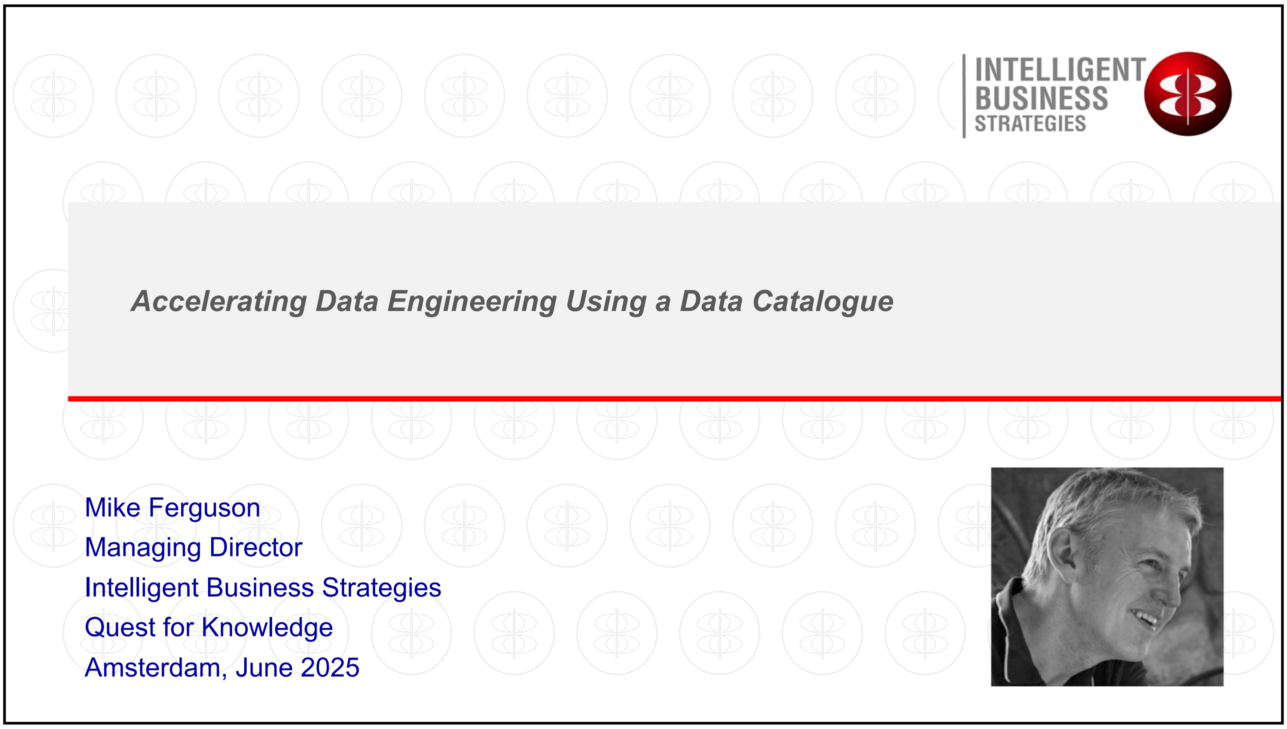
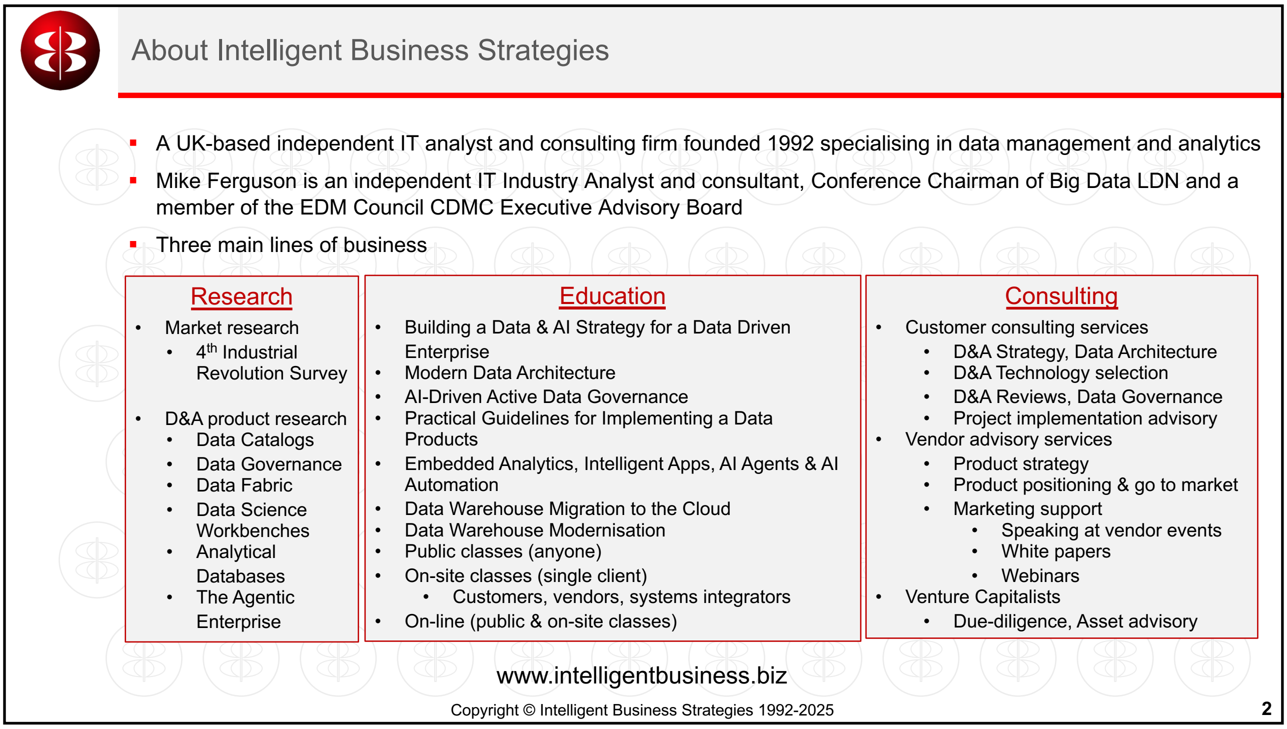


INTELLIGENT BUSINESS STRATEGIES

Accelerating Data Engineering Using a Data Catalogue

Mike Ferguson
Managing Director
Intelligent Business Strategies
Quest for Knowledge
Amsterdam, June 2025



About Intelligent Business Strategies

- A UK-based independent IT analyst and consulting firm founded 1992 specialising in data management and analytics
- Mike Ferguson is an independent IT Industry Analyst and consultant, Conference Chairman of Big Data LDN and a member of the EDM Council CDMC Executive Advisory Board
- Three main lines of business

Research

- Market research
  - 4th Industrial Revolution Survey
- D&A product research
  - Data Catalogs
  - Data Governance
  - Data Fabric
  - Data Science Workbenches
  - Analytical Databases
  - The Agentic Enterprise

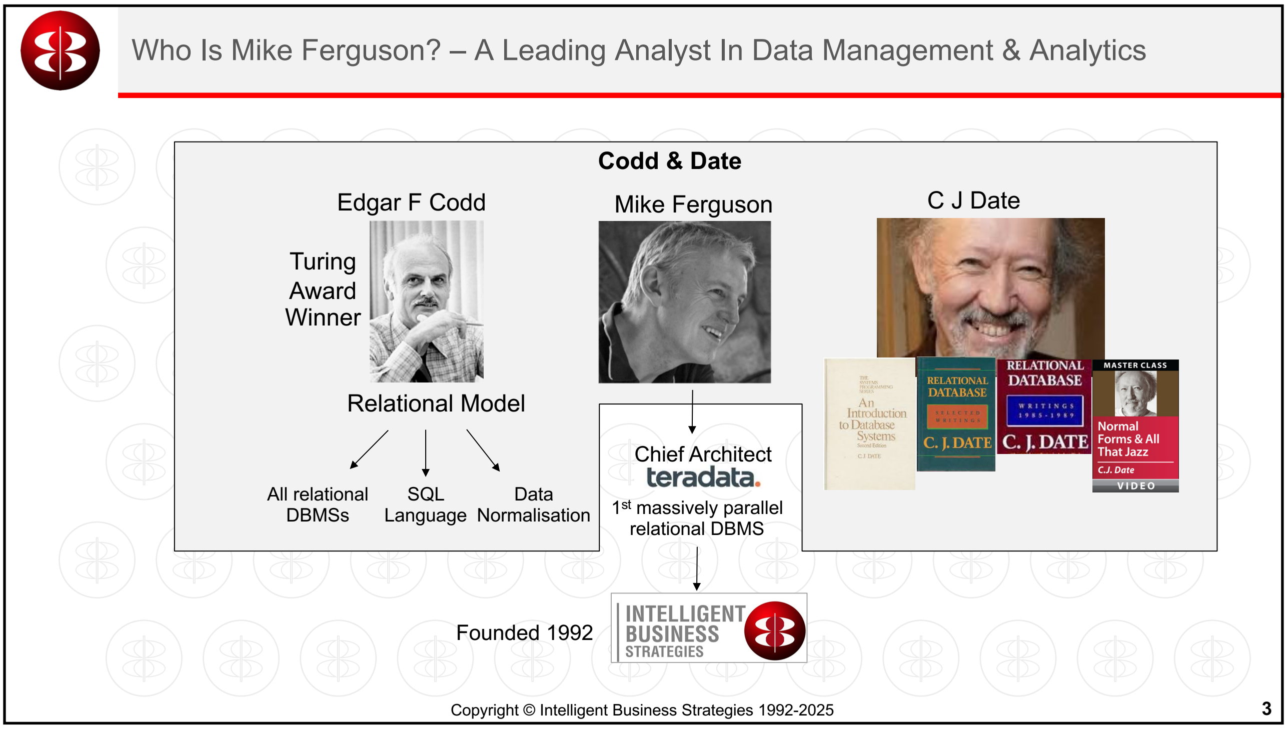
Education

- Building a Data & AI Strategy for a Data Driven Enterprise
- Modern Data Architecture
- AI-Driven Active Data Governance
- Practical Guidelines for Implementing a Data Products
- Embedded Analytics, Intelligent Apps, AI Agents & AI Automation
- Data Warehouse Migration to the Cloud
- Data Warehouse Modernisation
- Public classes (anyone)
- On-site classes (single client)
  - Customers, vendors, systems integrators
- On-line (public & on-site classes)

Consulting

- Customer consulting services
  - D&A Strategy, Data Architecture
  - D&A Technology selection
  - D&A Reviews, Data Governance
  - Project implementation advisory
- Vendor advisory services
  - Product strategy
  - Product positioning & go to market
  - Marketing support
    - Speaking at vendor events
    - White papers
    - Webinars
- Venture Capitalists
  - Due-diligence, Asset advisory

www.intelligentbusiness.biz



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

Founded 1992

INTELLIGENT BUSINESS STRATEGIES

C J Date

An Introduction to Database Systems
C. J. DATE
RELATIONAL DATABASE
SELECTED WRITINGS
C. J. DATE
RELATIONAL DATABASE
WRITINGS 1985-1989
C. J. DATE
MASTER CLASS
Normal Forms & All That Jazz
C.J. Date
VIDEO

Copyright © Intelligent Business Strategies 1992-2025 3

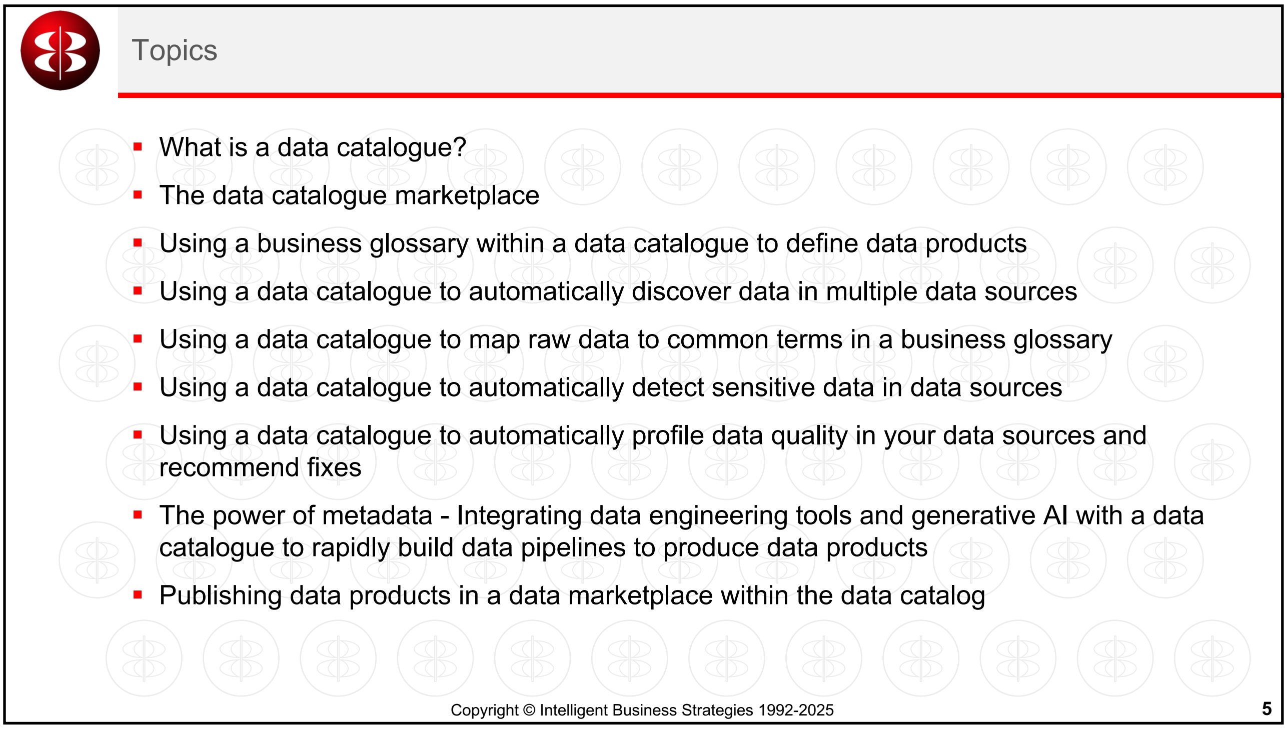
Mike Ferguson Is Europe's Leading Industry Analyst / Consultant In Data Management & Analytics And Conference Chairman Of Big Data LDN

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

- 20000 delegates
- 200+ vendors
- 14 theatres
- 350+ speakers

It is 6 x size of Gartner's D&A conference

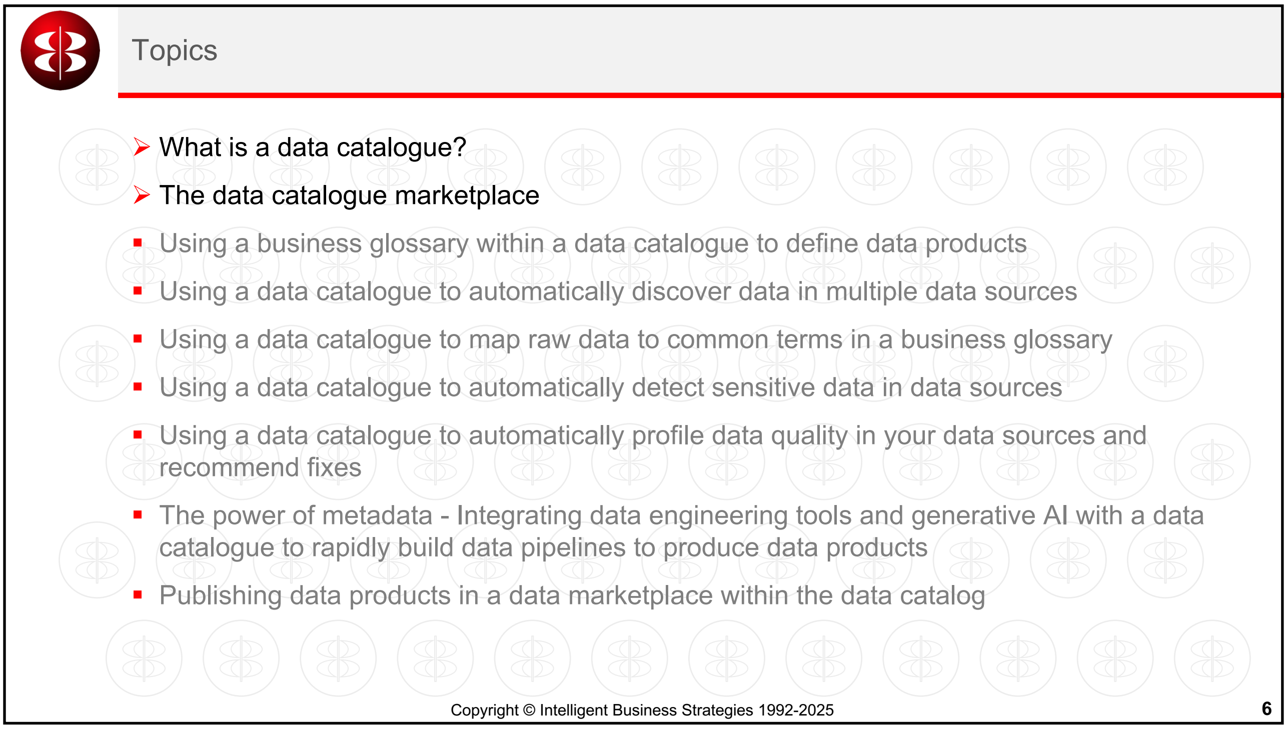
Copyright © Intelligent Business Strategies 1992-2025 4



Topics

- What is a data catalogue?
- The data catalogue marketplace
- Using a business glossary within a data catalogue to define data products
- Using a data catalogue to automatically discover data in multiple data sources
- Using a data catalogue to map raw data to common terms in a business glossary
- Using a data catalogue to automatically detect sensitive data in data sources
- Using a data catalogue to automatically profile data quality in your data sources and recommend fixes
- The power of metadata - Integrating data engineering tools and generative AI with a data catalogue to rapidly build data pipelines to produce data products
- Publishing data products in a data marketplace within the data catalog

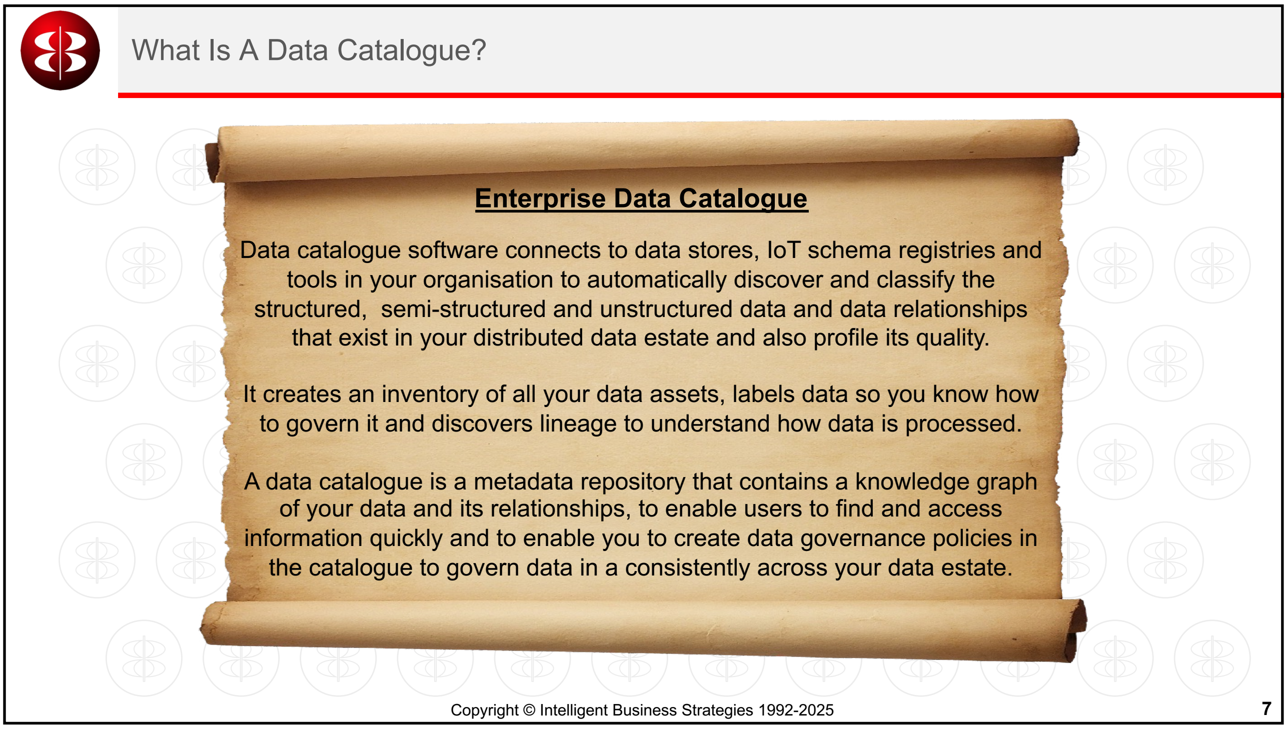
5



Topics

- ➢ What is a data catalogue?
- ➢ The data catalogue marketplace
- Using a business glossary within a data catalogue to define data products
- Using a data catalogue to automatically discover data in multiple data sources
- Using a data catalogue to map raw data to common terms in a business glossary
- Using a data catalogue to automatically detect sensitive data in data sources
- Using a data catalogue to automatically profile data quality in your data sources and recommend fixes
- The power of metadata - Integrating data engineering tools and generative AI with a data catalogue to rapidly build data pipelines to produce data products
- Publishing data products in a data marketplace within the data catalog

6



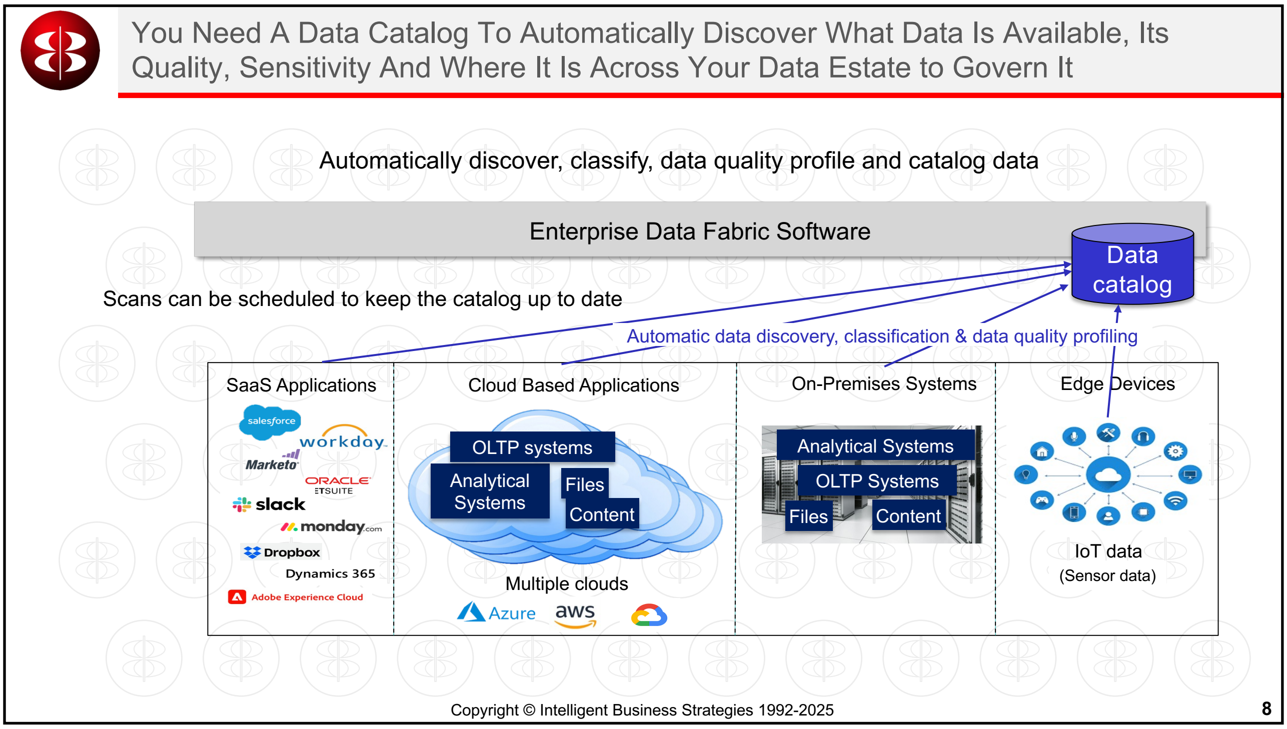
What Is A Data Catalogue?

Enterprise Data Catalogue

Data catalogue software connects to data stores, IoT schema registries and tools in your organisation to automatically discover and classify the structured, semi-structured and unstructured data and data relationships that exist in your distributed data estate and also profile its quality.

It creates an inventory of all your data assets, labels data so you know how to govern it and discovers lineage to understand how data is processed.

A data catalogue is a metadata repository that contains a knowledge graph of your data and its relationships, to enable users to find and access information quickly and to enable you to create data governance policies in the catalogue to govern data in a consistently across your data estate.



