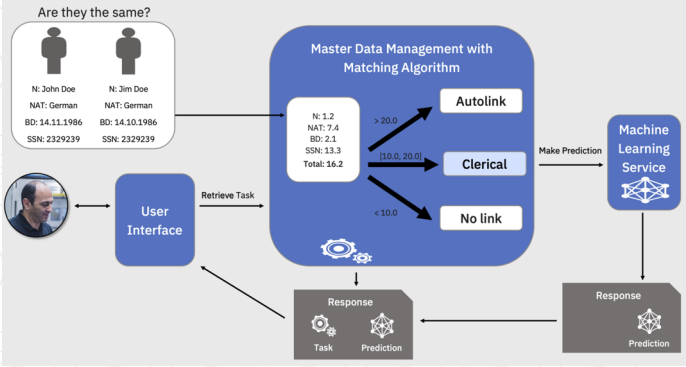


Using AI To Assist Data Stewards In Master Data Management Matching By Automating Suspected Duplicate Resolution

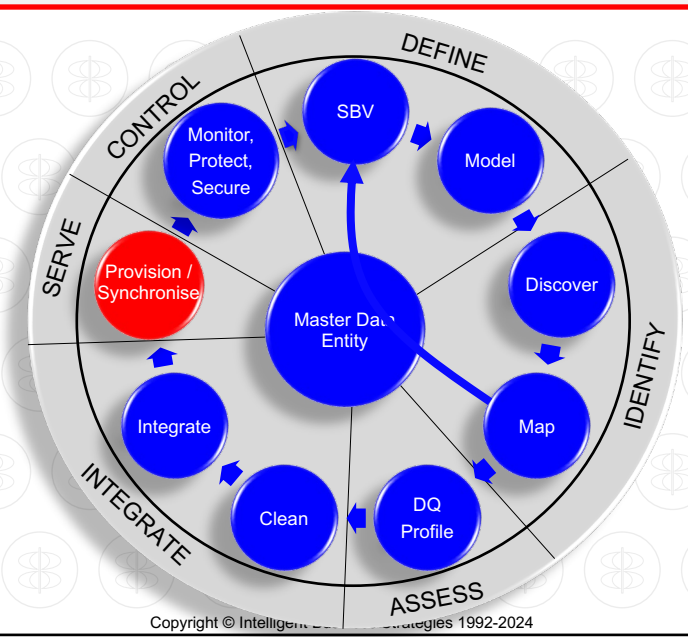


AI Assisted Data Stewardship – Phased Deployment Of AI

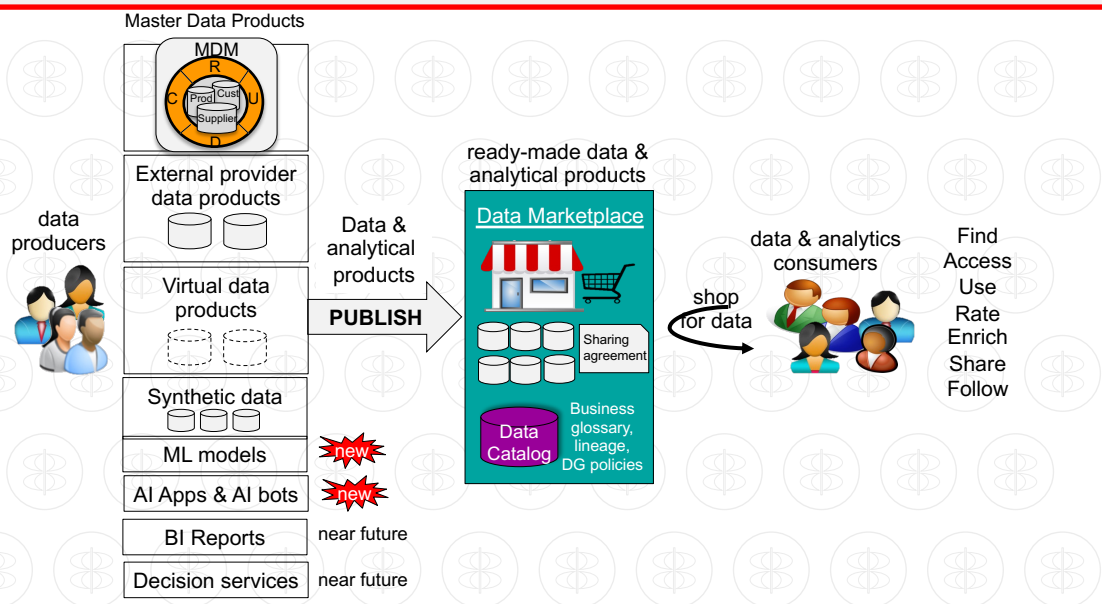
- Step 1
  - Introduce the trained model to recommend to the data stewards how to resolve the suspected duplicate
  - Data steward simply reviews and approves
- Step 2
  - Automate the process above a model prediction confidence level threshold (e.g. >95%) once data stewards trust it