You Need A Data Catalog To Automatically Discover What Data Is Available, Its Quality, Sensitivity And Where It Is Across Your Data Estate to Govern It

Automatically discover, classify, data quality profile and catalog data

Enterprise Data Fabric Software

Data catalog

Scans can be scheduled to keep the catalog up to date

Automatic data discovery, classification & data quality profiling

SaaS Applications

Cloud Based Applications

OLTP systems

Analytical Systems

Files

Content

Multiple clouds

On-Premises Systems

Analytical Systems

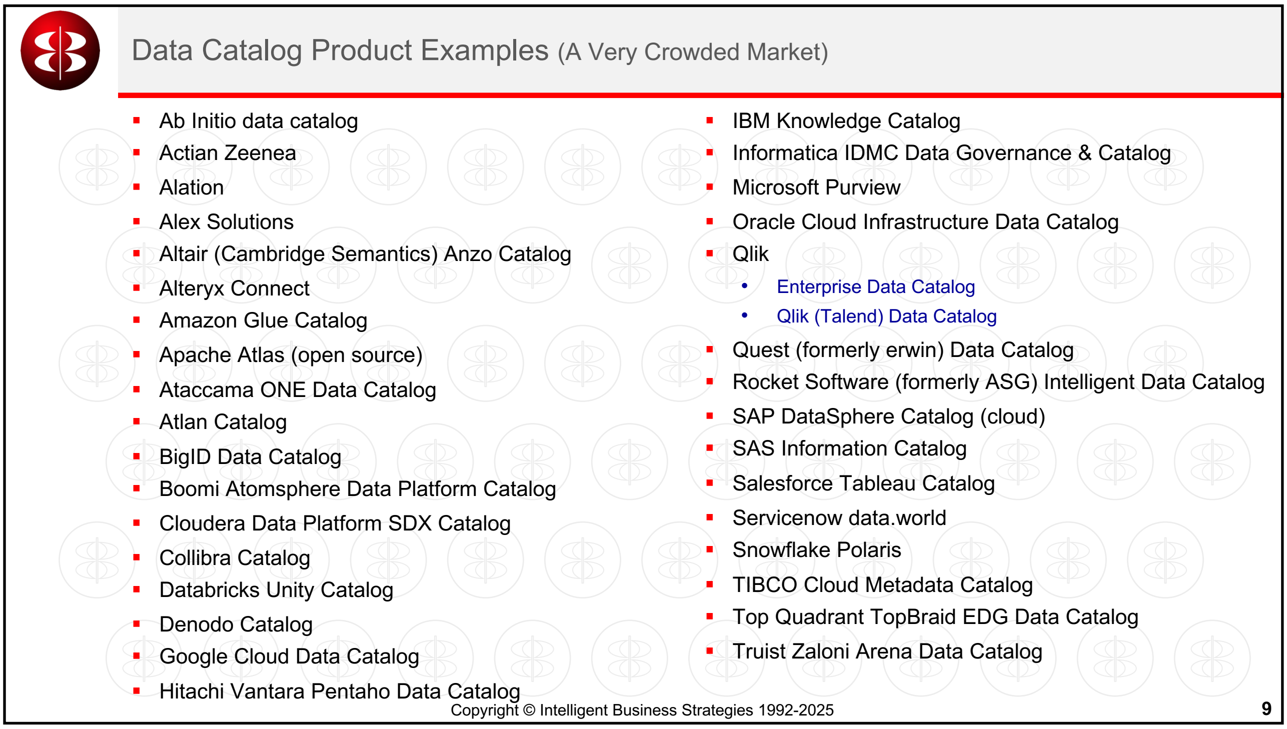
OLTP Systems

Files

Content

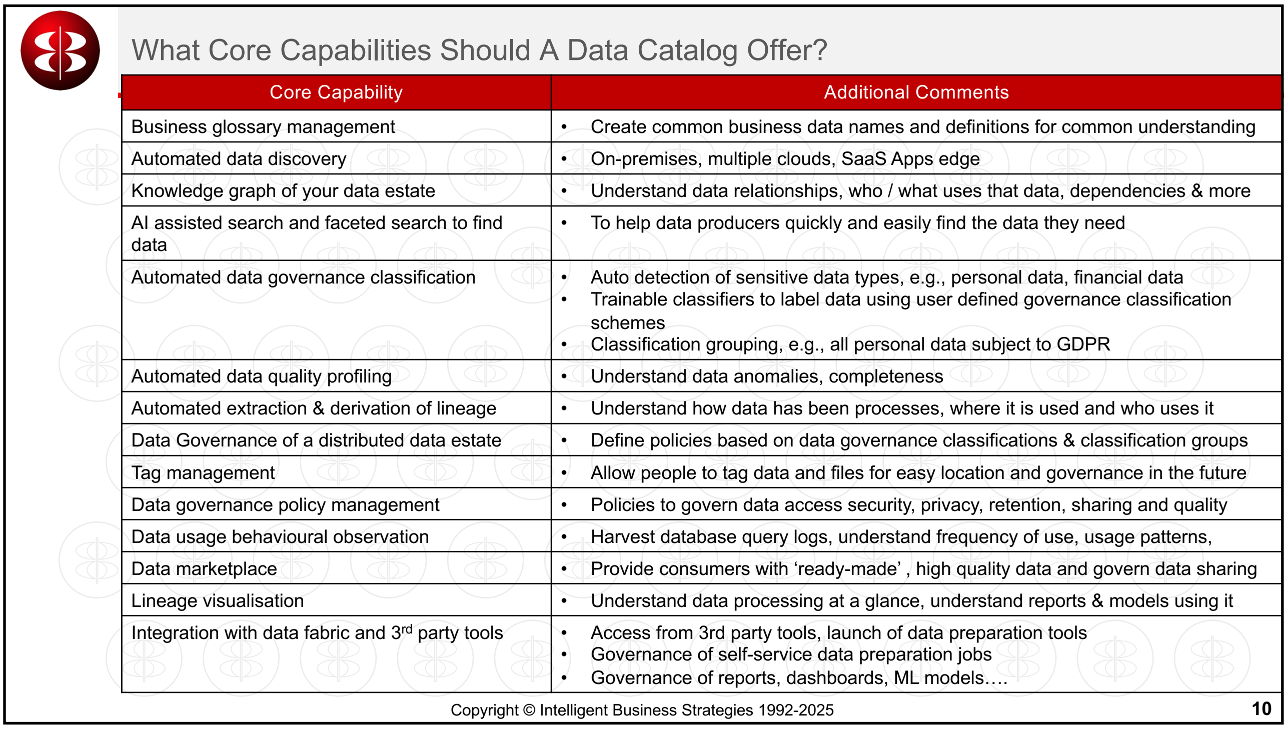
Edge Devices

IoT data
(Sensor data)



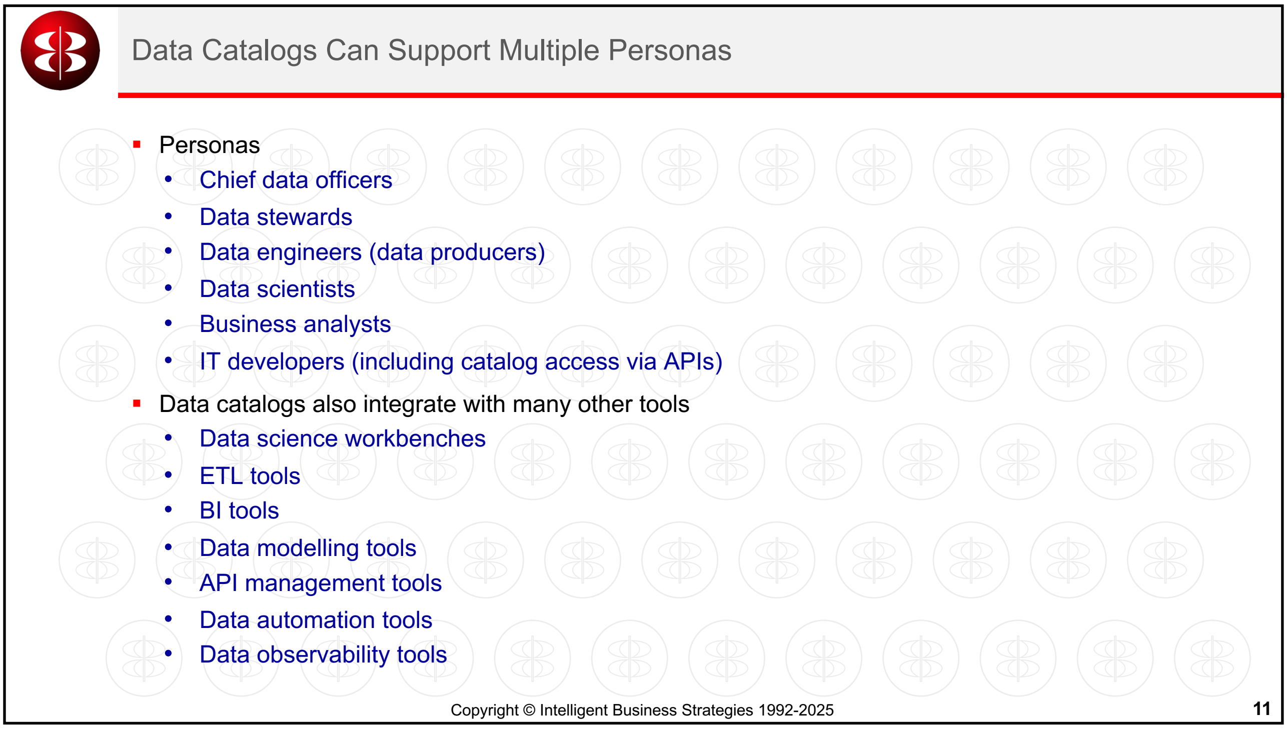
Data Catalog Product Examples (A Very Crowded Market)

- Ab Initio data catalog
- Actian Zeenea
- Alation
- Alex Solutions
- Altair (Cambridge Semantics) Anzo Catalog
- Alteryx Connect
- Amazon Glue Catalog
- Apache Atlas (open source)
- Ataccama ONE Data Catalog
- Atlan Catalog
- BigID Data Catalog
- Boomi Atomsphere Data Platform Catalog
- Cloudera Data Platform SDX Catalog
- Collibra Catalog
- Databricks Unity Catalog
- Denodo Catalog
- Google Cloud Data Catalog
- Hitachi Vantara Pentaho Data Catalog
- IBM Knowledge Catalog
- Informatica IDMC Data Governance & Catalog
- Microsoft Purview
- Oracle Cloud Infrastructure Data Catalog
- Qlik
  - Enterprise Data Catalog
  - Qlik (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
- Servicenow data.world
- Snowflake Polaris
- TIBCO Cloud Metadata Catalog
- Top Quadrant TopBraid EDG Data Catalog
- Truist Zaloni Arena Data Catalog



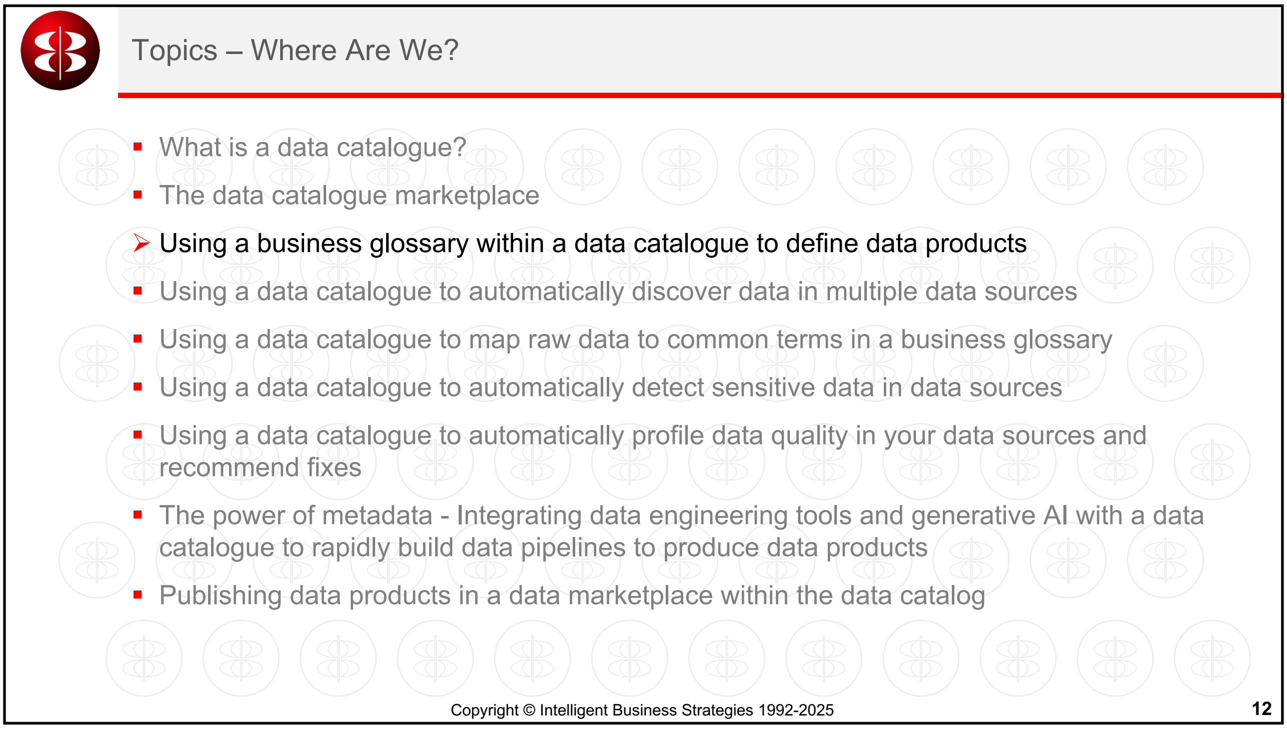
What Core Capabilities Should A Data Catalog Offer?

Core Capability	Additional Comments
Business glossary management	• Create common business data names and definitions for common understanding
Automated data discovery	• On-premises, multiple clouds, SaaS Apps edge
Knowledge graph of your data estate	• Understand data relationships, who / what uses that data, dependencies & more
AI assisted search and faceted search to find data	• To help data producers quickly and easily find the data they need
Automated data governance classification	• Auto detection of sensitive data types, e.g., personal data, financial data • Trainable classifiers to label data using user defined governance classification schemes • Classification grouping, e.g., all personal data subject to GDPR
Automated data quality profiling	• Understand data anomalies, completeness
Automated extraction & derivation of lineage	• Understand how data has been processes, where it is used and who uses it
Data Governance of a distributed data estate	• Define policies based on data governance classifications & classification groups
Tag management	• Allow people to tag data and files for easy location and governance in the future
Data governance policy management	• Policies to govern data access security, privacy, retention, sharing and quality
Data usage behavioural observation	• Harvest database query logs, understand frequency of use, usage patterns,
Data marketplace	• Provide consumers with 'ready-made' , high quality data and govern data sharing
Lineage visualisation	• Understand data processing at a glance, understand reports & models using it
Integration with data fabric and 3rd party tools	• Access from 3rd party tools, launch of data preparation tools • Governance of self-service data preparation jobs • Governance of reports, dashboards, ML models….



Data Catalogs Can Support Multiple Personas

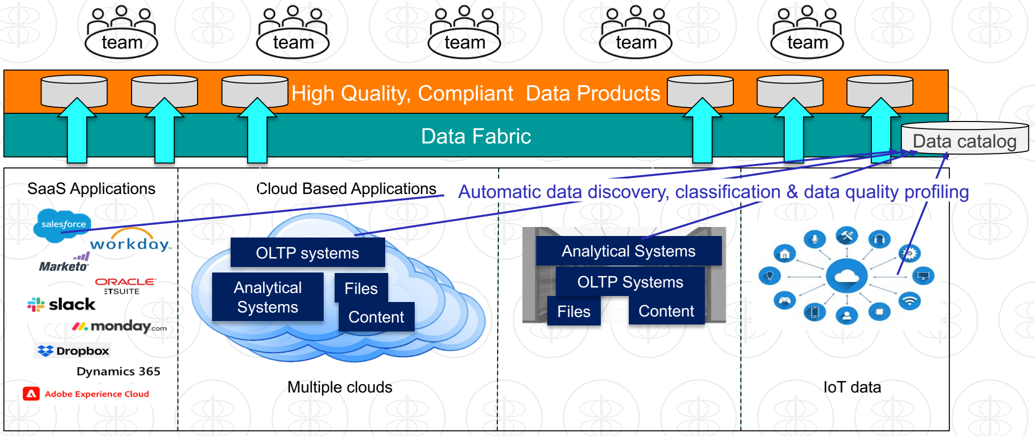
- Personas
  - Chief data officers
  - Data stewards
  - Data engineers (data producers)
  - Data scientists
  - Business analysts
  - IT developers (including catalog access via APIs)
- Data catalogs also integrate with many other tools
  - Data science workbenches
  - ETL tools
  - BI tools
  - Data modelling tools
  - API management tools
  - Data automation tools
  - Data observability tools



Topics – Where Are We?

- What is a data catalogue?
- The data catalogue marketplace
- ➤ Using a business glossary within a data catalogue to define data products
- Using a data catalogue to automatically discover data in multiple data sources
- Using a data catalogue to map raw data to common terms in a business glossary
- Using a data catalogue to automatically detect sensitive data in data sources
- Using a data catalogue to automatically profile data quality in your data sources and recommend fixes
- The power of metadata - Integrating data engineering tools and generative AI with a data catalogue to rapidly build data pipelines to produce data products
- Publishing data products in a data marketplace within the data catalog

Creating Reusable Data Products Is An Approach Fast Gaining Momentum As A Way of Incrementally Building Up A Secure And Compliant Data Foundation



team team team team team

High Quality, Compliant Data Products

Data Fabric

Data catalog

SaaS Applications

Cloud Based Applications

Automatic data discovery, classification & data quality profiling

OLTP systems

Analytical Systems

Files

Content

Multiple clouds

Analytical Systems

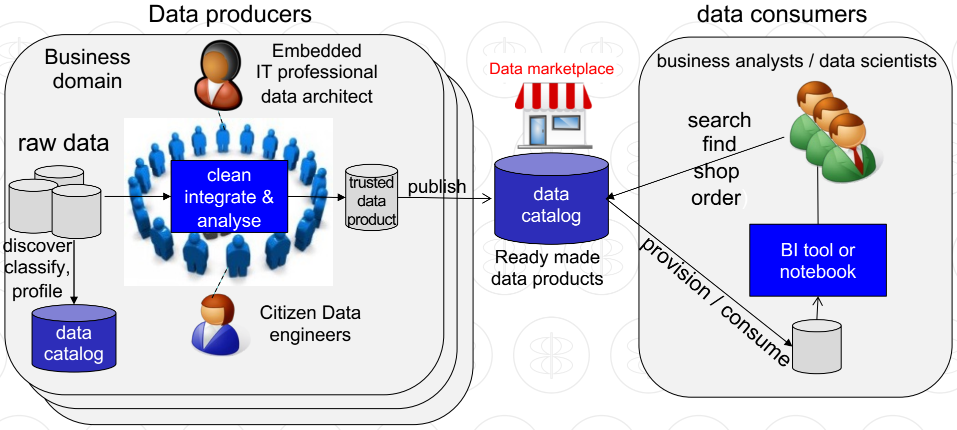
OLTP Systems

Files

Content

IoT data

Key Requirements – Organisational Implications If You Are To Become Data Driven – We Need Data Product Producers And Consumers



Data producers

data consumers

Business domain

Embedded IT professional data architect

raw data

clean integrate & analyse

trusted data product

publish

discover classify, profile

data catalog

Citizen Data engineers

Data marketplace

data catalog

Ready made data products

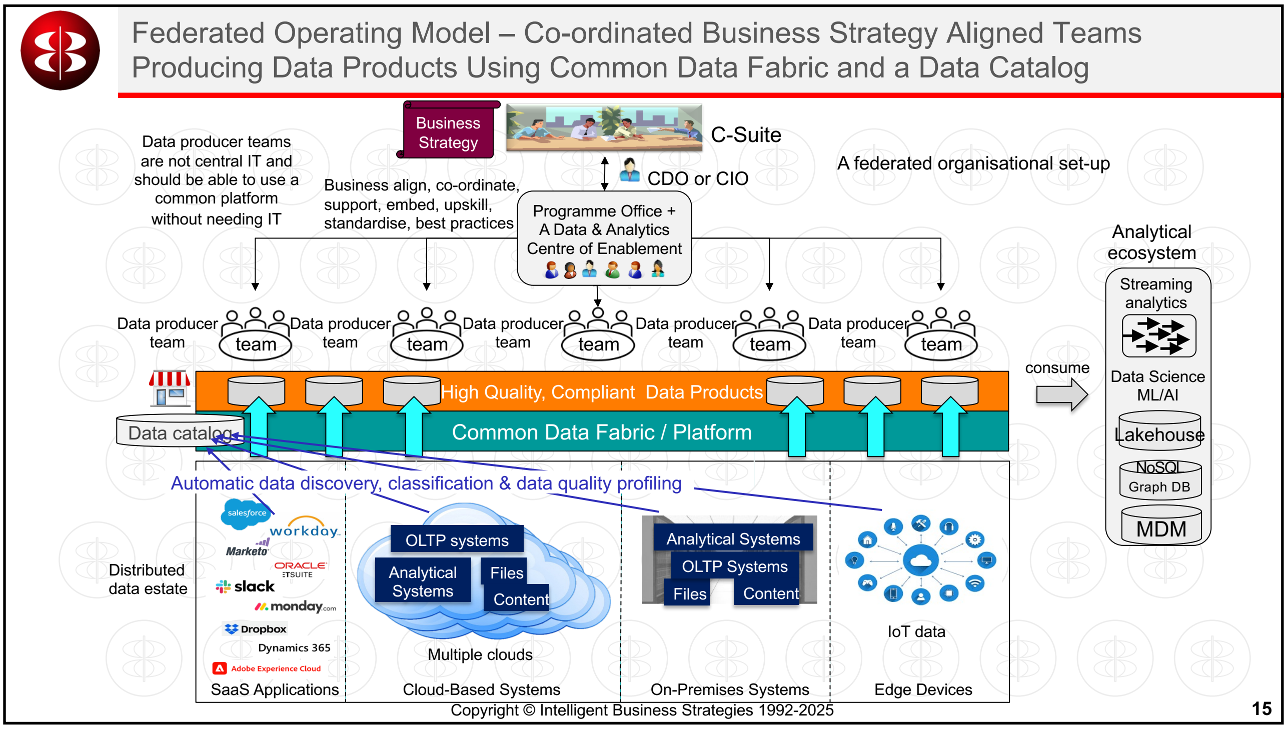
business analysts / data scientists

search find shop order

BI tool or notebook

provision / consume

- Need to make use of
  - A business glossary and data catalog
  - A collaborative approach to producing data and analytical products
  - A catalog / data marketplace to quickly find reusable trusted data products to drive business value



Federated Operating Model – Co-ordinated Business Strategy Aligned Teams Producing Data Products Using Common Data Fabric and a Data Catalog

Business Strategy

C-Suite

CDO or CIO

A federated organisational set-up

Data producer teams are not central IT and should be able to use a common platform without needing IT

Business align, co-ordinate, support, embed, upskill, standardise, best practices

Programme Office + A Data & Analytics Centre of Enablement

Data producer team — team

Data producer team — team

Data producer team — team

Data producer team — team

Data producer team — team

Data producer team — team

High Quality, Compliant Data Products

Common Data Fabric / Platform

Data catalog

consume

Analytical ecosystem

Streaming analytics

Data Science ML/AI

Lakehouse

NoSQL Graph DB

MDM

Automatic data discovery, classification & data quality profiling

Distributed data estate

salesforce, workday, Marketo, ORACLE NETSUITE, slack, monday.com, Dropbox, Dynamics 365, Adobe Experience Cloud

SaaS Applications

OLTP systems, Analytical Systems, Files, Content

Multiple clouds

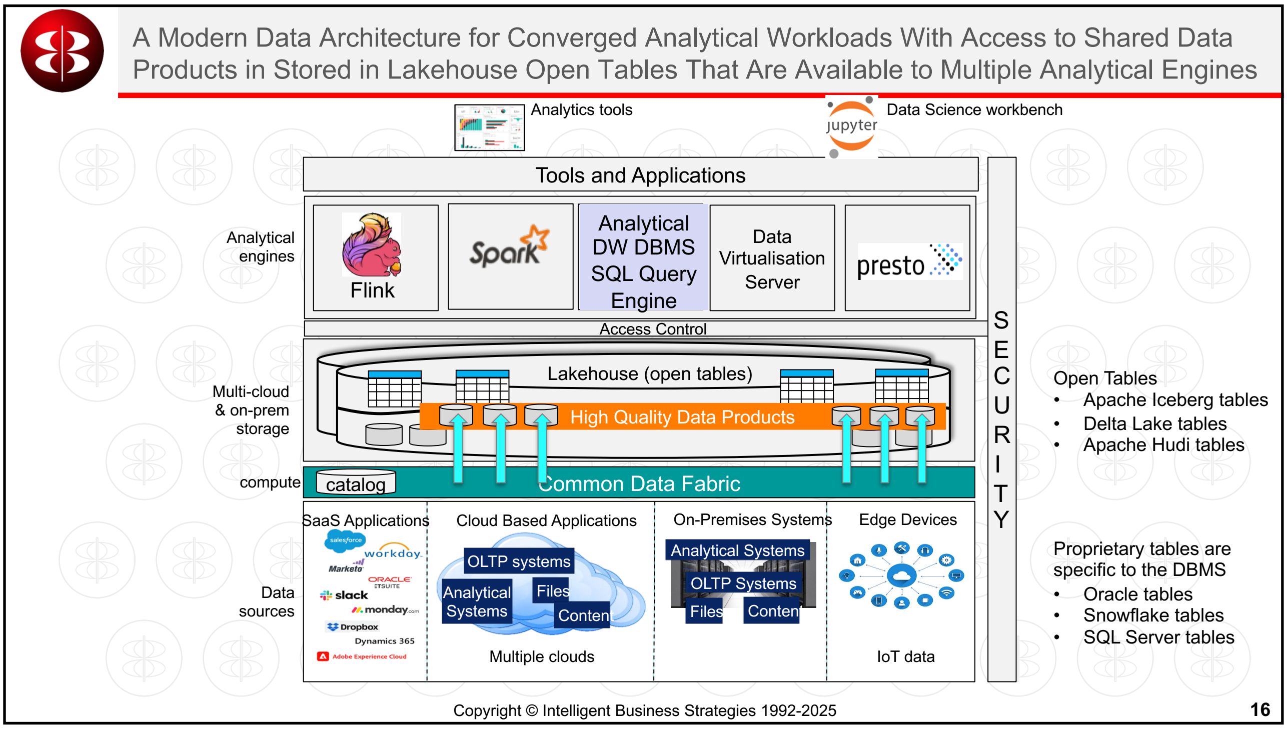
Cloud-Based Systems

Analytical Systems, OLTP Systems, Files, Content

On-Premises Systems

IoT data

Edge Devices



A Modern Data Architecture for Converged Analytical Workloads With Access to Shared Data Products in Stored in Lakehouse Open Tables That Are Available to Multiple Analytical Engines

Analytics tools

Data Science workbench

jupyter

Tools and Applications

Analytical engines

Flink

Spark

Analytical DW DBMS SQL Query Engine

Data Virtualisation Server

presto

Access Control

Multi-cloud & on-prem storage

Lakehouse (open tables)

High Quality Data Products

compute

catalog

Common Data Fabric

SECURITY

Data sources

SaaS Applications

salesforce, workday, Marketo, ORACLE NETSUITE, slack, monday.com, Dropbox, Dynamics 365, Adobe Experience Cloud

Cloud Based Applications

OLTP systems, Analytical Systems, Files, Conten

Multiple clouds

On-Premises Systems

Analytical Systems, OLTP Systems, Files, Conten

Edge Devices

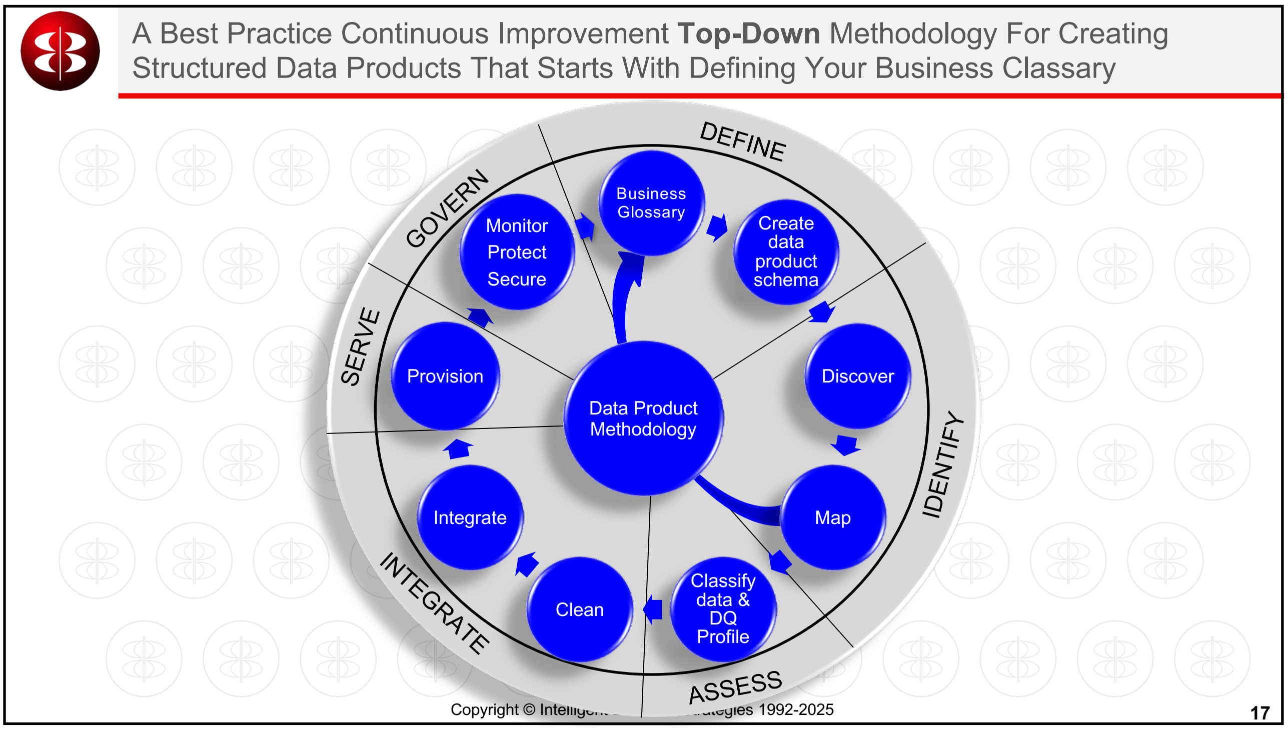
IoT data

Open Tables

- Apache Iceberg tables
- Delta Lake tables
- Apache Hudi tables

Proprietary tables are specific to the DBMS

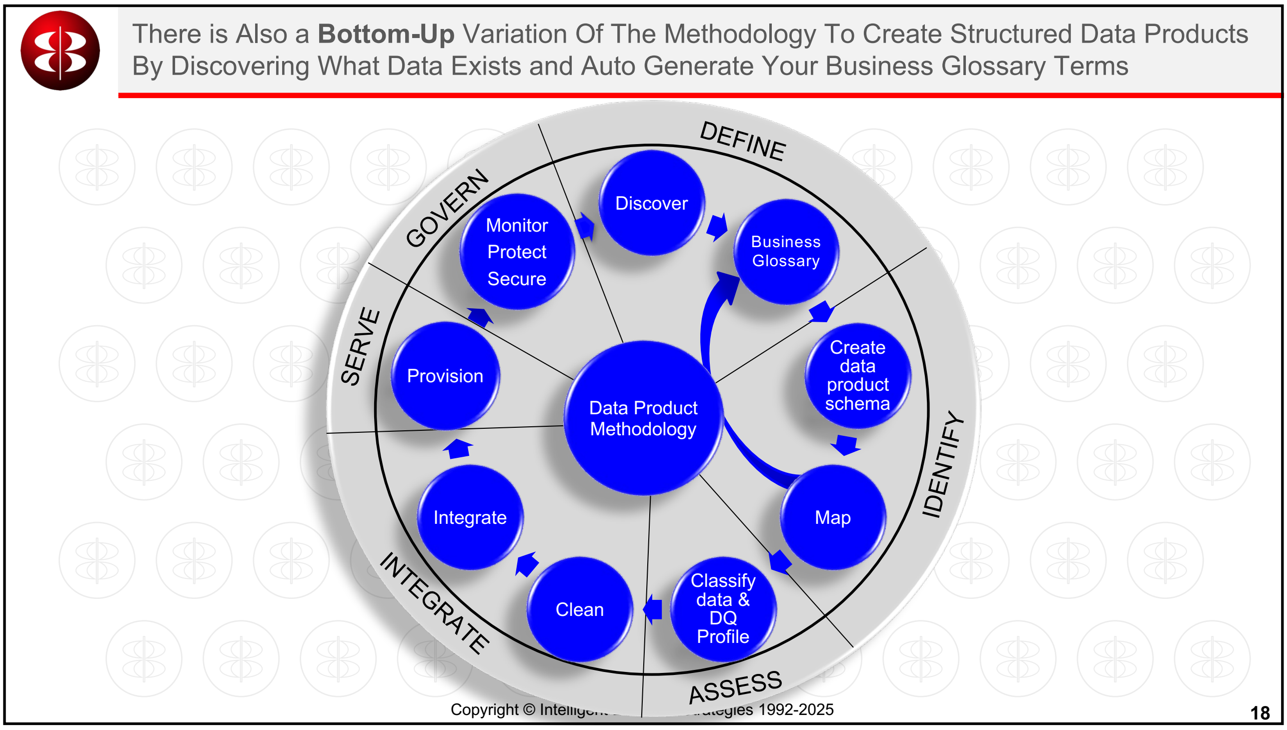
- Oracle tables
- Snowflake tables
- SQL Server tables



A Best Practice Continuous Improvement **Top-Down** Methodology For Creating Structured Data Products That Starts With Defining Your Business Classary

Copyright © Intelligen rategies 1992-2025

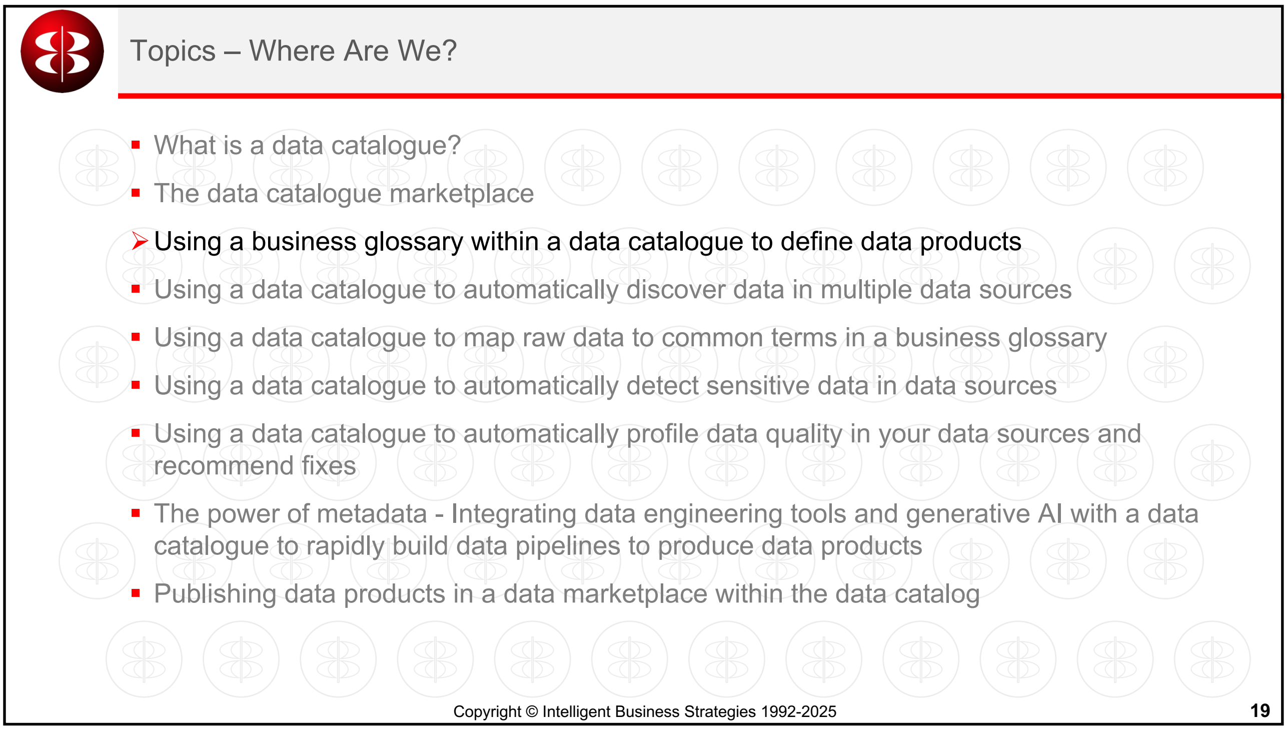
17



There is Also a **Bottom-Up** Variation Of The Methodology To Create Structured Data Products By Discovering What Data Exists and Auto Generate Your Business Glossary Terms

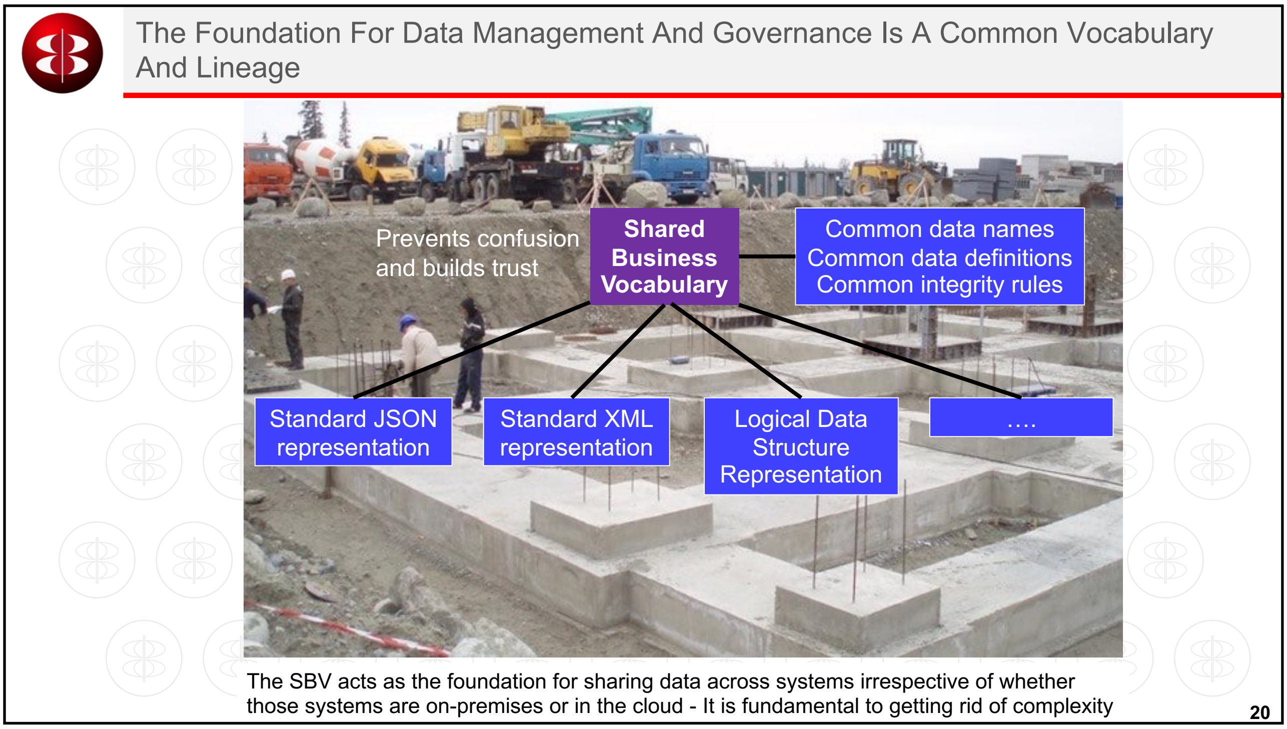
Copyright © Intelligen rategies 1992-2025

18



Topics – Where Are We?

- What is a data catalogue?
- The data catalogue marketplace
- ➢ Using a business glossary within a data catalogue to define data products
- Using a data catalogue to automatically discover data in multiple data sources
- Using a data catalogue to map raw data to common terms in a business glossary
- Using a data catalogue to automatically detect sensitive data in data sources
- Using a data catalogue to automatically profile data quality in your data sources and recommend fixes
- The power of metadata - Integrating data engineering tools and generative AI with a data catalogue to rapidly build data pipelines to produce data products
- Publishing data products in a data marketplace within the data catalog



The Foundation For Data Management And Governance Is A Common Vocabulary And Lineage

Prevents confusion and builds trust

Shared Business Vocabulary

Common data names
Common data definitions
Common integrity rules

Standard JSON representation

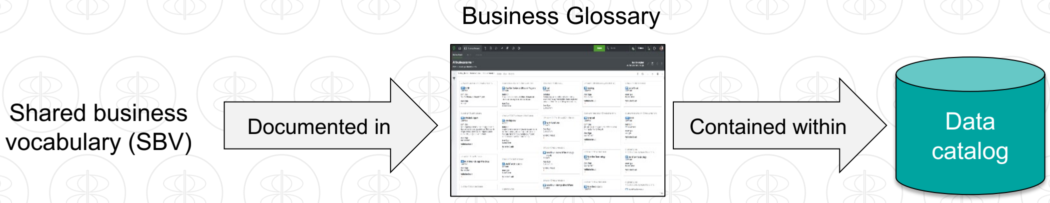
Standard XML representation

Logical Data Structure Representation

….

The SBV acts as the foundation for sharing data across systems irrespective of whether those systems are on-premises or in the cloud - It is fundamental to getting rid of complexity

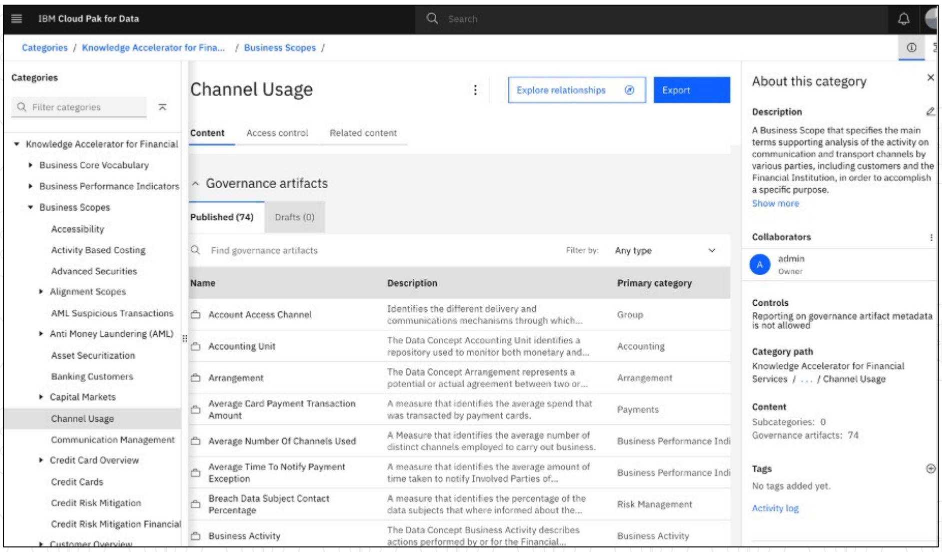
A Shared Business Vocabulary Is Defined In A Business Glossary Which Is Typically Part Of A Data Catalog



Options For Defining Your Own Shared Business Vocabulary

1. Adopt pre-built vocabularies as a standard
2. Leverage existing enterprise data models
   - If these models exist in your organisation, they are probably managed by data architects and may be widely adopted
3. Use a business data concept model to start with and incrementally define your own vocabulary from scratch
   - Use a data catalog to automatically discover data to help build out all attributes and to help automate the mapping of physical data assets to the terms in your glossary
4. Use Generative AI to automatically generated business glossary terms for the data it finds during automatic data discovery within a data catalog

Pre-Built Common Business Vocabularies Example - IBM Knowledge Accelerators - E.g. IBM Knowledge Accelerator for Financial Services



IBM Cloud Pak for Data

Search

Categories / Knowledge Accelerator for Fina... / Business Scopes /

Categories

Filter categories

- Knowledge Accelerator for Financial
  - Business Core Vocabulary
  - Business Performance Indicators
  - Business Scopes
    - Accessibility
    - Activity Based Costing
    - Advanced Securities
    - Alignment Scopes
    - AML Suspicious Transactions
    - Anti Money Laundering (AML)
    - Asset Securitization
    - Banking Customers
    - Capital Markets
    - Channel Usage
    - Communication Management
    - Credit Card Overview
    - Credit Cards
    - Credit Risk Mitigation
    - Credit Risk Mitigation Financial
    - Customer Overview

Channel Usage

Explore relationships | Export

Content | Access control | Related content

Governance artifacts

Published (74) | Drafts (0)

Find governance artifacts | Filter by: Any type

Name	Description	Primary category
Account Access Channel	Identifies the different delivery and communications mechanisms through which...	Group
Accounting Unit	The Data Concept Accounting Unit identifies a repository used to monitor both monetary and...	Accounting
Arrangement	The Data Concept Arrangement represents a potential or actual agreement between two or...	Arrangement
Average Card Payment Transaction Amount	A measure that identifies the average spend that was transacted by payment cards.	Payments
Average Number Of Channels Used	A Measure that identifies the average number of distinct channels employed to carry out business.	Business Performance Indi
Average Time To Notify Payment Exception	A measure that identifies the average amount of time taken to notify Involved Parties of...	Business Performance Indi
Breach Data Subject Contact Percentage	A measure that identifies the percentage of the data subjects that where informed about the...	Risk Management
Business Activity	The Data Concept Business Activity describes actions performed by or for the Financial...	Business Activity

About this category

Description

A Business Scope that specifies the main terms supporting analysis of the activity on communication and transport channels by various parties, including customers and the Financial Institution, in order to accomplish a specific purpose.

Show more

Collaborators

A admin
Owner

Controls

Reporting on governance artifact metadata is not allowed

Category path

Knowledge Accelerator for Financial Services / . . . / Channel Usage

Content

Subcategories: 0
Governance artifacts: 74

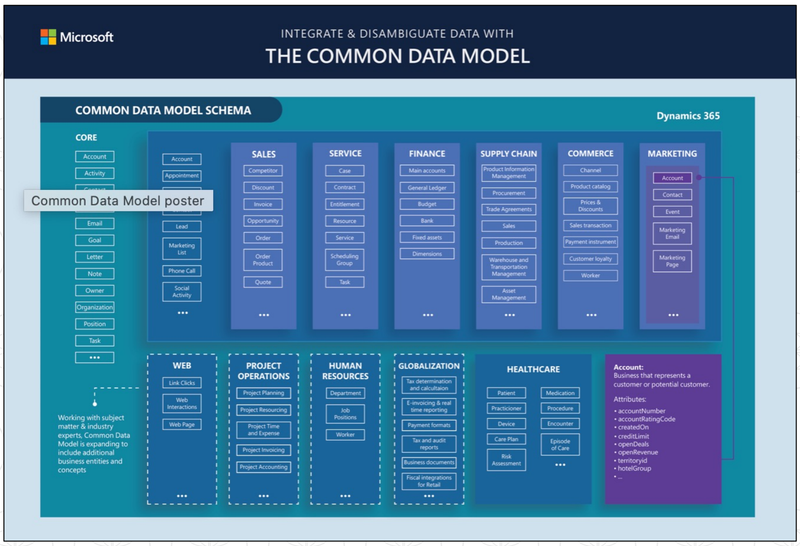
Tags

No tags added yet.

Activity log

Source: IBM

Pre-Built Vocabularies Example - Microsoft Common Data Model



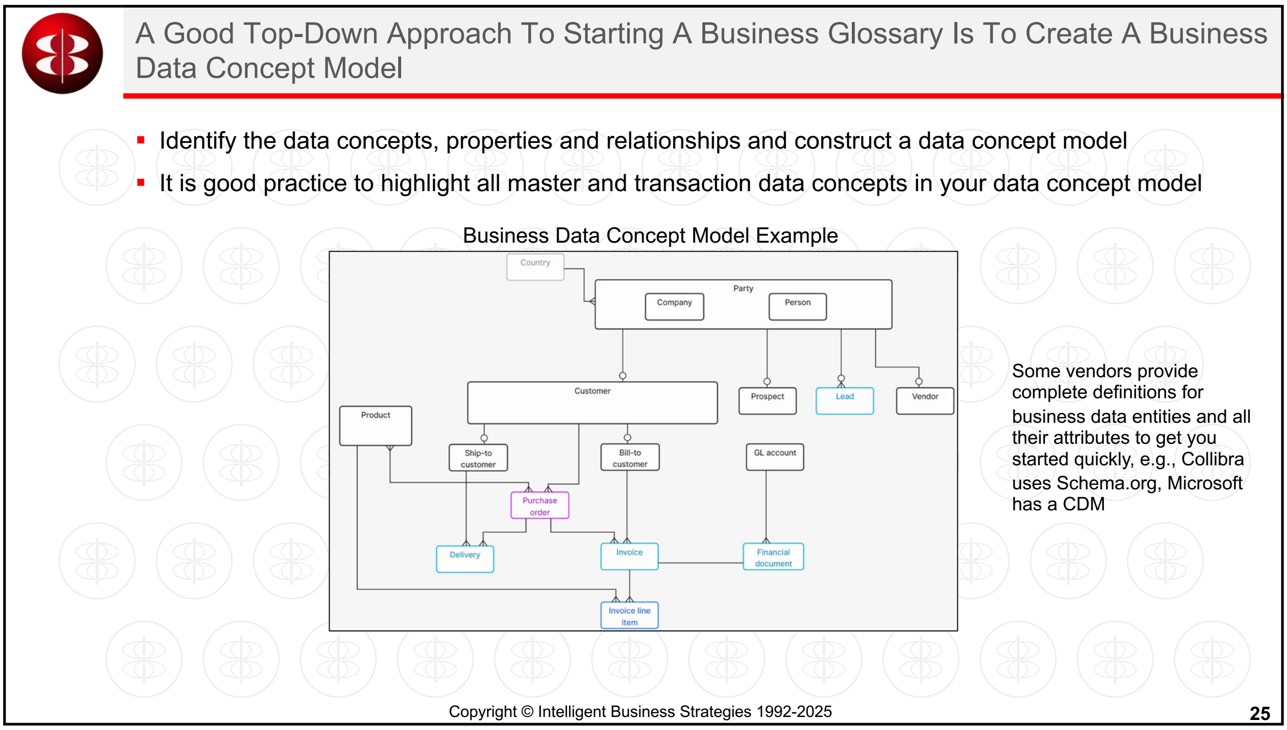
Working with subject matter & industry experts, Common Data Model is expanding to include additional business entities and concepts

Account:
Business that represents a customer or potential customer.

Attributes:

- accountNumber
- accountRatingCode
- createdOn
- creditLimit
- openDeals
- openRevenue
- territoryid
- hotelGroup
- ...