Where Are We?
– Provisioning / Synchronising High Quality, Commonly Understood Master Data



Enterprise Data Marketplace Operations - Facilitate The Publishing, Sharing and Governance of Data and Analytical Products

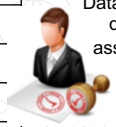


Key Elements In Enterprise Data Sharing – Governance Publishing Of Data Products

- Quality assurance and governance of data products is needed before they are published in a marketplace for consumption

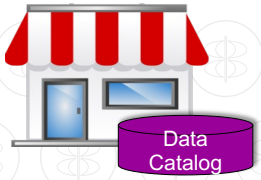
Business Glossary +
Data Lineage
+ Data Freshness
+ Data Owners

Data Quality score	✓
Common data names (business glossary)	✓
Full metadata lineage	✓
Access security policy defined	✓
Privacy policy defined	✓
Retention policy defined	✓
Version management	✓
Data sharing agreement defined	✓
Tagged by business objective	✓



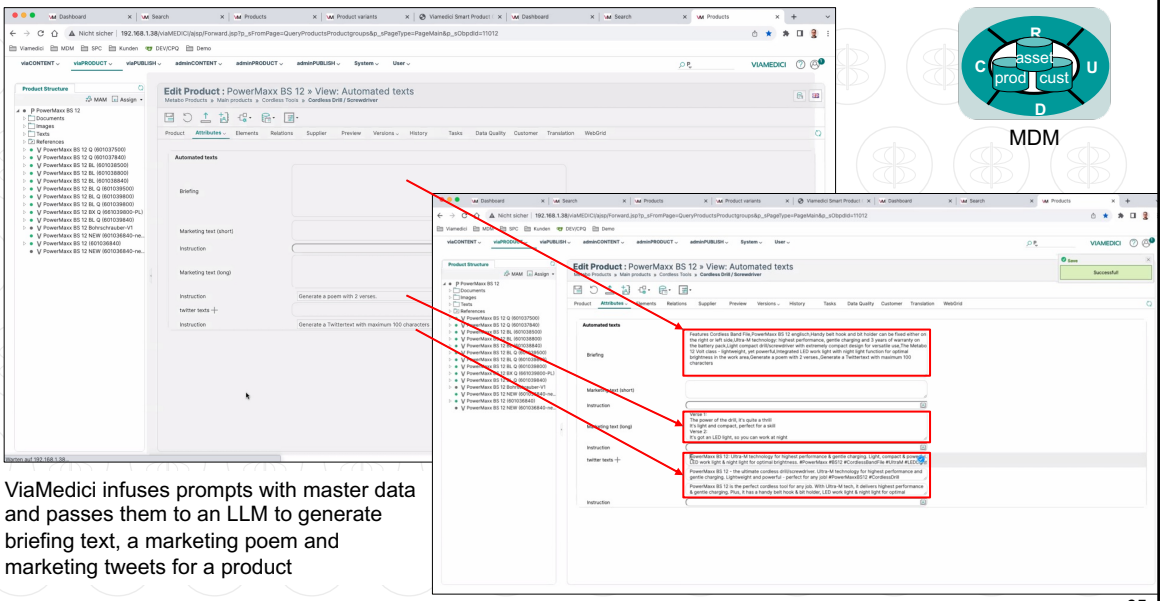
Data product quality assurance

Data product



Data marketplace

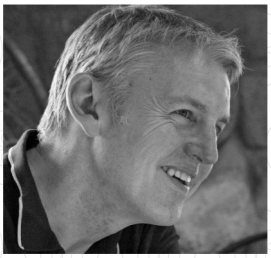
Generative AI In Master Data Management (MDM) – Product Example: ViaMedici Product MDM & AI-Powered PIM Text Generation



MDM

ViaMedici infuses prompts with master data and passes them to an LLM to generate briefing text, a marketing poem and marketing tweets for a product

About Mike Ferguson



www.intelligentbusiness.biz
mferguson@intelligentbusiness.biz
@mikeferguson1
(+44) 1625 520700

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an independent IT industry analyst and consultant, he specialises in BI / analytics and data management. With over 40 years of IT experience, Mike has consulted for dozens of companies on BI/Analytics, data strategy, technology selection, enterprise architecture, and data management. Mike is also conference chairman of Big Data LDN, the largest data and analytics conference in Europe and a member of the EDM Council CDMC Executive Advisory Board. He has spoken at events all over the world and written numerous articles. Formerly he was a principal and co-founder of Codd and Date – the inventors of the Relational Model (which caused the birth of relational databases and the SQL language), and Chief Architect at Teradata on the Teradata DBMS He teaches popular master classes in Data Strategy, Data Warehouse Modernisation, Practical Guidelines for Implementing a Data Mesh, Big Data, How to Govern Data across a Distributed Data Landscape, Machine Learning and Advanced Analytics, and Embedded Analytics, Intelligent Apps and AI Automation

Thank You!