Source: https://learn.microsoft.com/en-us/common-data-model/

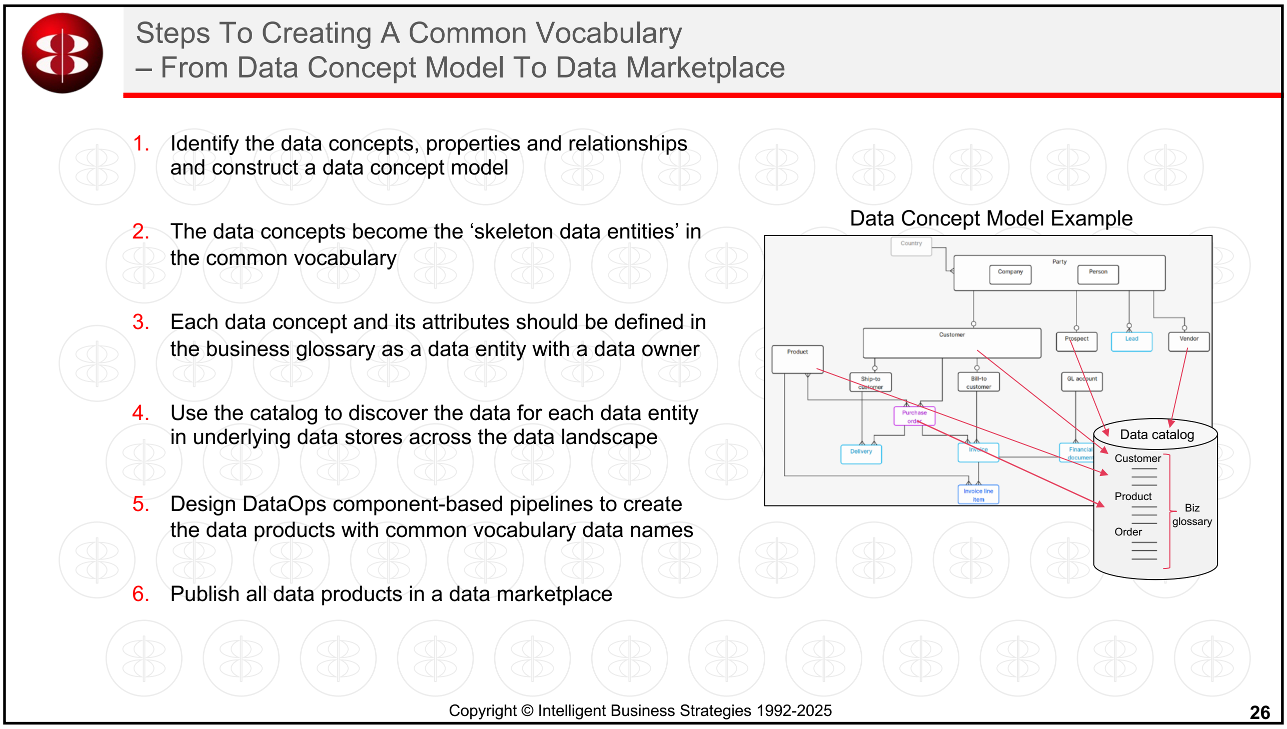


A Good Top-Down Approach To Starting A Business Glossary Is To Create A Business Data Concept Model

- 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



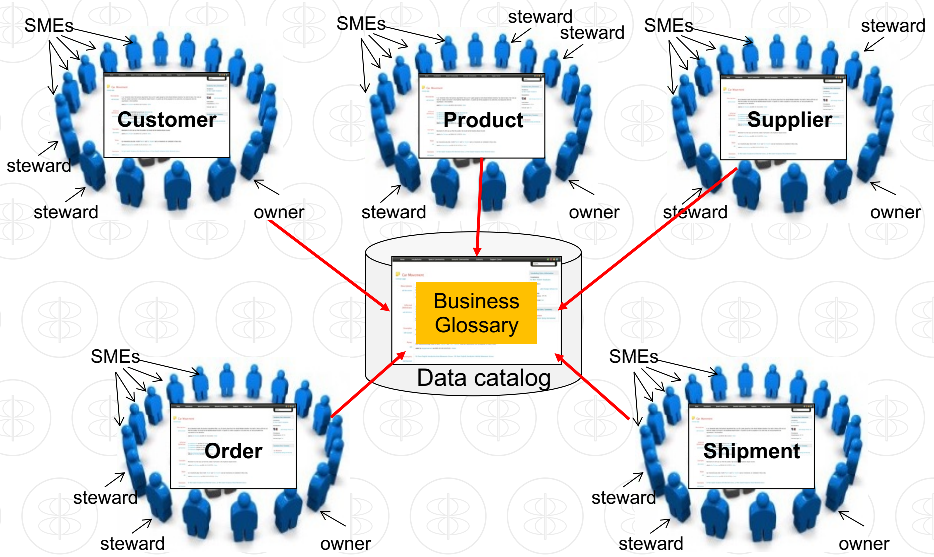
Steps To Creating A Common Vocabulary – From Data Concept Model To Data Marketplace

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
4. Use the catalog to discover the data for each data entity in underlying data stores across the data landscape
5. Design DataOps component-based pipelines to create the data products with common vocabulary data names
6. Publish all data products in a data marketplace

Data Concept Model Example

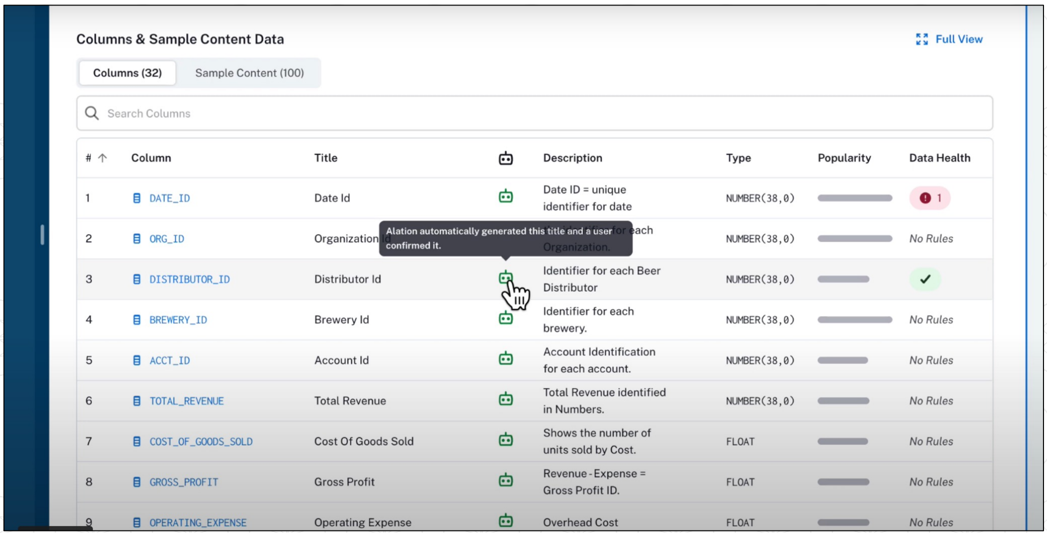
Several Communities Are Likely To Be Involved In Defining Common Data Names And Definitions For Data In A Business Glossary

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



Data Catalogs Can Also Use AI to Auto Generate Business Data Names for Your Business Glossary That Can Be Accepted or Rejected - E.g. User Confirmed Business Terms in Alation

This is a bottom-up approach to creating a business glossary



Columns & Sample Content Data

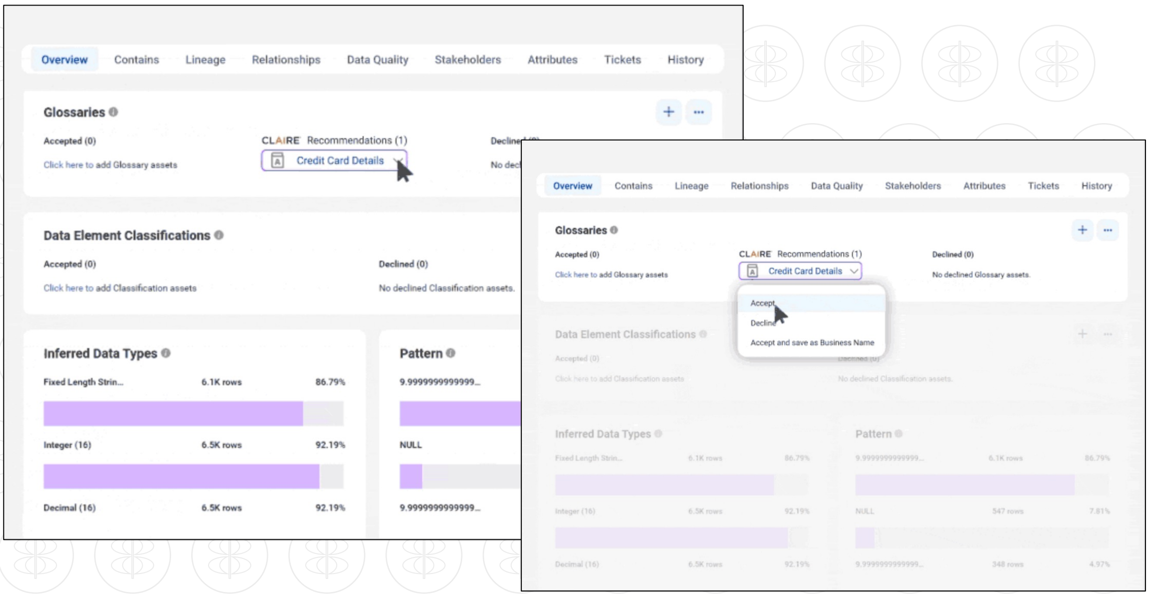
# ↑	Column	Title	🤖	Description	Type	Popularity	Data Health
1	DATE_ID	Date Id	🤖	Date ID = unique identifier for date	NUMBER(38,0)		! 1
2	ORG_ID	Organization Id		for each Organization.	NUMBER(38,0)		No Rules
3	DISTRIBUTOR_ID	Distributor Id	🤖	Identifier for each Beer Distributor	NUMBER(38,0)		✓
4	BREWERY_ID	Brewery Id	🤖	Identifier for each brewery.	NUMBER(38,0)		No Rules
5	ACCT_ID	Account Id	🤖	Account Identification for each account.	NUMBER(38,0)		No Rules
6	TOTAL_REVENUE	Total Revenue	🤖	Total Revenue identified in Numbers.	NUMBER(38,0)		No Rules
7	COST_OF_GOODS_SOLD	Cost Of Goods Sold	🤖	Shows the number of units sold by Cost.	FLOAT		No Rules
8	GROSS_PROFIT	Gross Profit	🤖	Revenue - Expense = Gross Profit ID.	FLOAT		No Rules
9	OPERATING_EXPENSE	Operating Expense	🤖	Overhead Cost	FLOAT		No Rules

Alation automatically generated this title and a user confirmed it.

Source: Alation

Green ‘robot heads’ indicate that the AI Generated business term has been confirmed by a business user

Data Catalogs Can Also Recommend Business Terms That You Can Accept or Decline - E.g., Informatica IDMC Data Governance and Catalog



Glossaries

Accepted (0)

Click here to add Glossary assets

CLAIRE Recommendations (1)

Credit Card Details

- Accept
- Decline
- Accept and save as Business Name

Declined (0)

No declined Glossary assets.

Data Element Classifications

Accepted (0)

Click here to add Classification assets

Declined (0)

No declined Classification assets.

Inferred Data Types

Fixed Length Strin...	6.1K rows	86.79%
Integer (16)	6.5K rows	92.19%
Decimal (16)	6.5K rows	92.19%

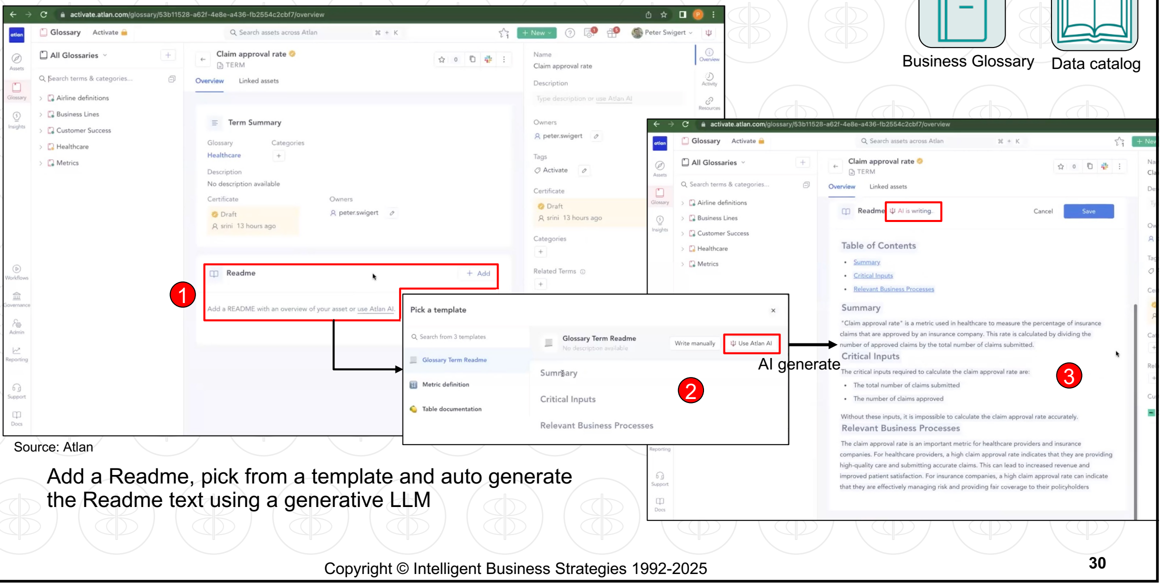
Pattern

9.9999999999999...	6.1K rows	86.79%
NULL	547 rows	7.81%
9.9999999999999...	348 rows	4.97%

Using Generative AI For Business Glossary Metadata Enrichment - Auto-Generation of Business Term Descriptions in Atlan Business Glossary

Business Glossary

Data catalog



Claim approval rate

TERM

Term Summary

Glossary: Healthcare

Categories

Description: No description available

Certificate: Draft, srini 13 hours ago

Owners: peter.swigert

Name: Claim approval rate

Description: Type description or use Atlan AI

Tags: Activate

Related Terms

Readme + Add

Add a README with an overview of your asset or use Atlan AI.

Pick a template

Search from 3 templates

- Glossary Term Readme
- Metric definition
- Table documentation

Glossary Term Readme — No description available — Write manually — Use Atlan AI

Summary

Critical Inputs

Relevant Business Processes

Readme — AI is writing.. — Cancel — Save

Table of Contents

- Summary
- Critical Inputs
- Relevant Business Processes

Summary

"Claim approval rate" is a metric used in healthcare to measure the percentage of insurance claims that are approved by an insurance company. This rate is calculated by dividing the number of approved claims by the total number of claims submitted.

Critical Inputs

The critical inputs required to calculate the claim approval rate are:

- The total number of claims submitted
- The number of claims approved

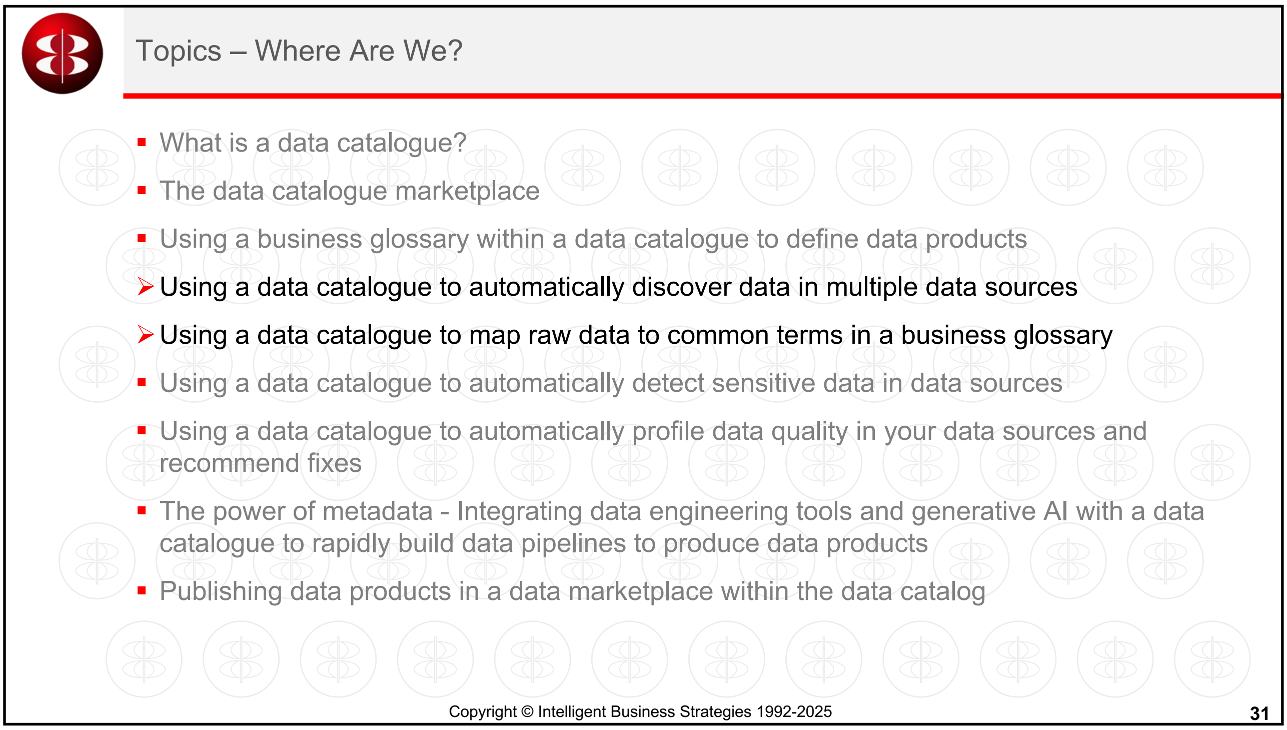
Without these inputs, it is impossible to calculate the claim approval rate accurately.

Relevant Business Processes

The claim approval rate is an important metric for healthcare providers and insurance companies. For healthcare providers, a high claim approval rate indicates that they are providing high-quality care and submitting accurate claims. This can lead to increased revenue and improved patient satisfaction. For insurance companies, a high claim approval rate can indicate that they are effectively managing risk and providing fair coverage to their policyholders

Source: Atlan

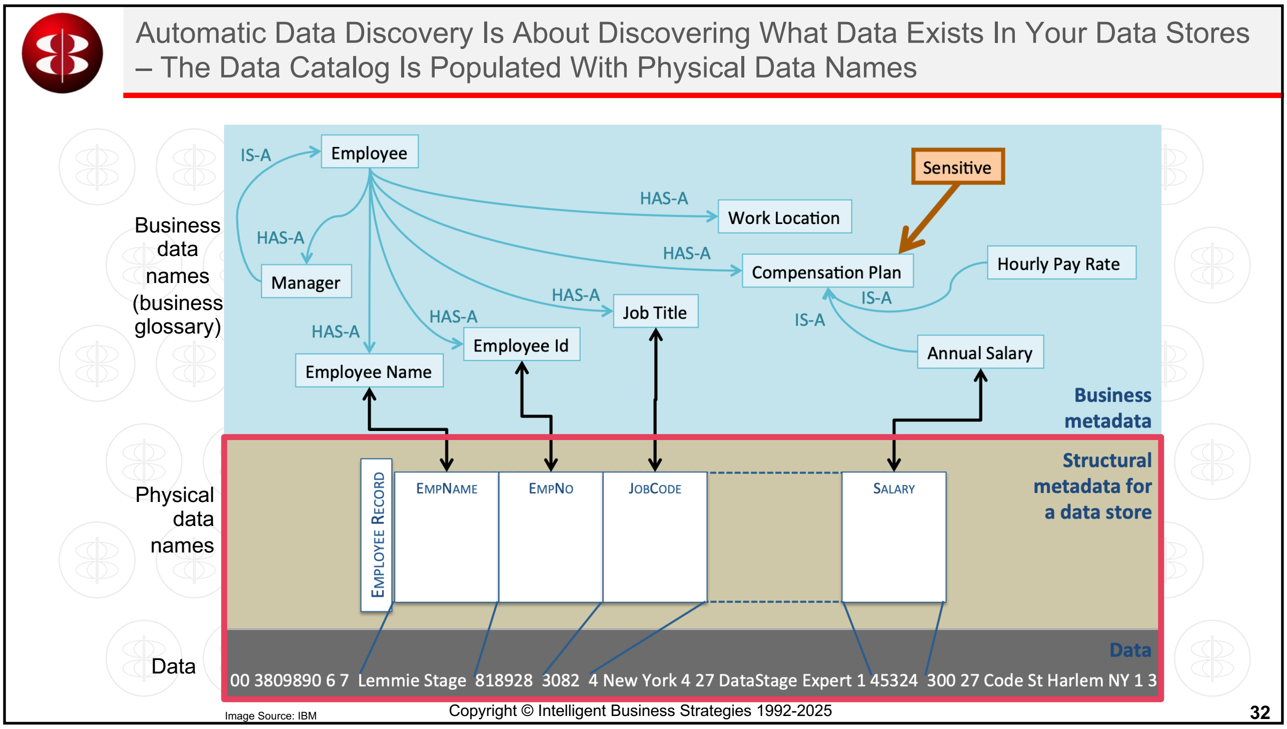
Add a Readme, pick from a template and auto generate the Readme text using a generative LLM



Topics – Where Are We?

- What is a data catalogue?
- The data catalogue marketplace
- Using a business glossary within a data catalogue to define data products
- ➢ Using a data catalogue to automatically discover data in multiple data sources
- ➢ Using a data catalogue to map raw data to common terms in a business glossary
- Using a data catalogue to automatically detect sensitive data in data sources
- Using a data catalogue to automatically profile data quality in your data sources and recommend fixes
- The power of metadata - Integrating data engineering tools and generative AI with a data catalogue to rapidly build data pipelines to produce data products
- Publishing data products in a data marketplace within the data catalog

31

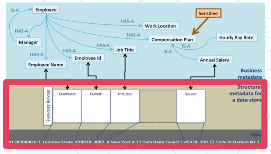


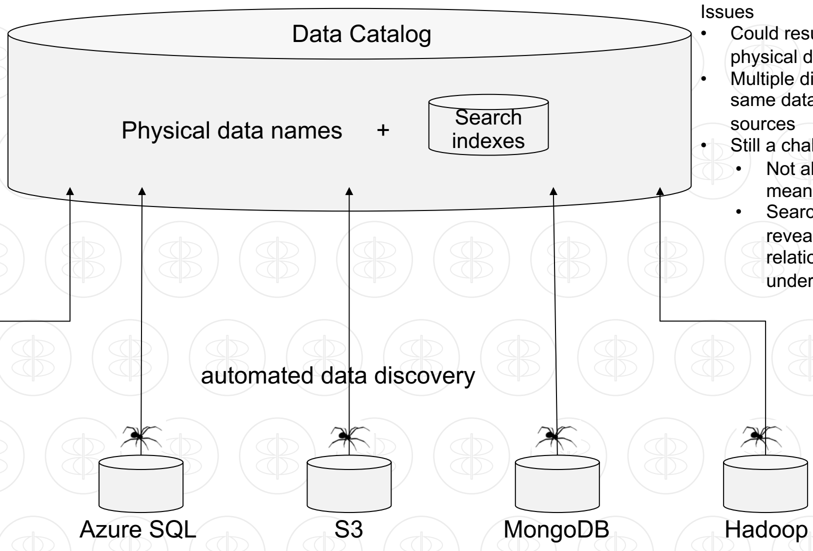
Automatic Data Discovery Is About Discovering What Data Exists In Your Data Stores – The Data Catalog Is Populated With Physical Data Names

Image Source: IBM

32

Automated Data Discovery Using Data Catalog Crawlers (Scanners) - Scan The Data Stores, Populate The Data Catalog And Build Search Indexes





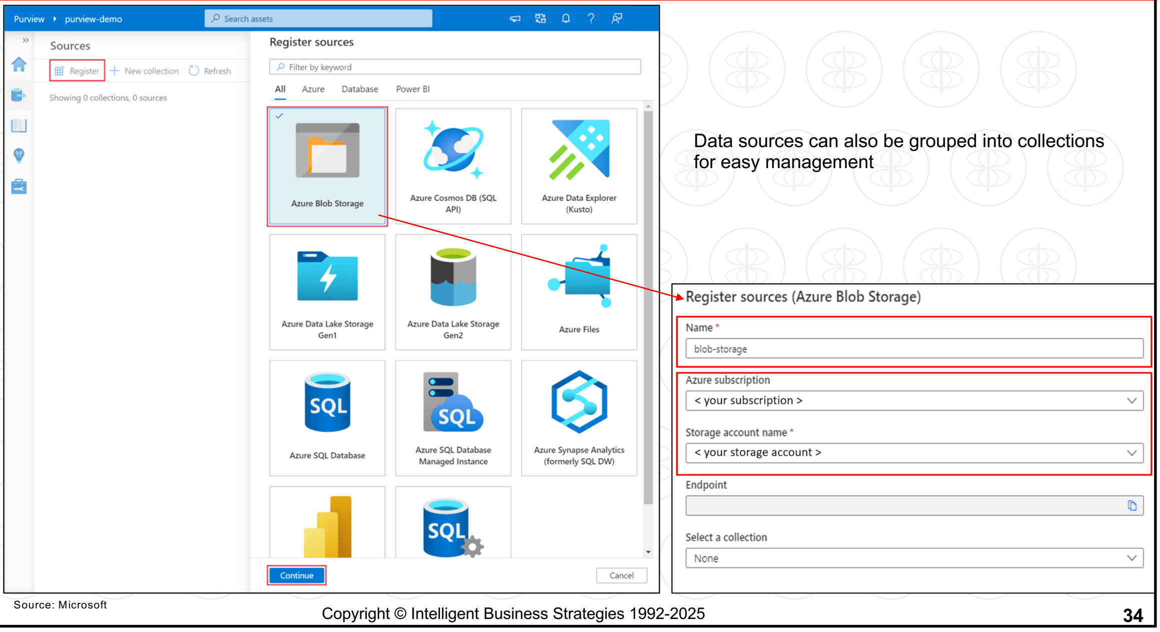
Scan time will vary depending on data volumes, scope of the scan, variety of the data and whether or not an incremental or full scan has been selected

Issues

- Could result in millions of physical data names
- Multiple different names for same data in different data sources
- Still a challenge to find data
  - Not all data has meaningful data names
  - Searching may not reveal all unless relationships are understood

Copyright © Intelligent Business Strategies 1992-2025 33

Registering Data Sources In A Data Catalog For Automated Discovery – Product Example: Microsoft Purview



Showing 0 collections, 0 sources

Register sources (Azure Blob Storage)

Name *: blob-storage

Azure subscription: < your subscription >

Storage account name *: < your storage account >

Endpoint:

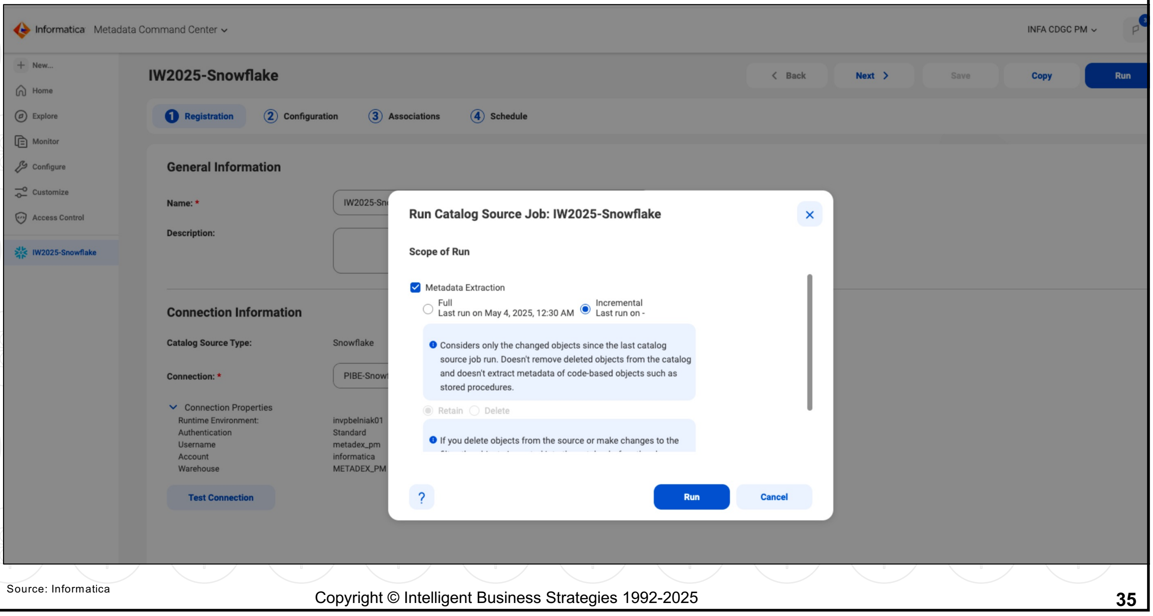
Select a collection: None

Data sources can also be grouped into collections for easy management

Source: Microsoft

Copyright © Intelligent Business Strategies 1992-2025 34

Some Data Catalog Products Offer Incremental Data Discover Scans - E.g., Informatica IDMC Data Governance and Catalog



Informatica Metadata Command Center

INFA CDGC PM

New... Home Explore Monitor Configure Customize Access Control IW2025-Snowflake

IW2025-Snowflake

Back | Next | Save | Copy | Run

1 Registration 2 Configuration 3 Associations 4 Schedule

General Information

Name: * IW2025-Sn

Description:

Connection Information

Catalog Source Type: Snowflake

Connection: * PIBE-Snow

Connection Properties

Runtime Environment:	invpbelniak01
Authentication	Standard
Username	metadex_pm
Account	informatica
Warehouse	METADEX_PM

Test Connection

Run Catalog Source Job: IW2025-Snowflake

Scope of Run

☑ Metadata Extraction

○ Full Last run on May 4, 2025, 12:30 AM ● Incremental Last run on -

Considers only the changed objects since the last catalog source job run. Doesn't remove deleted objects from the catalog and doesn't extract metadata of code-based objects such as stored procedures.

● Retain ○ Delete

If you delete objects from the source or make changes to the

? Run Cancel

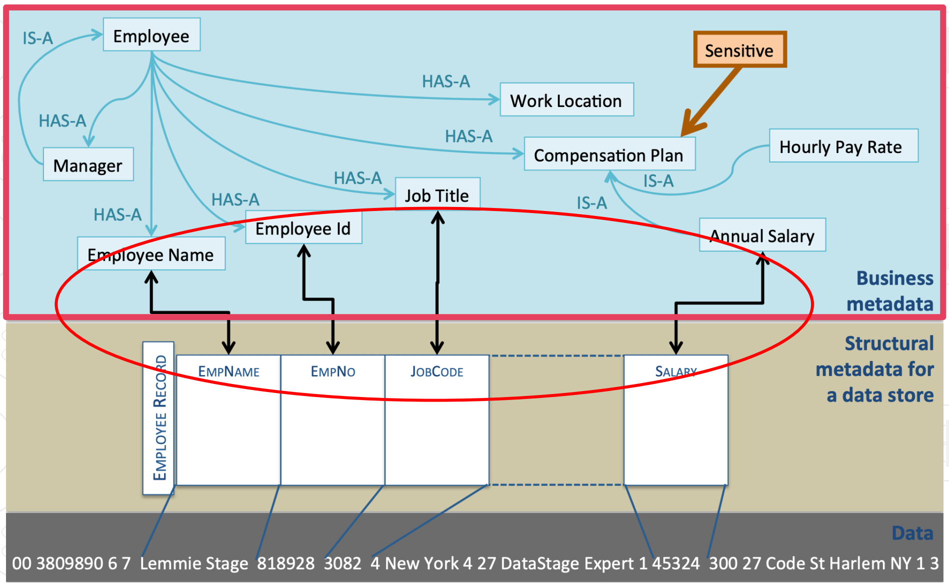
Source: Informatica

The Mapping Of Discovered Data In A Data Store To Business Metadata In A Business Glossary Enables Us To Understand The Meaning Of Data

Business data names (business glossary)

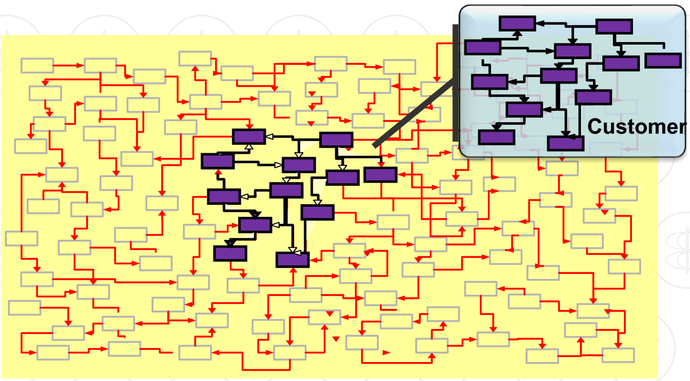
Physical data names

Data



Source: IBM

Automated Mapping of Raw Data to Business Terms Seeks to Determine the Meaning of Discovered Data and to Identify Complete Data Entities From Data Across Your Data Estate

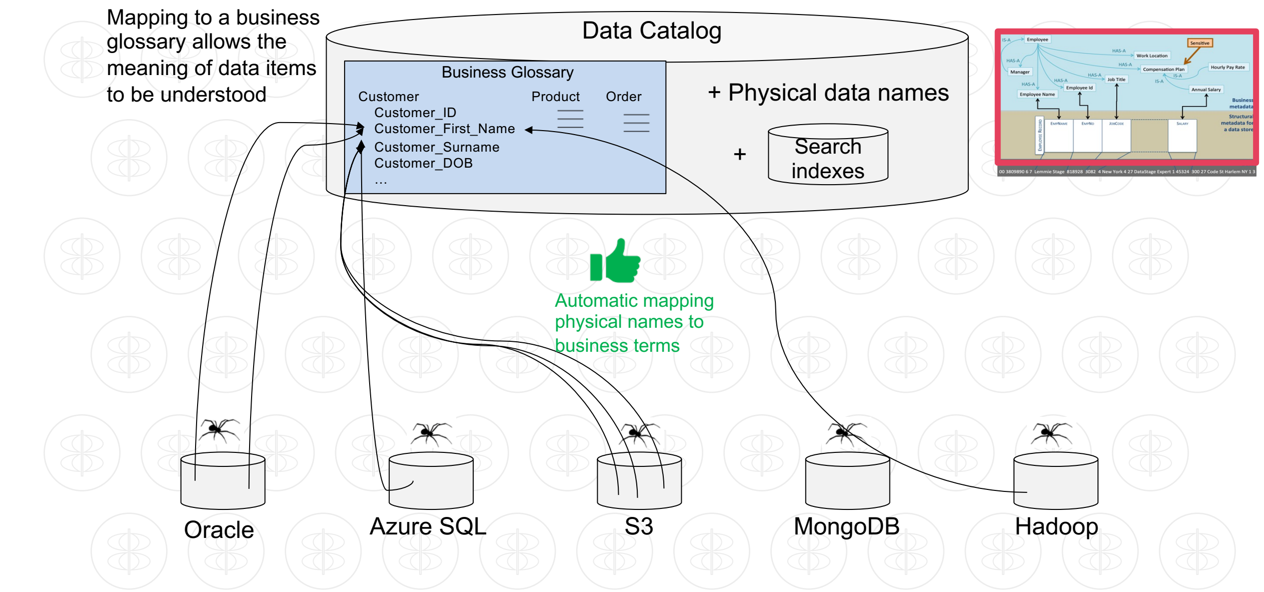


- Automated determination and grouping of data across multiple systems into business entities

How Do You Map Physical Data Names To Business Terms In A Business Glossary?

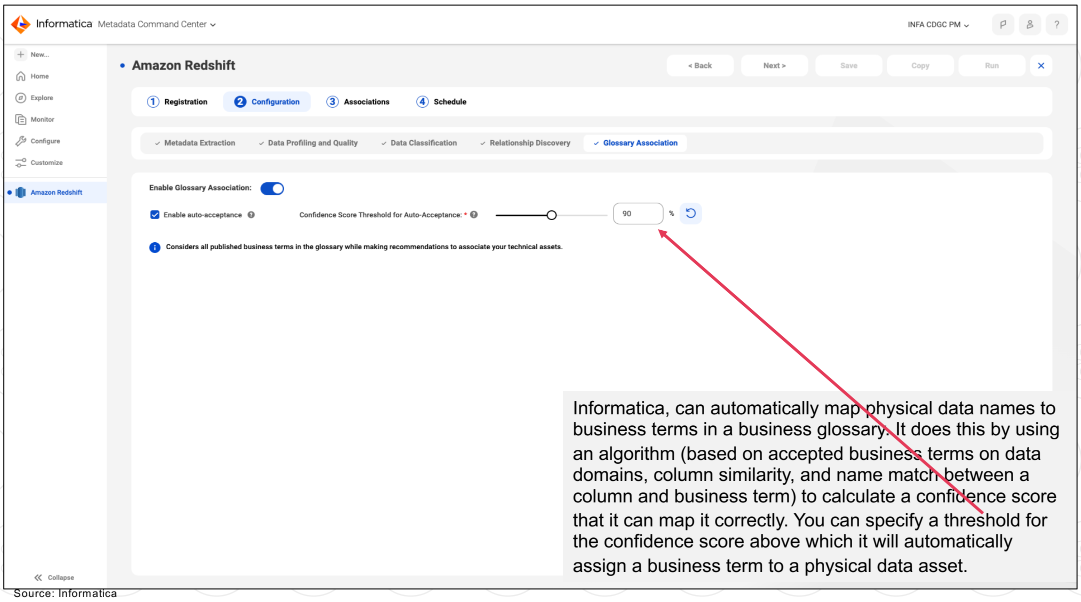
- Manual mapping done by business domain subject matter experts and data stewards
  - For occasional use and overriding incorrect automated mappings
  - Very time consuming and not practical for millions of files and thousands of tables and columns
  - Highly unlikely that any employee will understand data assets in SaaS applications
- **Automatically mapping** using several techniques
  - Pre-built machine learning models (AI-driven algorithmic scoring)
  - AI assisted mapping
    - Auto data discovery
    - Auto data clustering of similar data
    - Observability of manual mapping to glossary terms
    - Auto labelling of similar data based on observations
    - Training machine learning models to do it automatically
  - User defined rules, e.g., regular expressions
  - Reference data

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


Mapping to a business glossary allows the meaning of data items to be understood

Automatic mapping physical names to business terms

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



Amazon Redshift

1 Registration 2 Configuration 3 Associations 4 Schedule

✓ Metadata Extraction ✓ Data Profiling and Quality ✓ Data Classification ✓ Relationship Discovery ✓ Glossary Association

Enable Glossary Association: ●

☑ Enable auto-acceptance

Confidence Score Threshold for Auto-Acceptance: * 90 %

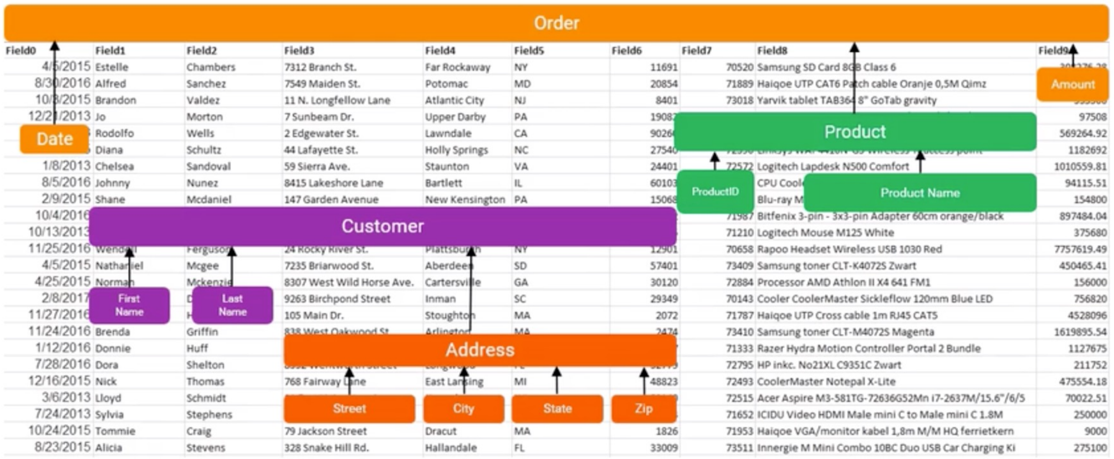
Considers all published business terms in the glossary while making recommendations to associate your technical assets.

Informatica, can automatically map physical data names to business terms in a business glossary. It does this by using an algorithm (based on accepted business terms on data domains, column similarity, and name match between a column and business term) to calculate a confidence score that it can map it correctly. You can specify a threshold for the confidence score above which it will automatically assign a business term to a physical data asset.

Source: Informatica

Auto Discovery of Complete Data Entities Such as Customer, Address, Product is Also Possible in Informatica Data Governance and Catalog Using Entity Classifications

Data entities are foundational data products



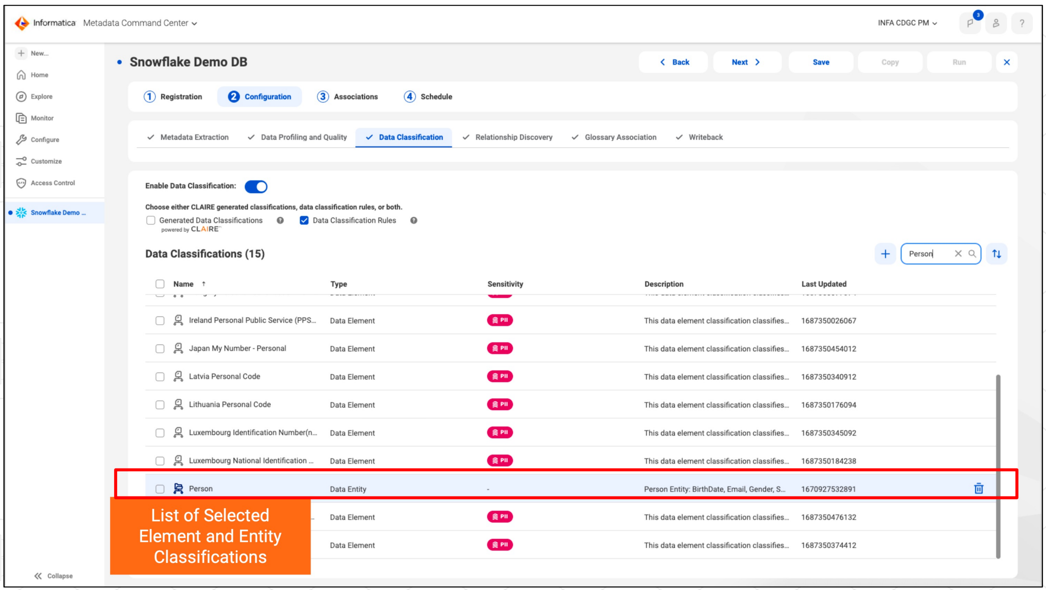
Field0	Field1	Field2	Field3	Field4	Field5	Field6	Field7	Field8	Field9
4/5/2015	Estelle	Chambers	7312 Branch St.	Far Rockaway	NY	11691	70520	Samsung SD Card 8GB Class 6	[illegible]
8/30/2016	Alfred	Sanchez	7549 Maiden St.	Potomac	MD	20854	71889	Haiqoe UTP CAT6 Patch cable Oranje 0,5M Qimz	[illegible]
10/8/2015	Brandon	Valdez	11 N. Longfellow Lane	Atlantic City	NJ	8401	73018	Yarvik tablet TAB364 8" GoTab gravity	[illegible]
12/21/2013	Jo	Morton	7 Sunbeam Dr.	Upper Darby	PA	1908[illegible]	[illegible]	[illegible]	97508
[illegible]	Rodolfo	Wells	2 Edgewater St.	Lawndale	CA	9026[illegible]	[illegible]	[illegible]	569264.92
[illegible]	Diana	Schultz	44 Lafayette St.	Holly Springs	NC	27540	[illegible]	[illegible]	1182692
1/8/2013	Chelsea	Sandoval	59 Sierra Ave.	Staunton	VA	24401	72572	Logitech Lapdesk N500 Comfort	1010559.81
8/5/2016	Johnny	Nunez	8415 Lakeshore Lane	Bartlett	IL	60103	[illegible]	CPU Cool[illegible]	94115.51
2/9/2015	Shane	Mcdaniel	147 Garden Avenue	New Kensington	PA	15068	[illegible]	Blu-ray M[illegible]	154800
10/4/2016	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	71987	Bitfenix 3-pin - 3x3-pin Adapter 60cm orange/black	897484.04
10/13/2013	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	71210	Logitech Mouse M125 White	375680
11/25/2016	Wendell	Ferguson	24 Rocky River St.	Plattsburgh	NY	12901	70658	Rapoo Headset Wireless USB 1030 Red	7757619.49
4/5/2015	Nathaniel	Mcgee	7235 Briarwood St.	Aberdeen	SD	57401	73409	Samsung toner CLT-K4072S Zwart	450465.41
4/25/2015	Norman	Mckenzie	8307 West Wild Horse Ave.	Cartersville	GA	30120	72884	Processor AMD Athlon II X4 641 FM1	156000
2/8/2017	[illegible]	D[illegible]	9263 Birchpond Street	Inman	SC	29349	70143	Cooler CoolerMaster Sickleflow 120mm Blue LED	756820
11/27/2016	[illegible]	H[illegible]	105 Main Dr.	Stoughton	MA	2072	71787	Haiqoe UTP Cross cable 1m RJ45 CAT5	4528096
11/24/2016	Brenda	Griffin	838 West Oakwood St.	Arlington	MA	2474	73410	Samsung toner CLT-M4072S Magenta	1619895.54
1/12/2016	Donnie	Huff	[illegible]	[illegible]	[illegible]	[illegible]	71333	Razer Hydra Motion Controller Portal 2 Bundle	1127675
7/28/2016	Dora	Shelton	[illegible]	[illegible]	[illegible]	[illegible]	72795	HP inkc. No21XL C9351C Zwart	211752
12/16/2015	Nick	Thomas	768 Fairway Lane	East Lansing	MI	48823	72493	CoolerMaster Notepal X-Lite	475554.18
3/6/2013	Lloyd	Schmidt	[illegible]	[illegible]	[illegible]	[illegible]	72515	Acer Aspire M3-581TG-72636G52Mn i7-2637M/15.6"/6/5	70022.51
7/24/2013	Sylvia	Stephens	[illegible]	[illegible]	[illegible]	[illegible]	71652	ICIDU Video HDMI Male mini C to Male mini C 1.8M	250000
10/24/2015	Tommie	Craig	79 Jackson Street	Dracut	MA	1826	71953	Haiqoe VGA/monitor kabel 1,8m M/M HQ ferrietkern	9000
8/23/2015	Alicia	Stevens	328 Snake Hill Rd.	Hallandale	FL	33009	73511	Innergie M Mini Combo 10BC Duo USB Car Charging Ki	275100

Source: Informatica

You define data entities that you want it to identify and then run the classifications

41

Identifying Complete Data Entities - Informatica IDMC Data Governance and Catalog Can Classify Data Elements and Complete Data Entities



Snowflake Demo DB

Enable Data Classification: ●

Choose either CLAIRE generated classifications, data classification rules, or both.

□ Generated Data Classifications ☑ Data Classification Rules

Data Classifications (15)

Search: Person

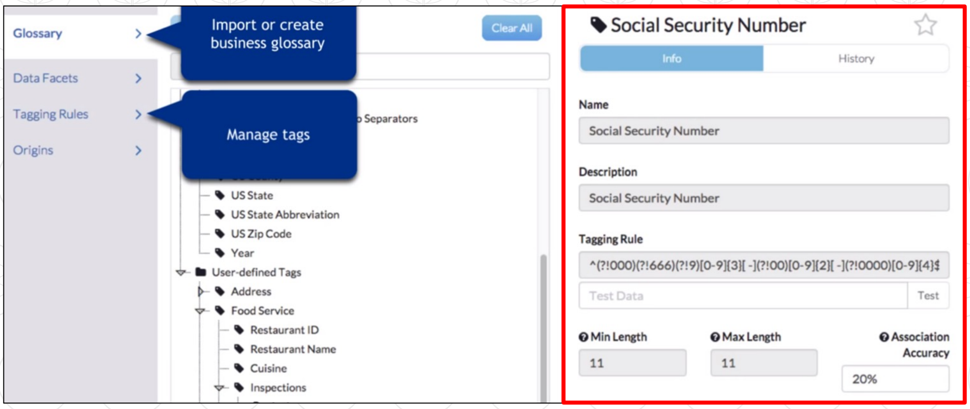
□	Name ↑	Type	Sensitivity	Description	Last Updated
□	Ireland Personal Public Service (PPS...	Data Element	PII	This data element classification classifies...	1687350026067
□	Japan My Number - Personal	Data Element	PII	This data element classification classifies...	1687350454012
□	Latvia Personal Code	Data Element	PII	This data element classification classifies...	1687350340912
□	Lithuania Personal Code	Data Element	PII	This data element classification classifies...	1687350176094
□	Luxembourg Identification Number(n...	Data Element	PII	This data element classification classifies...	1687350345092
□	Luxembourg National Identification ...	Data Element	PII	This data element classification classifies...	1687350184238
□	Person	Data Entity	-	Person Entity: BirthDate, Email, Gender, S...	1670927532891
[illegible]	[illegible]	Data Element	PII	This data element classification classifies...	1687350476132
[illegible]	[illegible]	Data Element	PII	This data element classification classifies...	1687350374412

List of Selected Element and Entity Classifications

Source: Informatica

42

You Can Also Add Your Own Rules to Extend Data Catalogs - Using Regex in Pentaho Data Catalog to Identify Social Security Number Data and Map It Your Business Glossary Terms



Social Security Number

Name: Social Security Number

Description: Social Security Number

Tagging Rule: ^(?!000)(?!666)(?!9)[0-9]{3}[-](?!00)[0-9]{2}[-](?!0000)[0-9]{4}$

Min Length: 11

Max Length: 11

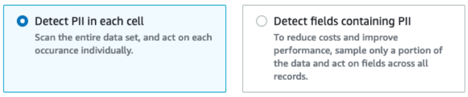
Association Accuracy: 20%

Source: Hitachi Vantara

Topics – Where Are We?

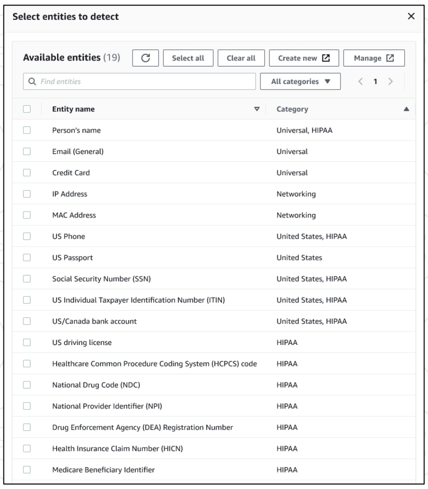
- What is a data catalogue?
- The data catalogue marketplace
- Using a business glossary within a data catalogue to define data products
- Using a data catalogue to automatically discover data in multiple data sources
- Using a data catalogue to map raw data to common terms in a business glossary
- ➢ Using a data catalogue to automatically detect sensitive data in data sources
- ➢ Using a data catalogue to automatically profile data quality in your data sources and recommend fixes
- The power of metadata - Integrating data engineering tools and generative AI with a data catalogue to rapidly build data pipelines to produce data products
- Publishing data products in a data marketplace within the data catalog

Several Vendors Support Predefined Classifiers To Classify Sensitive Data Types – E.g., AWS Glue Studio Can Detect And Process Sensitive Data



● Detect PII in each cell
Scan the entire data set, and act on each occurance individually.

○ Detect fields containing PII
To reduce costs and improve performance, sample only a portion of the data and act on fields across all records.



Select entities to detect

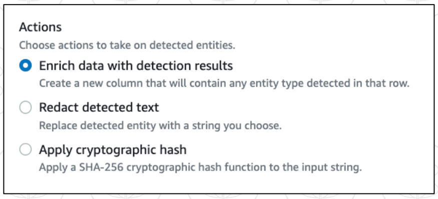
Available entities (19)

□	Entity name	Category
□	Person's name	Universal, HIPAA
□	Email (General)	Universal
□	Credit Card	Universal
□	IP Address	Networking
□	MAC Address	Networking
□	US Phone	United States, HIPAA
□	US Passport	United States
□	Social Security Number (SSN)	United States, HIPAA
□	US Individual Taxpayer Identification Number (ITIN)	United States, HIPAA
□	US/Canada bank account	United States, HIPAA
□	US driving license	HIPAA
□	Healthcare Common Procedure Coding System (HCPCS) code	HIPAA
□	National Drug Code (NDC)	HIPAA
□	National Provider Identifier (NPI)	HIPAA
□	Drug Enforcement Agency (DEA) Registration Number	HIPAA
□	Health Insurance Claim Number (HICN)	HIPAA
□	Medicare Beneficiary Identifier	HIPAA

Source: AWS

- Universal (e.g., Email, Credit Card)
- HIPAA (e.g., US Driving License, Healthcare Common Procedure Coding System (HCPCS) code)
- Networking (e.g., IP Address, MAC Address)
- United States (e.g., US Phone, US Passport)
- United Kingdom (e.g., UK Bank Account, UK VAT)
- Japan (e.g., Japan My Number, Japan Passport)

Choose what you want to do with identified PII data



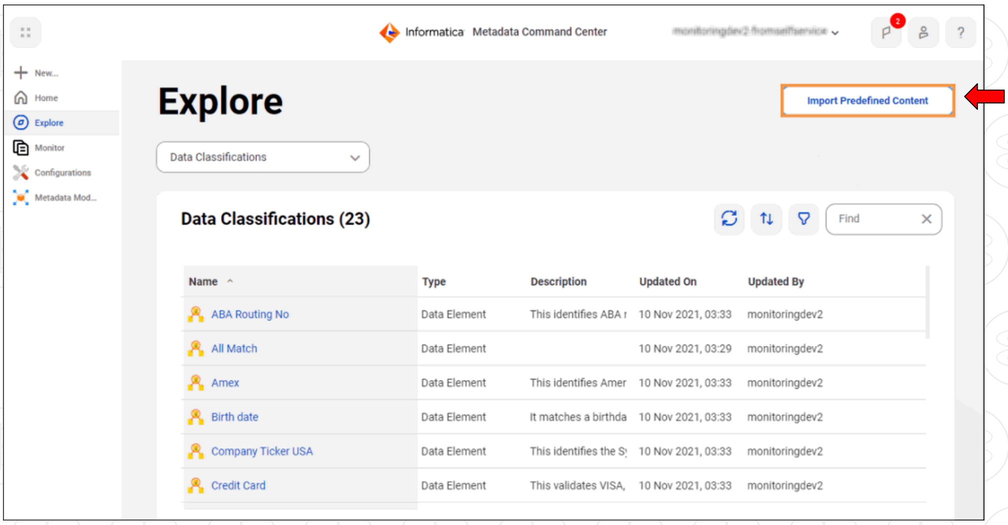
Actions
Choose actions to take on detected entities.

● Enrich data with detection results
Create a new column that will contain any entity type detected in that row.

○ Redact detected text
Replace detected entity with a string you choose.

○ Apply cryptographic hash
Apply a SHA-256 cryptographic hash function to the input string.

Informatica Metadata Command Center Let's You Import Over 200 Predefined Classifications That You Can Use To Identify Country Specific Personal Data In Sources



Explore

Data Classifications

Data Classifications (23)

Name	Type	Description	Updated On	Updated By
ABA Routing No	Data Element	This identifies ABA r	10 Nov 2021, 03:33	monitoringdev2
All Match	Data Element		10 Nov 2021, 03:29	monitoringdev2
Amex	Data Element	This identifies Amer	10 Nov 2021, 03:33	monitoringdev2
Birth date	Data Element	It matches a birthda	10 Nov 2021, 03:33	monitoringdev2
Company Ticker USA	Data Element	This identifies the S	10 Nov 2021, 03:33	monitoringdev2
Credit Card	Data Element	This validates VISA,	10 Nov 2021, 03:33	monitoringdev2

Source: Informatica

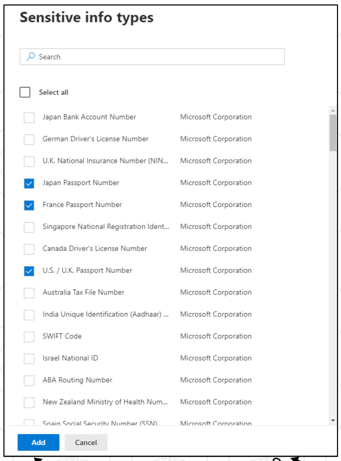
Note that Informatica predefined classifications cannot be modified

Several Vendors Support Predefined Classifiers To Classify Sensitive Data Types – E.g. 200+ Pre-Defined 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

….



Sensitive info types

Search

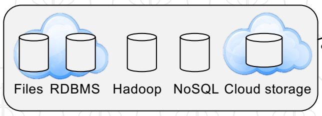
☐ Select all

☐	Japan Bank Account Number	Microsoft Corporation
☐	German Driver's License Number	Microsoft Corporation
☐	U.K. National Insurance Number (NIN...	Microsoft Corporation
☑	Japan Passport Number	Microsoft Corporation
☑	France Passport Number	Microsoft Corporation
☐	Singapore National Registration Ident...	Microsoft Corporation
☐	Canada Driver's License Number	Microsoft Corporation
☑	U.S. / U.K. Passport Number	Microsoft Corporation
☐	Australia Tax File Number	Microsoft Corporation
☐	India Unique Identification (Aadhaar) ...	Microsoft Corporation
☐	SWIFT Code	Microsoft Corporation
☐	Israel National ID	Microsoft Corporation
☐	ABA Routing Number	Microsoft Corporation
☐	New Zealand Ministry of Health Num...	Microsoft Corporation
☐	Spain Social Security Number (SSN)	Microsoft Corporation

These are examples of very specific sensitive data classifications

You can also add your own sensitive data types

Data

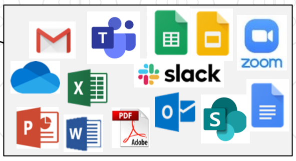


Scan & classify

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

Scan & classify

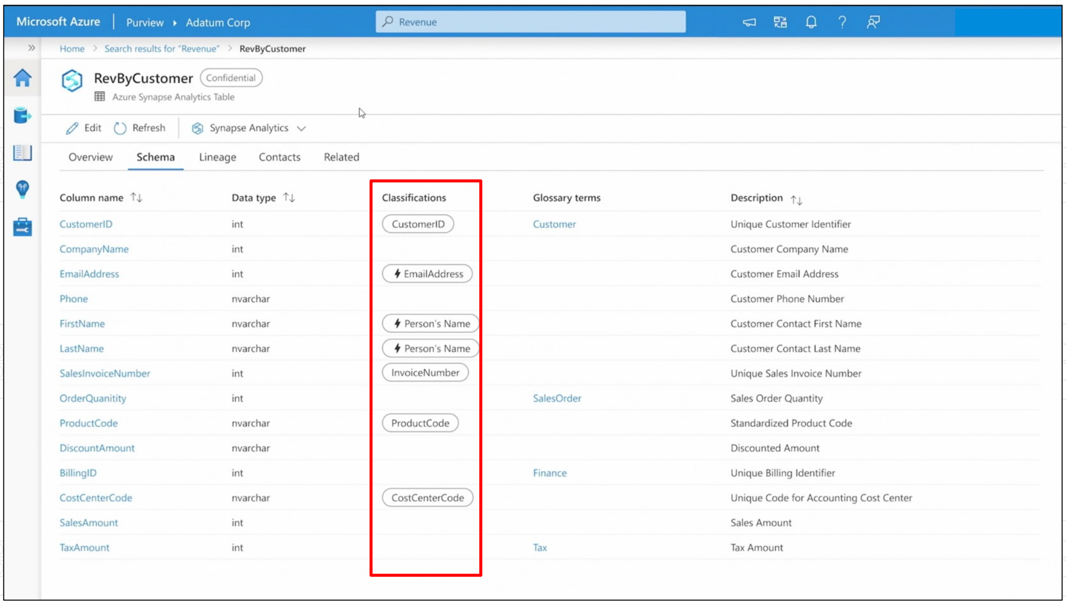
Content



Copyright © Intelligent Business Strategies 1992-2025

47

Microsoft Purview Data Catalog – Table Schema And Automated Classification Of A Discovered Data Asset



RevByCustomer (Confidential)

Azure Synapse Analytics Table

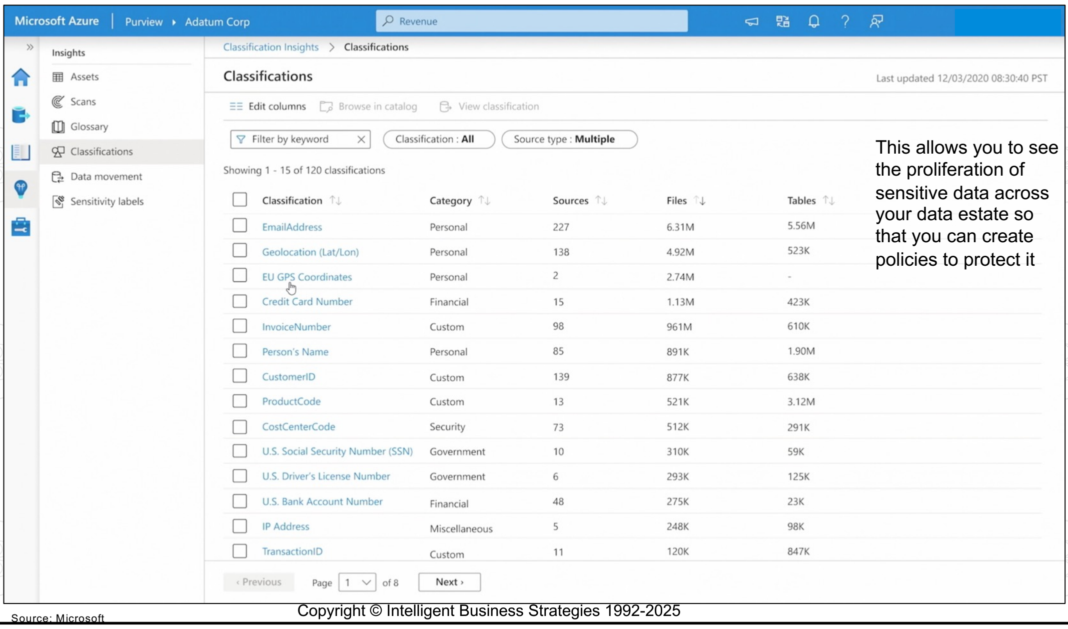
Column name ↑↓	Data type ↑↓	Classifications	Glossary terms	Description ↑↓
CustomerID	int	CustomerID	Customer	Unique Customer Identifier
CompanyName	int			Customer Company Name
EmailAddress	int	⚡ EmailAddress		Customer Email Address
Phone	nvarchar			Customer Phone Number
FirstName	nvarchar	⚡ Person's Name		Customer Contact First Name
LastName	nvarchar	⚡ Person's Name		Customer Contact Last Name
SalesInvoiceNumber	int	InvoiceNumber		Unique Sales Invoice Number
OrderQuantity	int		SalesOrder	Sales Order Quantity
ProductCode	nvarchar	ProductCode		Standardized Product Code
DiscountAmount	nvarchar			Discounted Amount
BillingID	int		Finance	Unique Billing Identifier
CostCenterCode	nvarchar	CostCenterCode		Unique Code for Accounting Cost Center
SalesAmount	int			Sales Amount
TaxAmount	int		Tax	Tax Amount

Source: Microsoft

Copyright © Intelligent Business Strategies 1992-2025

48

Microsoft Purview Data Catalog – Classification Counts Show How Many Data Sources, Files And Tables Have That Data Of That Classification Type


Microsoft Azure | Purview ▸ Adatum Corp
Revenue
Insights
Assets
Scans
Glossary
Classifications
Data movement
Sensitivity labels
Classification Insights > Classifications

Classifications

Last updated 12/03/2020 08:30:40 PST

Showing 1 - 15 of 120 classifications

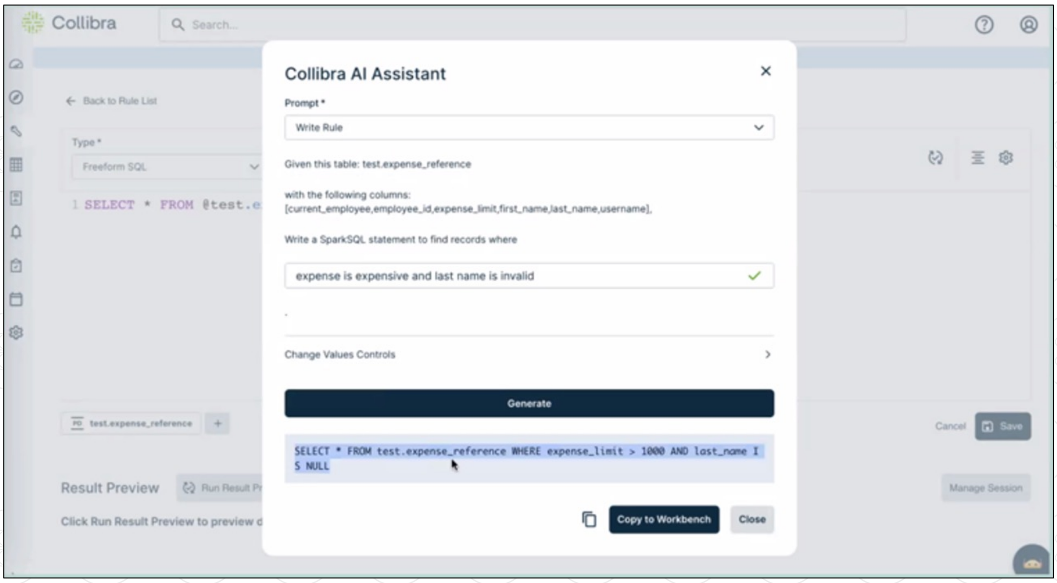
□	Classification	Category	Sources	Files	Tables
□	EmailAddress	Personal	227	6.31M	5.56M
□	Geolocation (Lat/Lon)	Personal	138	4.92M	523K
□	EU GPS Coordinates	Personal	2	2.74M	-
□	Credit Card Number	Financial	15	1.13M	423K
□	InvoiceNumber	Custom	98	961M	610K
□	Person's Name	Personal	85	891K	1.90M
□	CustomerID	Custom	139	877K	638K
□	ProductCode	Custom	13	521K	3.12M
□	CostCenterCode	Security	73	512K	291K
□	U.S. Social Security Number (SSN)	Government	10	310K	59K
□	U.S. Driver's License Number	Government	6	293K	125K
□	U.S. Bank Account Number	Financial	48	275K	23K
□	IP Address	Miscellaneous	5	248K	98K
□	TransactionID	Custom	11	120K	847K

This allows you to see the proliferation of sensitive data across your data estate so that you can create policies to protect it

Source: Microsoft

49

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



Collibra AI Assistant

Prompt *

Write Rule

Given this table: test.expense_reference

with the following columns:
[current_employee,employee_id,expense_limit,first_name,last_name,username],

Write a SparkSQL statement to find records where

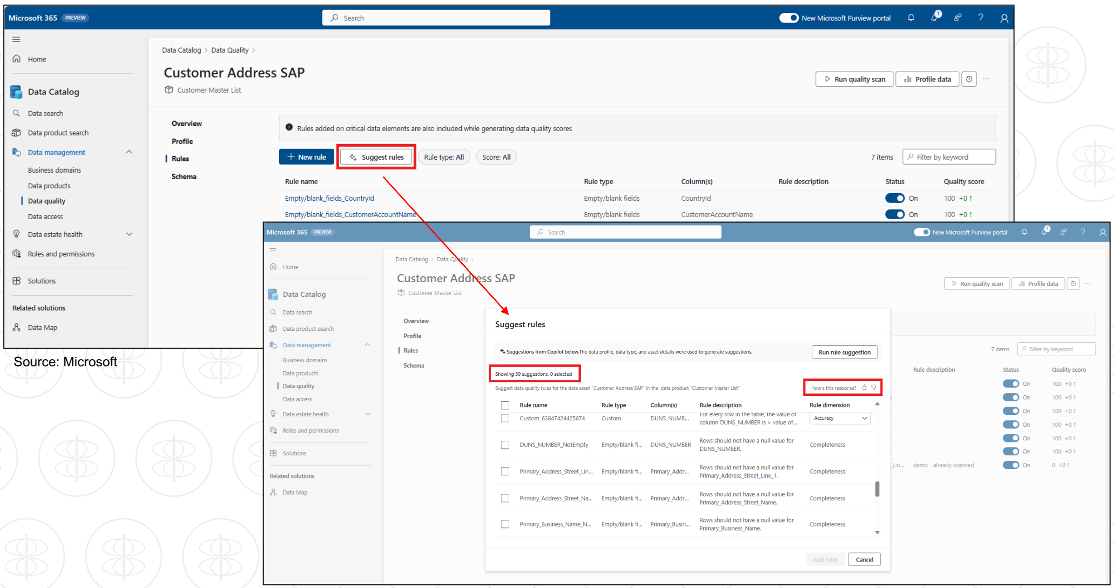
expense is expensive and last name is invalid ✓

```
SELECT * FROM test.expense_reference WHERE expense_limit > 1000 AND last_name I
S NULL
```

Source: Collibra

50

Some Data Catalogs Also Offer AI-Assisted Auto-generation of Data Quality Rules – e.g. Microsoft AI-Suggested Data Quality Rules



Customer Address SAP

Customer Master List

Rules added on critical data elements are also included while generating data quality scores

7 items

Rule name	Rule type	Column(s)	Rule description	Status	Quality score
Empty/blank_fields_CountryId	Empty/blank fields	CountryId		On	100 +0 ↑
Empty/blank_fields_CustomerAccountName	Empty/blank fields	CustomerAccountName		On	100 +0 ↑

Suggest rules

Suggestions from Copilot below.The data profile, data type, and asset details were used to generate suggestions.

Showing 39 suggestions, 0 selected

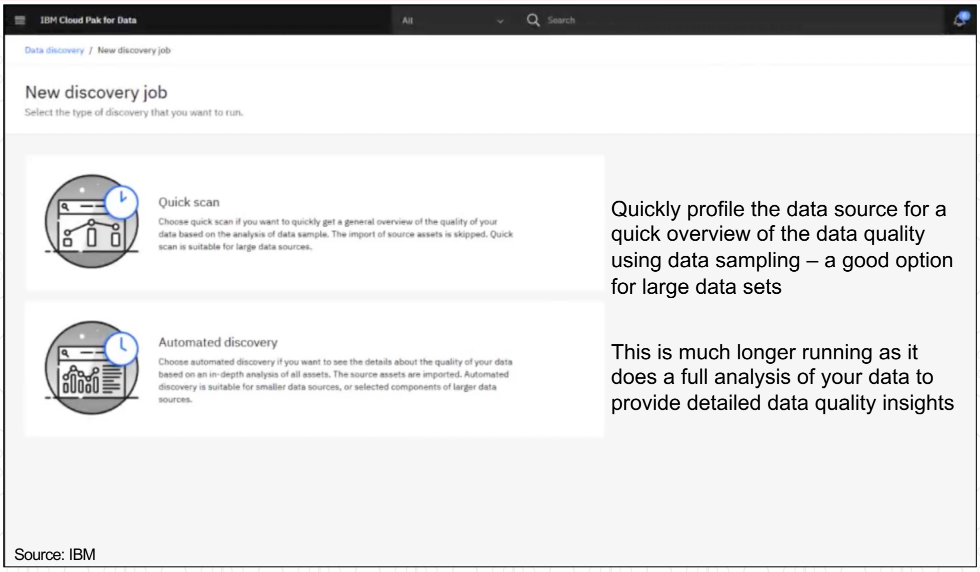
Suggest data quality rules for the data asset "Customer Address SAP" in the data product "Customer Master List"

□	Rule name	Rule type	Column(s)	Rule description	Rule dimension
□	Custom_63847424425674	Custom	DUNS_NUMB...	For every row in the table, the value of column DUNS_NUMBER is > value of...	Accuracy
□	DUNS_NUMBER_NotEmpty	Empty/blank fi...	DUNS_NUMBER	Rows should not have a null value for DUNS_NUMBER.	Completeness
□	Primary_Address_Street_Lin...	Empty/blank fi...	Primary_Addr...	Rows should not have a null value for Primary_Address_Street_Line_1.	Completeness
□	Primary_Address_Street_Na...	Empty/blank fi...	Primary_Addr...	Rows should not have a null value for Primary_Address_Street_Name.	Completeness
□	Primary_Business_Name_N...	Empty/blank fi...	Primary_Busin...	Rows should not have a null value for Primary_Business_Name.	Completeness

Source: Microsoft

Source: Microsoft

Data Catalogs Can Perform Automatic Data Quality Profiling and Generate Metadata In the Data Catalog – E.g. IBM Knowledge Catalog



New discovery job

Select the type of discovery that you want to run.

Quick scan

Choose quick scan if you want to quickly get a general overview of the quality of your data based on the analysis of data sample. The import of source assets is skipped. Quick scan is suitable for large data sources.

Automated discovery

Choose automated discovery if you want to see the details about the quality of your data based on an in-depth analysis of all assets. The source assets are imported. Automated discovery is suitable for smaller data sources, or selected components of larger data sources.

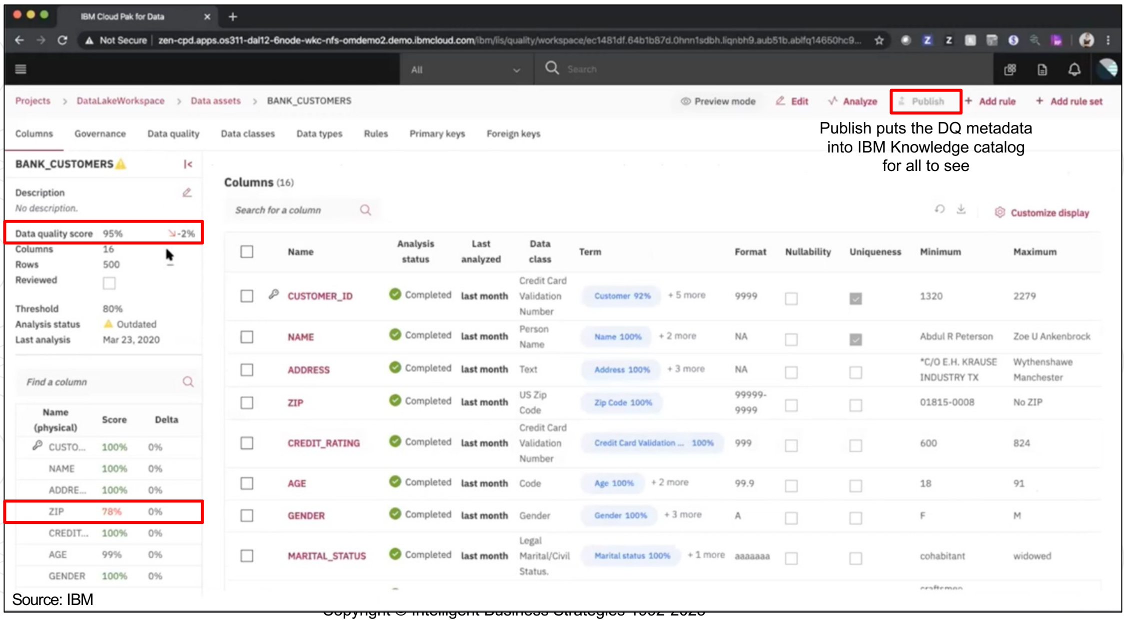
Quickly profile the data source for a quick overview of the data quality using data sampling – a good option for large data sets

This is much longer running as it does a full analysis of your data to provide detailed data quality insights

Source: IBM

Automated full and incremental data quality profiling is also supported by other vendors e.g. Informatica EDC

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



IBM Cloud Pak for Data
Not Secure | zen-cpd.apps.os311-dal12-6node-wkc-nfs-omdemo2.demo.ibmcloud.com/ibm/iis/quality/workspace/ec1481df.64b1b87d.0hnn1sdbh.liqnbh9.aub51b.ablfq14650hc9...
All
Search
Projects > DataLakeWorkspace > Data assets > BANK_CUSTOMERS
Preview mode
Edit
Analyze
Publish
Add rule
Add rule set
Columns
Governance
Data quality
Data classes
Data types
Rules
Primary keys
Foreign keys
Customize display

Publish puts the DQ metadata into IBM Knowledge catalog for all to see

BANK_CUSTOMERS

Description
No description.

Data quality score	95%	-2%
Columns	16	
Rows	500	
Reviewed	□	
Threshold	80%	
Analysis status	Outdated	
Last analysis	Mar 23, 2020	

Find a column

Name (physical)	Score	Delta
CUSTO...	100%	0%
NAME	100%	0%
ADDRE...	100%	0%
ZIP	78%	0%
CREDIT...	100%	0%
AGE	99%	0%
GENDER	100%	0%

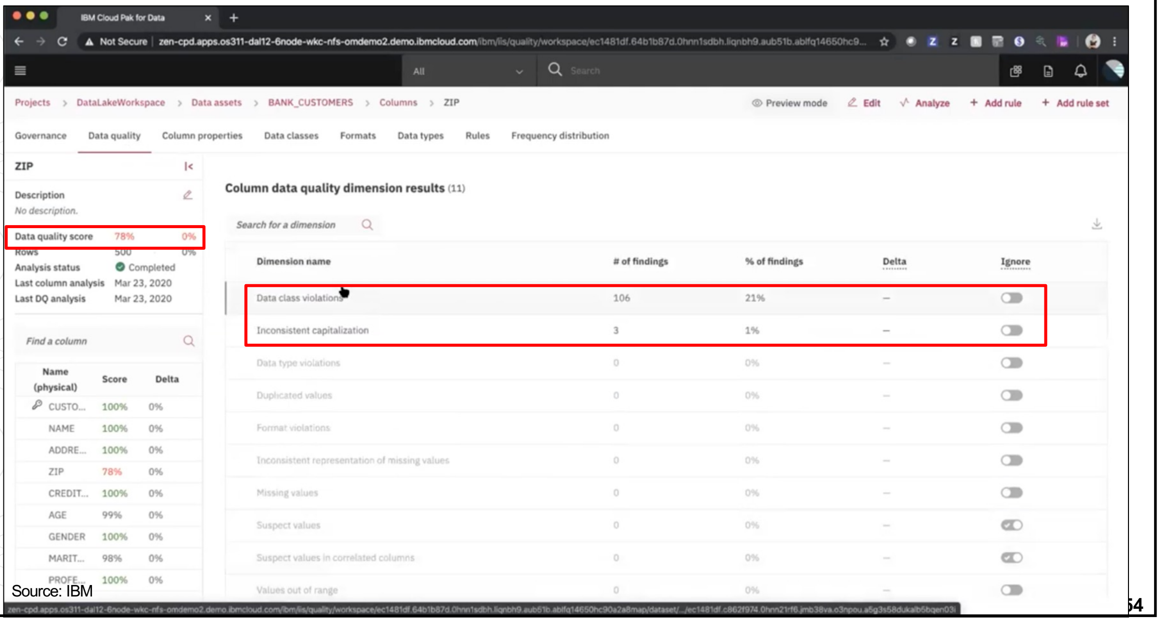
Columns (16)

Search for a column

□	Name	Analysis status	Last analyzed	Data class	Term	Format	Nullability	Uniqueness	Minimum	Maximum
□	CUSTOMER_ID	Completed	last month	Credit Card Validation Number	Customer 92% + 5 more	9999	□	☑	1320	2279
□	NAME	Completed	last month	Person Name	Name 100% + 2 more	NA	□	☑	Abdul R Peterson	Zoe U Ankenbrock
□	ADDRESS	Completed	last month	Text	Address 100% + 3 more	NA	□	□	*C/O E.H. KRAUSE INDUSTRY TX	Wythenshawe Manchester
□	ZIP	Completed	last month	US Zip Code	Zip Code 100%	99999-9999	□	□	01815-0008	No ZIP
□	CREDIT_RATING	Completed	last month	Credit Card Validation Number	Credit Card Validation ... 100%	999	□	□	600	824
□	AGE	Completed	last month	Code	Age 100% + 2 more	99.9	□	□	18	91
□	GENDER	Completed	last month	Gender	Gender 100% + 3 more	A	□	□	F	M
□	MARITAL_STATUS	Completed	last month	Legal Marital/Civil Status.	Marital status 100% + 1 more	aaaaaaa	□	□	cohabitant	widowed

Source: IBM

IBM Cloud Pak For Data Automated Data Quality Profiling Highlights Data Quality Issues, E.g., Data Violates The Matching Rules Set Up In A Data Class



IBM Cloud Pak for Data
Not Secure | zen-cpd.apps.os311-dal12-6node-wkc-nfs-omdemo2.demo.ibmcloud.com/ibm/iis/quality/workspace/ec1481df.64b1b87d.0hnn1sdbh.liqnbh9.aub51b.ablfq14650hc9...
All
Search
Projects > DataLakeWorkspace > Data assets > BANK_CUSTOMERS > Columns > ZIP
Preview mode
Edit
Analyze
Add rule
Add rule set
Governance
Data quality
Column properties
Data classes
Formats
Data types
Rules
Frequency distribution
zen-cpd.apps.os311-dal12-6node-wkc-nfs-omdemo2.demo.ibmcloud.com/ibm/iis/quality/workspace/ec1481df.64b1b87d.0hnn1sdbh.liqnbh9.aub61b.ablfq14650hc90a2a8map/dataset/.../ec1481df.c862f974.0hnn21rf6.jmb38va.o3npou.a5g3s58dukalb5bqen03i

ZIP

Description
No description.

Data quality score	78%	0%
Rows	500	0%
Analysis status	Completed	
Last column analysis	Mar 23, 2020	
Last DQ analysis	Mar 23, 2020	

Find a column

Name (physical)	Score	Delta
CUSTO...	100%	0%
NAME	100%	0%
ADDRE...	100%	0%
ZIP	78%	0%
CREDIT...	100%	0%
AGE	99%	0%
GENDER	100%	0%
MARIT...	98%	0%
PROFE...	100%	0%

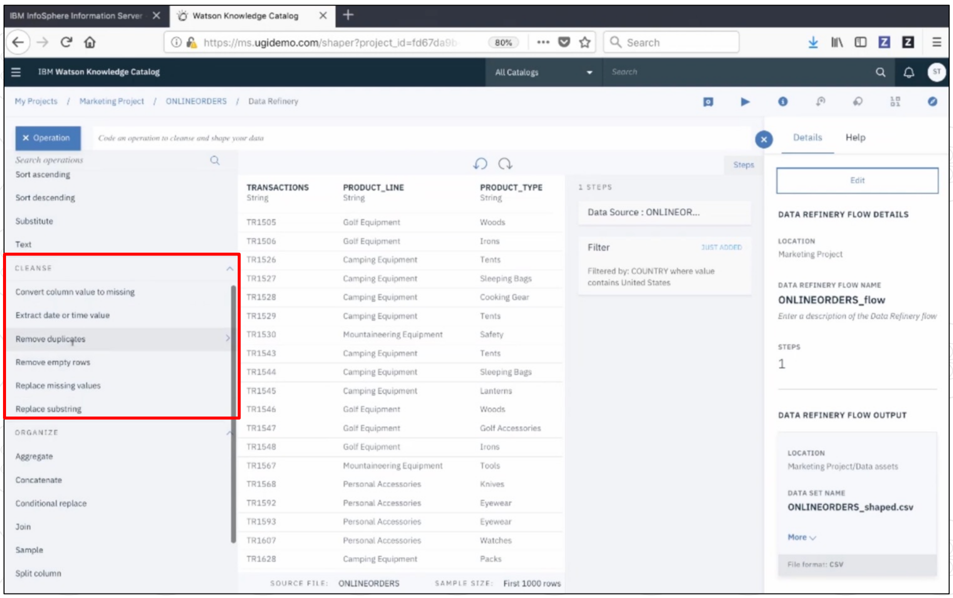
Column data quality dimension results (11)

Search for a dimension

Dimension name	# of findings	% of findings	Delta	Ignore
Data class violations	106	21%	–	○
Inconsistent capitalization	3	1%	–	○
Data type violations	0	0%	–	○
Duplicated values	0	0%	–	○
Format violations	0	0%	–	○
Inconsistent representation of missing values	0	0%	–	○
Missing values	0	0%	–	○
Suspect values	0	0%	–	✓
Suspect values in correlated columns	0	0%	–	✓
Values out of range	0	0%	–	○

Source: IBM

IBM Knowledge Catalog Allows You To Add Poor Quality Data To A Project And Immediately Start Cleaning It Using Self-Service Data Preparation


IBM InfoSphere Information Server
Watson Knowledge Catalog
https://ms.ugidemo.com/shaper?project_id=fd67da9b
80%
Search
IBM Watson Knowledge Catalog
All Catalogs
Search
ST

My Projects / Marketing Project / ONLINEORDERS / Data Refinery

✕ Operation

Code an operation to cleanse and shape your data

Search operations

- Sort ascending
- Sort descending
- Substitute
- Text

CLEANSE

- Convert column value to missing
- Extract date or time value
- Remove duplicates
- Remove empty rows
- Replace missing values
- Replace substring

ORGANIZE

- Aggregate
- Concatenate
- Conditional replace
- Join
- Sample
- Split column

TRANSACTIONS String	PRODUCT_LINE String	PRODUCT_TYPE String
TR1505	Golf Equipment	Woods
TR1506	Golf Equipment	Irons
TR1526	Camping Equipment	Tents
TR1527	Camping Equipment	Sleeping Bags
TR1528	Camping Equipment	Cooking Gear
TR1529	Camping Equipment	Tents
TR1530	Mountaineering Equipment	Safety
TR1543	Camping Equipment	Tents
TR1544	Camping Equipment	Sleeping Bags
TR1545	Camping Equipment	Lanterns
TR1546	Golf Equipment	Woods
TR1547	Golf Equipment	Golf Accessories
TR1548	Golf Equipment	Irons
TR1567	Mountaineering Equipment	Tools
TR1568	Personal Accessories	Knives
TR1592	Personal Accessories	Eyewear
TR1593	Personal Accessories	Eyewear
TR1607	Personal Accessories	Watches
TR1628	Camping Equipment	Packs

SOURCE FILE: ONLINEORDERS SAMPLE SIZE: First 1000 rows

Steps

1 STEPS

Data Source : ONLINEOR...

Filter JUST ADDED

Filtered by: COUNTRY where value contains United States

Details Help

Edit

DATA REFINERY FLOW DETAILS

LOCATION
Marketing Project

DATA REFINERY FLOW NAME
ONLINEORDERS_flow
Enter a description of the Data Refinery flow

STEPS
1

DATA REFINERY FLOW OUTPUT

LOCATION
Marketing Project/Data assets

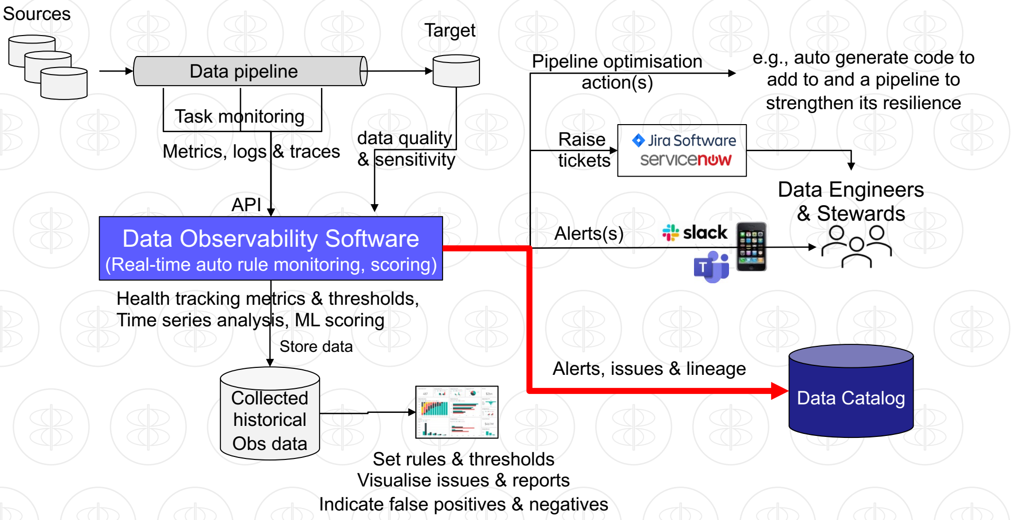
DATA SET NAME
ONLINEORDERS_shaped.csv

More

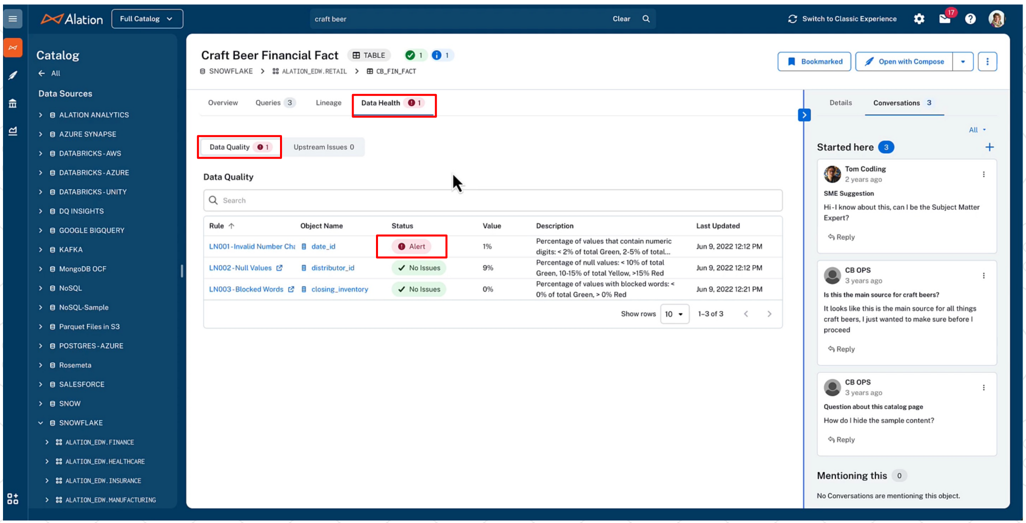
File format: CSV

Source: IBM

Data Observability Can Also Monitor Executing Data Pipelines, Cause Alerts and Pass Issues to Data Catalogs


Sources
Data pipeline
Target
Task monitoring
Metrics, logs & traces
data quality
& sensitivity
API
Data Observability Software
(Real-time auto rule monitoring, scoring)
Health tracking metrics & thresholds,
Time series analysis, ML scoring
Store data
Collected historical Obs data
Set rules & thresholds
Visualise issues & reports
Indicate false positives & negatives
Pipeline optimisation action(s)
e.g., auto generate code to add to and a pipeline to strengthen its resilience
Raise tickets
Jira Software
servicenow
Data Engineers & Stewards
Alerts(s)
slack
Alerts, issues & lineage
Data Catalog

Real-Time AI-Driven Active Data Governance In Data Quality Can Come From Integration Between Data Catalogs, Data Observability and DBMSs – E.g. Alation DQ Alerts



Craft Beer Financial Fact TABLE 1 1

SNOWFLAKE > ALATION_EDW.RETAIL > CB_FIN_FACT

Data Quality 1 Upstream Issues 0

Data Quality

Search

Rule	Object Name	Status	Value	Description	Last Updated
LN001 - Invalid Number Cha	date_id	Alert	1%	Percentage of values that contain numeric digits: < 2% of total Green, 2-5% of total...	Jun 9, 2022 12:12 PM
LN002 - Null Values	distributor_id	No Issues	9%	Percentage of null values: < 10% of total Green, 10-15% of total Yellow, >15% Red	Jun 9, 2022 12:12 PM
LN003 - Blocked Words	closing_inventory	No Issues	0%	Percentage of values with blocked words: < 0% of total Green, > 0% Red	Jun 9, 2022 12:21 PM

Show rows 10 1-3 of 3

Started here 3

Tom Codling
2 years ago
SME Suggestion
Hi - I know about this, can I be the Subject Matter Expert?
Reply

CB OPS
3 years ago
Is this the main source for craft beers?
It looks like this is the main source for all things craft beers, I just wanted to make sure before I proceed
Reply

CB OPS
3 years ago
Question about this catalog page
How do I hide the sample content?
Reply

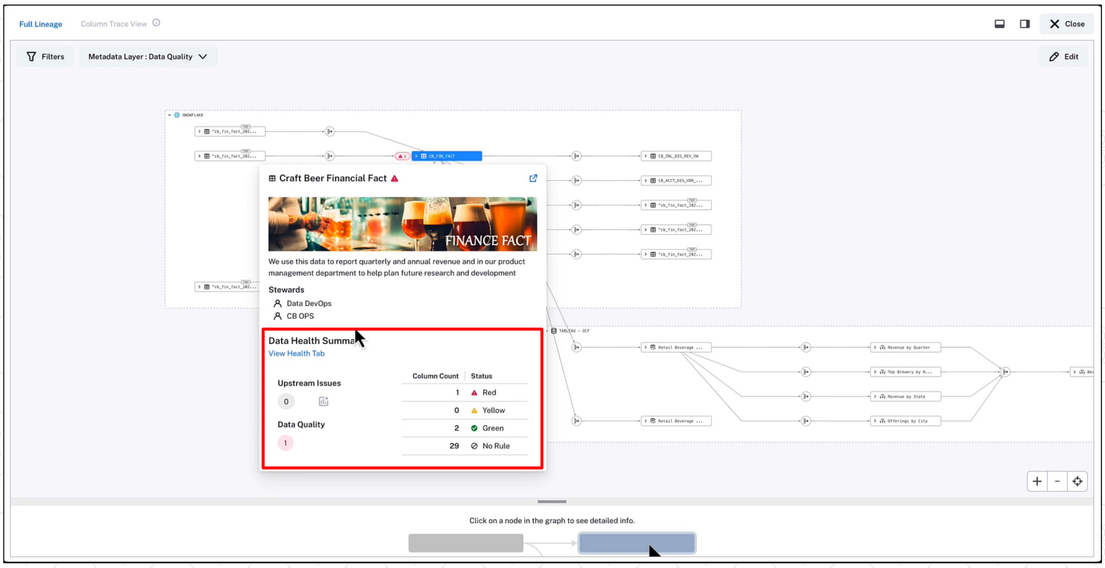
Mentioning this 0

No Conversations are mentioning this object.

Source: Alation

Integrates with SODA, Monte Carlo, Snowflake DMFs, Databricks etc.
Alation has DQ processes for Snowflake and Databricks to help govern DQ in these platforms

In Some Data Catalogs, Data Quality Can Also Be Overlayed on Top of Data Asset Lineage to See Data Quality Issues at a Glance – Alation Data Quality Overlay



Craft Beer Financial Fact

We use this data to report quarterly and annual revenue and in our product management department to help plan future research and development

Stewards
Data DevOps
CB OPS

Data Health Summa
View Health Tab

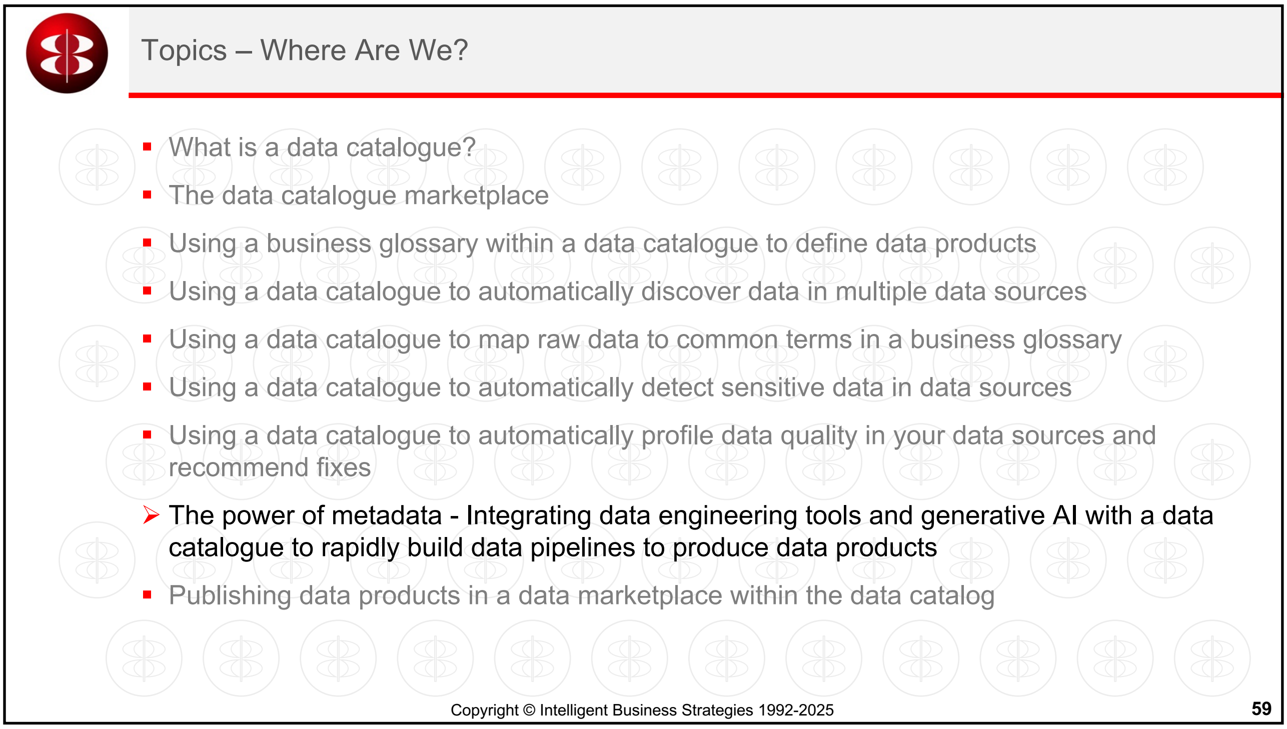
Upstream Issues
0

Data Quality
1

Column Count	Status
1	Red
0	Yellow
2	Green
29	No Rule

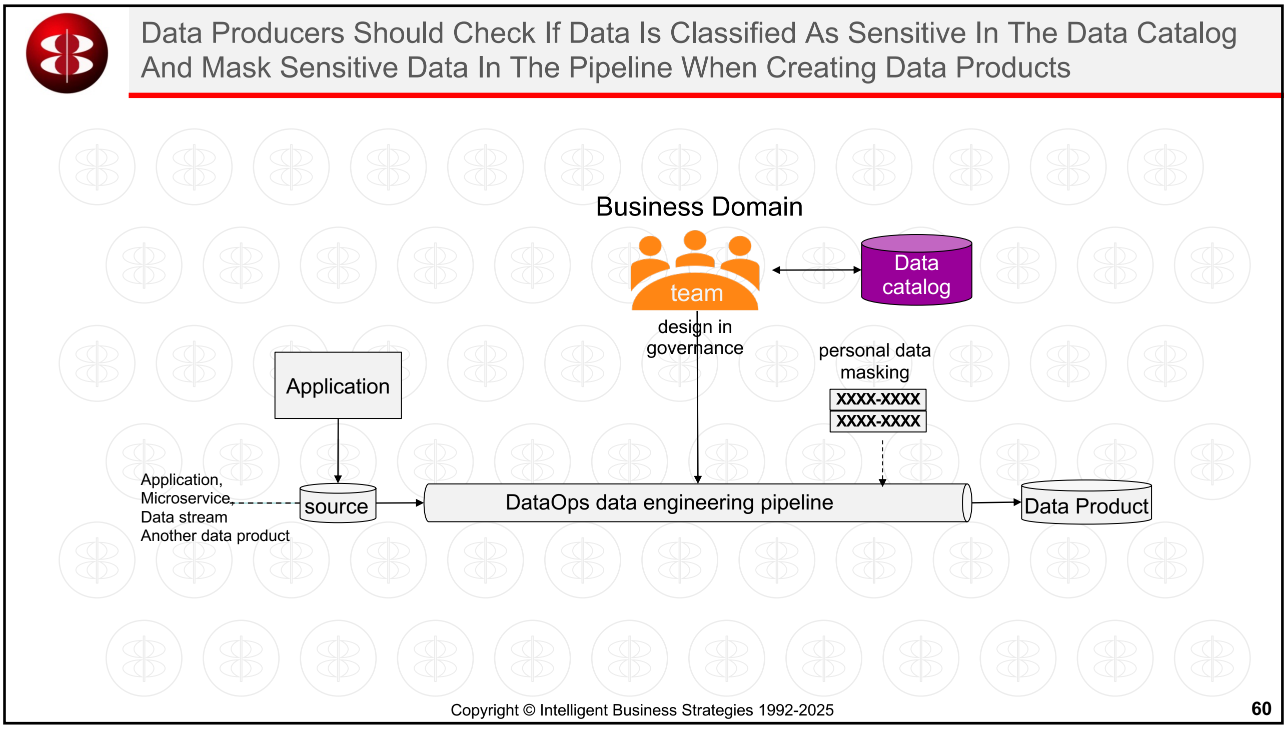
Click on a node in the graph to see detailed info.

Source: Alation



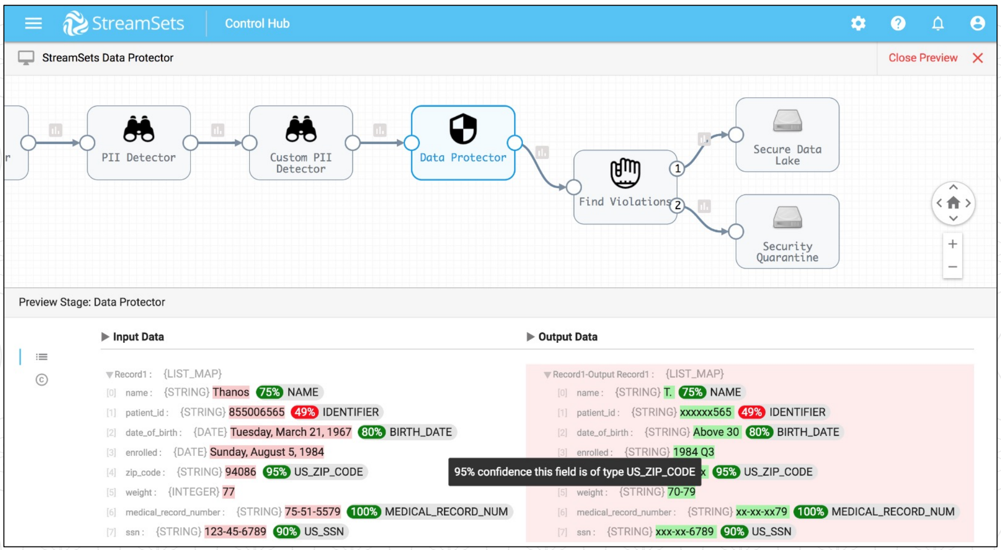
Topics – Where Are We?

- What is a data catalogue?
- The data catalogue marketplace
- Using a business glossary within a data catalogue to define data products
- Using a data catalogue to automatically discover data in multiple data sources
- Using a data catalogue to map raw data to common terms in a business glossary
- Using a data catalogue to automatically detect sensitive data in data sources
- Using a data catalogue to automatically profile data quality in your data sources and recommend fixes
- ➢ The power of metadata - Integrating data engineering tools and generative AI with a data catalogue to rapidly build data pipelines to produce data products
- Publishing data products in a data marketplace within the data catalog



Data Producers Should Check If Data Is Classified As Sensitive In The Data Catalog And Mask Sensitive Data In The Pipeline When Creating Data Products

Data Masking In A Pipeline That Produces a Data Product – E.g. IBM StreamSets



Preview Stage: Data Protector

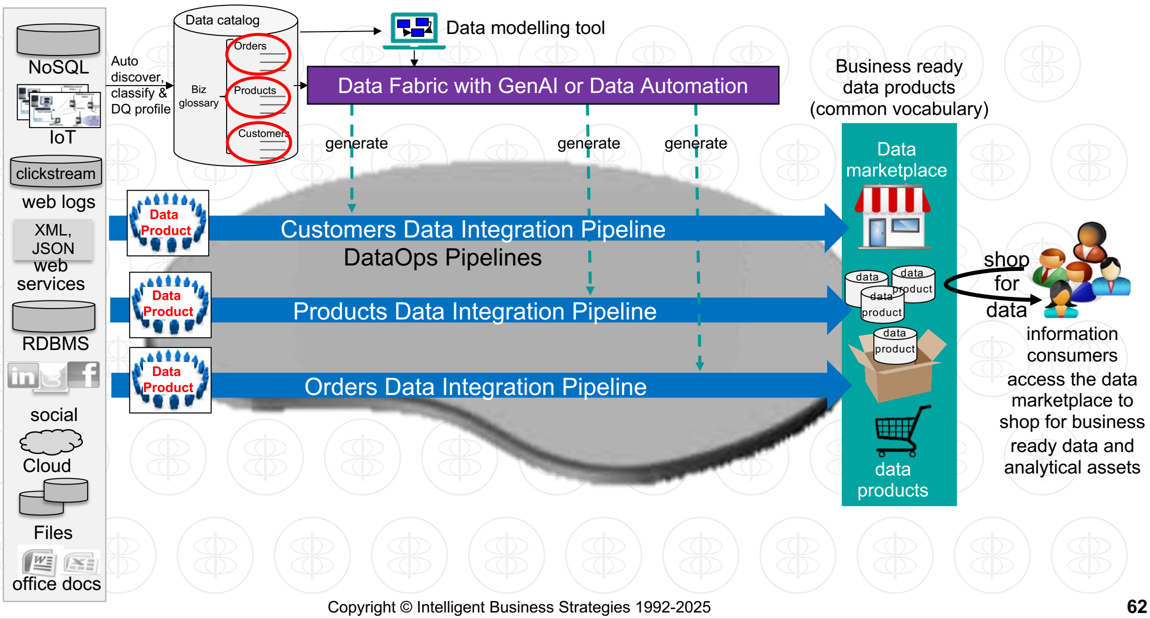
```
▶ Input Data
▼Record1 :  {LIST_MAP}
[0]  name :  {STRING} Thanos  75% NAME
[1]  patient_id :  {STRING} 855006565  49% IDENTIFIER
[2]  date_of_birth :  {DATE} Tuesday, March 21, 1967  80% BIRTH_DATE
[3]  enrolled :  {DATE} Sunday, August 5, 1984
[4]  zip_code :  {STRING} 94086  95% US_ZIP_CODE
[5]  weight :  {INTEGER} 77
[6]  medical_record_number :  {STRING} 75-51-5579  100% MEDICAL_RECORD_NUM
[7]  ssn :  {STRING} 123-45-6789  90% US_SSN

▶ Output Data
▼Record1-Output Record1 :  {LIST_MAP}
[0]  name :  {STRING} T.  75% NAME
[1]  patient_id :  {STRING} xxxxxx565  49% IDENTIFIER
[2]  date_of_birth :  {STRING} Above 30  80% BIRTH_DATE
[3]  enrolled :  {STRING} 1984 Q3
95% confidence this field is of type US_ZIP_CODE x  95% US_ZIP_CODE
[5]  weight :  {STRING} 70-79
[6]  medical_record_number :  {STRING} xx-xx-xx79  100% MEDICAL_RECORD_NUM
[7]  ssn :  {STRING} xxx-xx-6789  90% US_SSN
```

Source: StreamSets

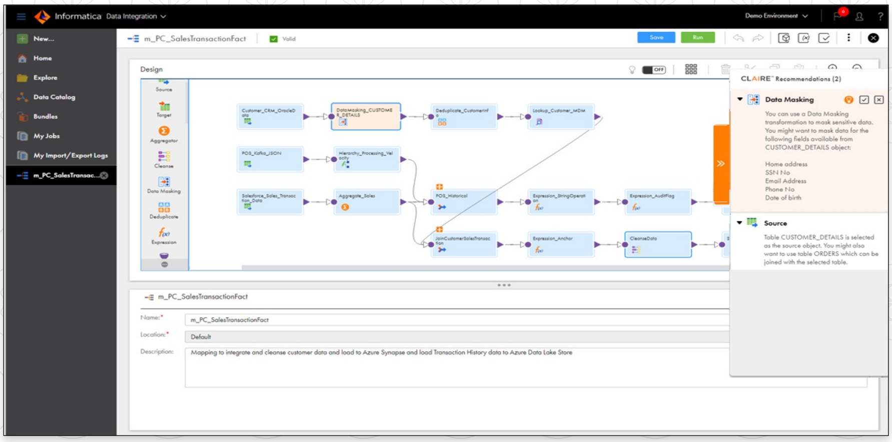
Masking capability is available in several data fabric software offerings

Use Metadata, Gen AI Data Automation To Generate Pipelines To Produce A Data Product For Each Entity Defined In The Business Glossary Within A Data Catalog



information consumers access the data marketplace to shop for business ready data and analytical assets

AI Assisted Data Fabric Software Can Now Automatically Generate Pipelines – Informatica IDMC CLAIRE Co-Pilot AI-Driven Automated Pipeline Generation



CLAIRE™ Recommendations (2)

Data Masking

You can use a Data Masking transformation to mask sensitive data. You might want to mask data for the following fields available from CUSTOMER_DETAILS object:

Home address
SSN No
Email Address
Phone No
Date of birth

Source

Table CUSTOMER_DETAILS is selected as the source object. You might also want to use table ORDERS which can be joined with the selected table.

m_PC_SalesTransactionFact

Name:* m_PC_SalesTransactionFact

Location:* Default

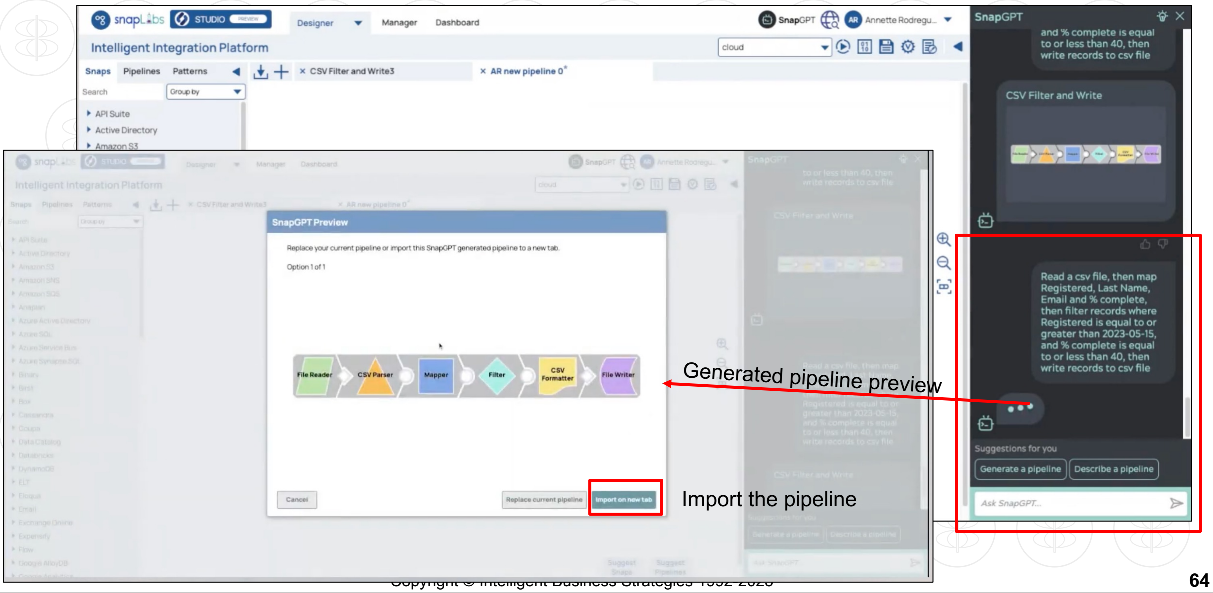
Description: Mapping to integrate and cleanse customer data and load to Azure Synapse and load Transaction History data to Azure Data Lake Store

Source: Informatica

- Fully automated data pipeline generation
- Support for unstructured, semi-structured and hierarchical data sources
- Automated creation data masking, cleansing and standardization rules
- Support for join, union, normalization and denormalization of data sources

- Simplifies and accelerates development
- Improves productivity

Generative AI In Data Engineering Opens Up Data Product Development To Citizen Data Engineers – SnapLogic SnapGPT Prompt Based Data Engineering



Replace your current pipeline or import this SnapGPT generated pipeline to a new tab.

Option 1 of 1

Generated pipeline preview

Import the pipeline

SnapGPT

and % complete is equal to or less than 40, then write records to csv file

CSV Filter and Write

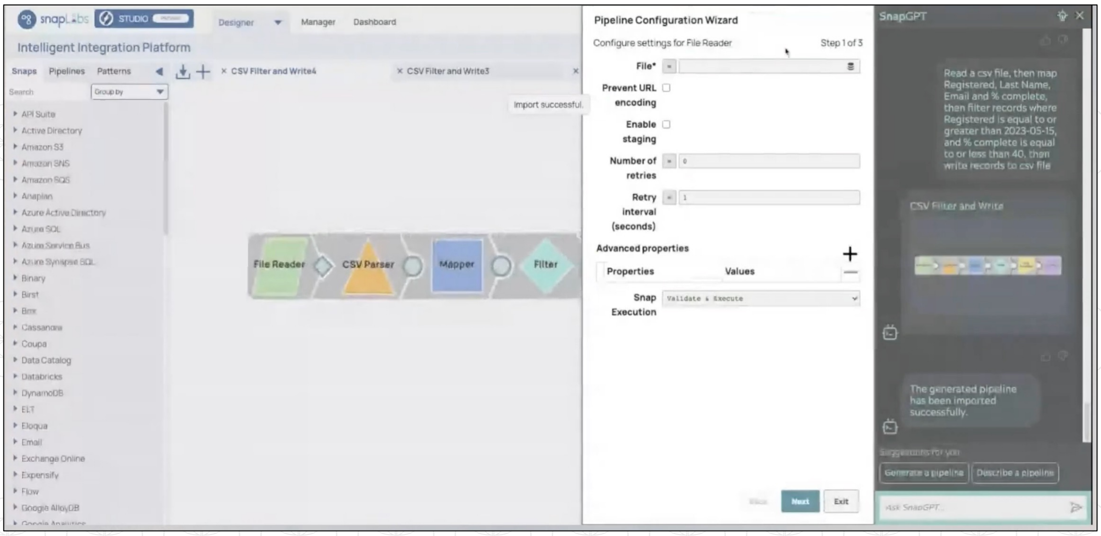
Read a csv file, then map Registered, Last Name, Email and % complete, then filter records where Registered is equal to or greater than 2023-05-15, and % complete is equal to or less than 40, then write records to csv file

Suggestions for you

Generate a pipeline | Describe a pipeline

Ask SnapGPT...

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



Import successful.

Pipeline Configuration Wizard

Configure settings for File Reader — Step 1 of 3

File*:

Prevent URL encoding: □

Enable staging: □

Number of retries: 0

Retry interval (seconds): 1

Advanced properties

Properties	Values

Snap Execution: Validate & Execute

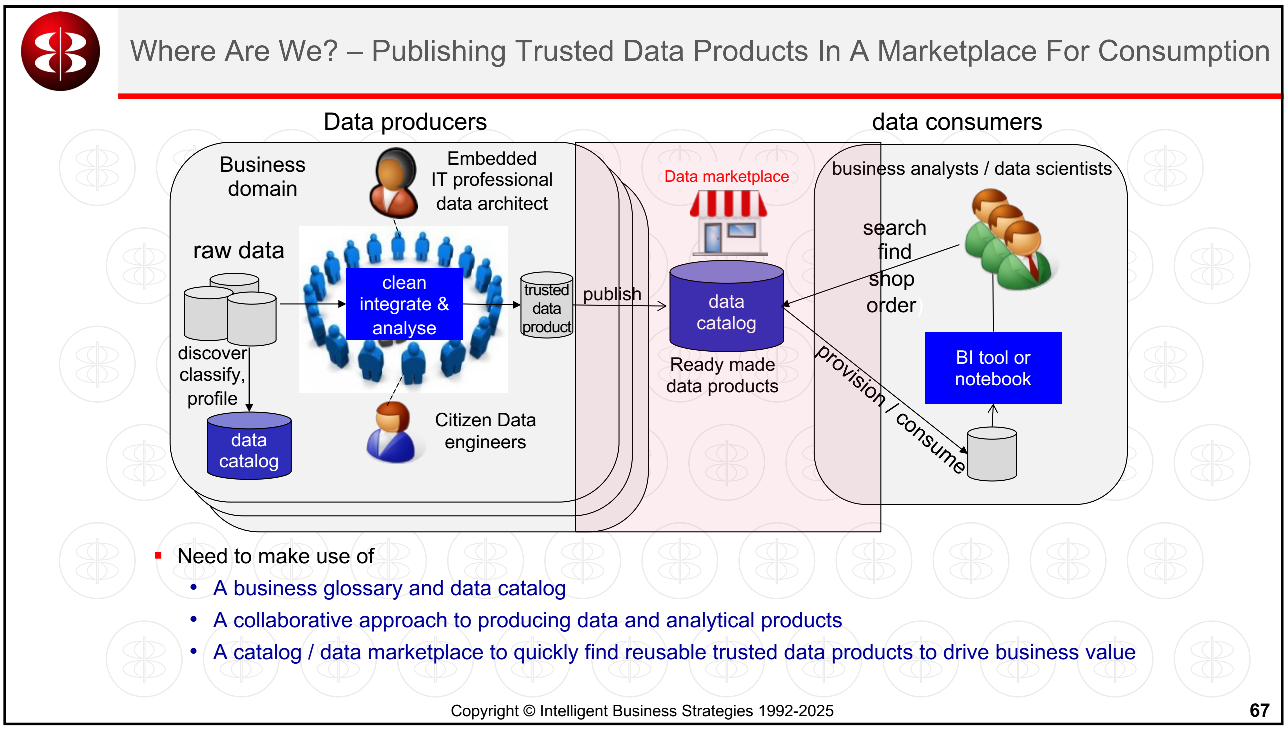
Read a csv file, then map Registered, Last Name, Email and % complete, then filter records where Registered is equal to or greater than 2023-05-15, and % complete is equal to or less than 40, then write records to csv file

The generated pipeline has been imported successfully.

Source: SnapLogic

Topics – Where Are We?

- What is a data catalogue?
- The data catalogue marketplace
- Using a business glossary within a data catalogue to define data products
- Using a data catalogue to automatically discover data in multiple data sources
- Using a data catalogue to map raw data to common terms in a business glossary
- Using a data catalogue to automatically detect sensitive data in data sources
- Using a data catalogue to automatically profile data quality in your data sources and recommend fixes
- The power of metadata - Integrating data engineering tools and generative AI with a data catalogue to rapidly build data pipelines to produce data products
- ➢ Publishing data products in a data marketplace within the data catalog

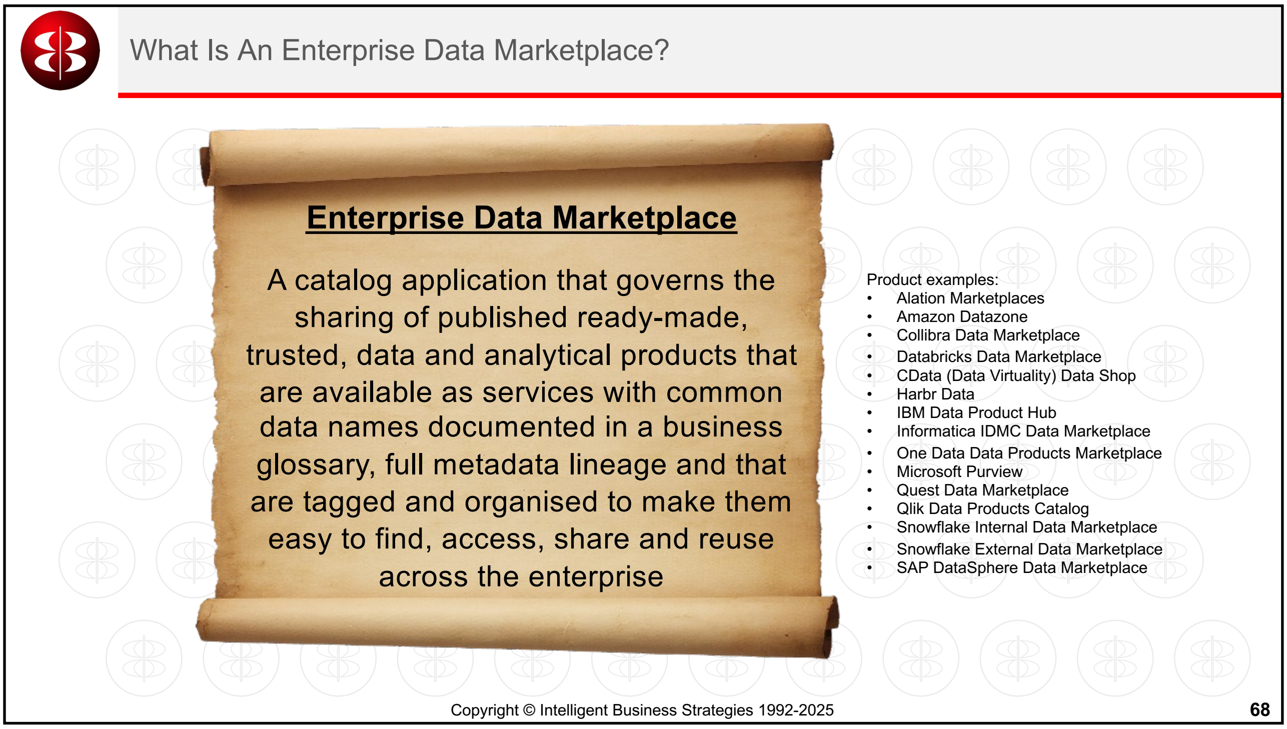


Where Are We? – Publishing Trusted Data Products In A Marketplace For Consumption

- Need to make use of
  - A business glossary and data catalog
  - A collaborative approach to producing data and analytical products
  - A catalog / data marketplace to quickly find reusable trusted data products to drive business value

Copyright © Intelligent Business Strategies 1992-2025

67



What Is An Enterprise Data Marketplace?

Enterprise Data Marketplace

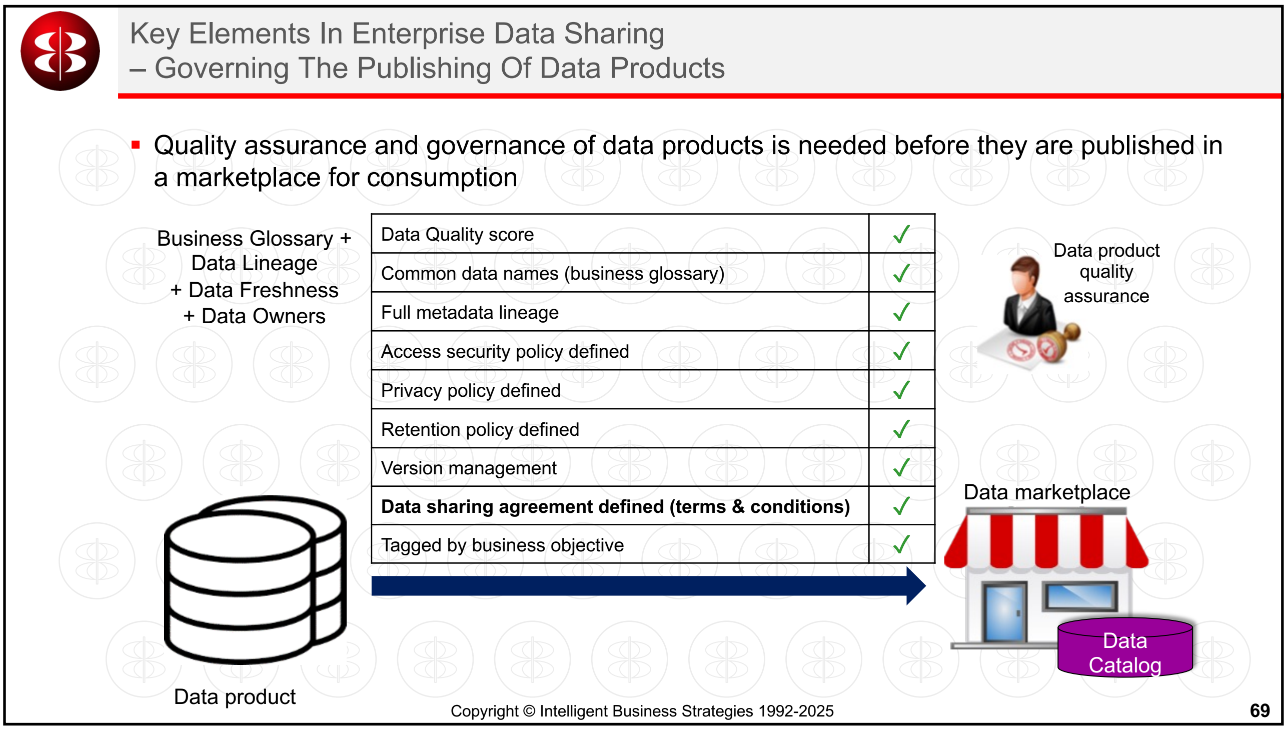
A catalog application that governs the sharing of published ready-made, trusted, data and analytical products that are available as services with common data names documented in a business glossary, full metadata lineage and that are tagged and organised to make them easy to find, access, share and reuse across the enterprise

Product examples:

- Alation Marketplaces
- Amazon Datazone
- Collibra Data Marketplace
- Databricks Data Marketplace
- CData (Data Virtuality) Data Shop
- Harbr Data
- IBM Data Product Hub
- Informatica IDMC Data Marketplace
- One Data Data Products Marketplace
- Microsoft Purview
- Quest Data Marketplace
- Qlik Data Products Catalog
- Snowflake Internal Data Marketplace
- Snowflake External Data Marketplace
- SAP DataSphere Data Marketplace

Copyright © Intelligent Business Strategies 1992-2025

68



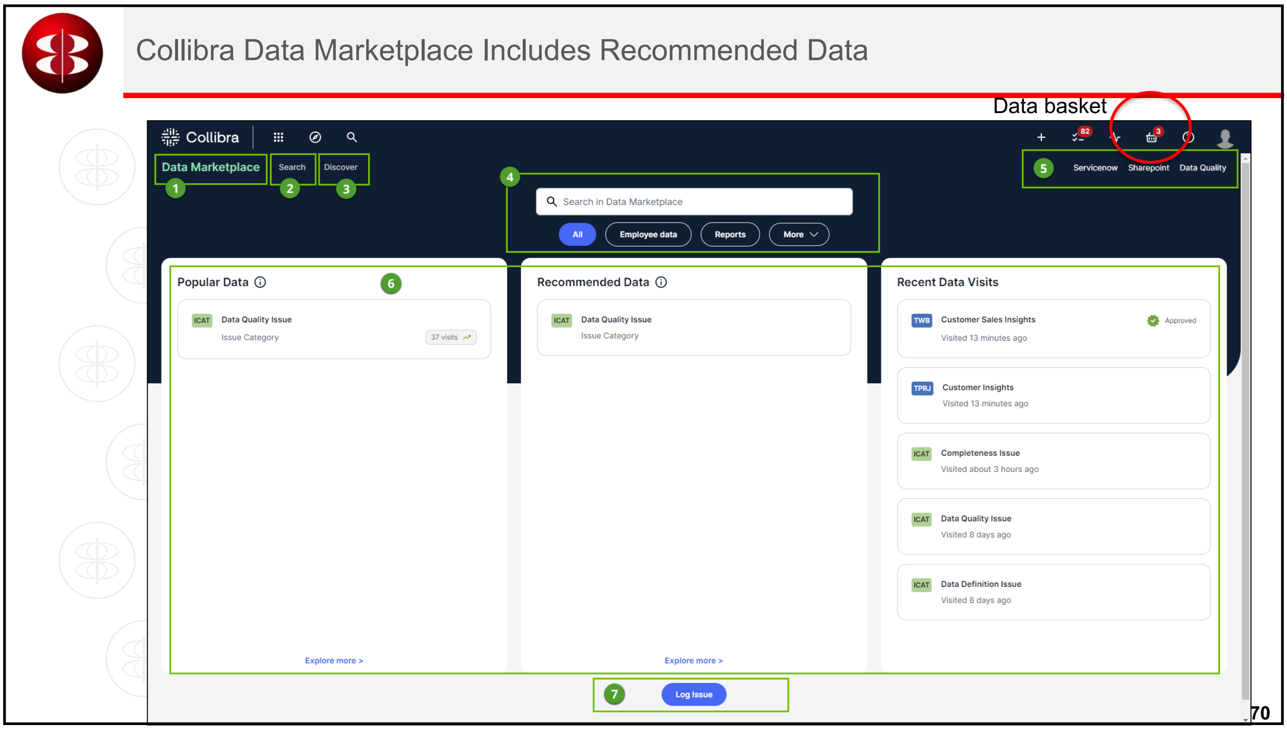
Key Elements In Enterprise Data Sharing – Governing The Publishing Of Data Products

- 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 (terms & conditions)	✓
Tagged by business objective	✓

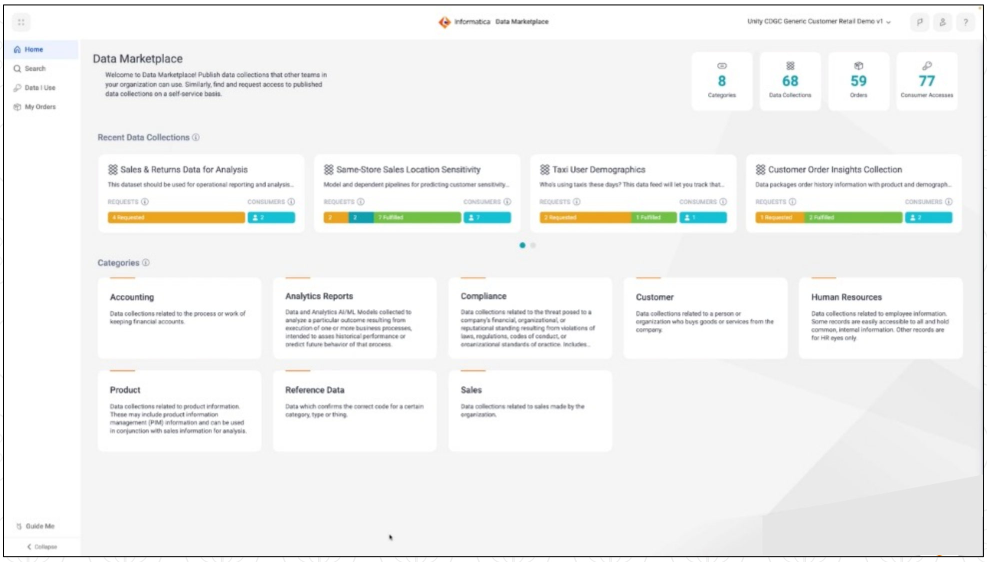
Copyright © Intelligent Business Strategies 1992-2025 69



Collibra Data Marketplace Includes Recommended Data

70

Information Consumers Shopping For Data In An Enterprise Data Marketplace – E.g. Informatica IDMC Data Marketplace Data Categories


Informatica Data Marketplace
Unity CDGC Generic Customer Retail Demo v1
Home
Search
Data I Use
My Orders
Guide Me
Collapse

Data Marketplace

Welcome to Data Marketplace! Publish data collections that other teams in your organization can use. Similarly, find and request access to published data collections on a self-service basis.

8 Categories | 68 Data Collections | 59 Orders | 77 Consumer Accesses

Recent Data Collections

Sales & Returns Data for Analysis
This dataset should be used for operational reporting and analysis...
REQUESTS: 4 Requested — CONSUMERS: 2

Same-Store Sales Location Sensitivity
Model and dependent pipelines for predicting customer sensitivity...
REQUESTS: 2 / 2 / 7 Fulfilled — CONSUMERS: 7

Taxi User Demographics
Who's using taxis these days? This data feed will let you track that...
REQUESTS: 2 Requested / 1 Fulfilled — CONSUMERS: 1

Customer Order Insights Collection
Data packages order history information with product and demograph...
REQUESTS: 1 Requested / 2 Fulfilled — CONSUMERS: 2

Categories

Accounting
Data collections related to the process or work of keeping financial accounts.

Analytics Reports
Data and Analytics AI/ML Models collected to analyze a particular outcome resulting from execution of one or more business processes, intended to asses historical performance or predict future behavior of that process.

Compliance
Data collections related to the threat posed to a company's financial, organizational, or reputational standing resulting from violations of laws, regulations, codes of conduct, or organizational standards of practice. Includes...

Customer
Data collections related to a person or organization who buys goods or services from the company.

Human Resources
Data collections related to employee information. Some records are easily accessible to all and hold common, internal information. Other records are for HR eyes only.

Product
Data collections related to product information. These may include product information management (PIM) information and can be used in conjunction with sales information for analysis.

Reference Data
Data which confirms the correct code for a certain category, type or thing.

Sales
Data collections related to sales made by the organization.

Audit history, reporting, analytics on data consumption patterns

Categories contain data collections

Data collections contain data assets

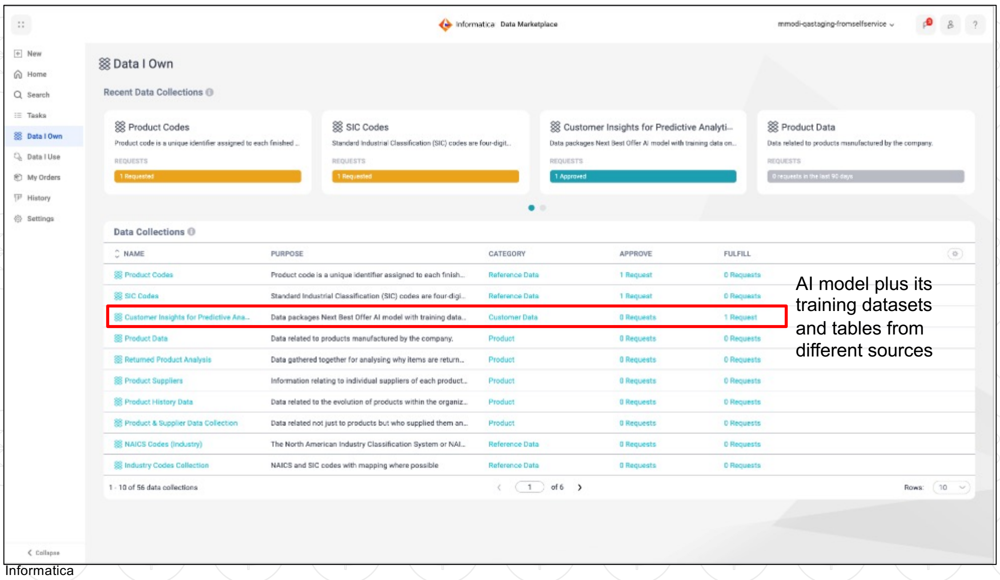
Data assets can contain one or more data elements

Source: Informatica

Copyright © Intelligent Business Strategies 1992-2025

71

Informatica IDMC Data Marketplace – Includes Data Products And Machine Learning Models



Data I Own

Recent Data Collections

Product Codes
Product code is a unique identifier assigned to each finished ...
REQUESTS: 1 Requested

SIC Codes
Standard Industrial Classification (SIC) codes are four-digit...
REQUESTS: 1 Requested

Customer Insights for Predictive Analyti...
Data packages Next Best Offer AI model with training data on...
REQUESTS: 1 Approved

Product Data
Data related to products manufactured by the company.
REQUESTS: 0 requests in the last 90 days

Data Collections

NAME	PURPOSE	CATEGORY	APPROVE	FULFILL
Product Codes	Product code is a unique identifier assigned to each finish...	Reference Data	1 Request	0 Requests
SIC Codes	Standard Industrial Classification (SIC) codes are four-digi...	Reference Data	1 Request	0 Requests
Customer Insights for Predictive Ana...	Data packages Next Best Offer AI model with training data...	Customer Data	0 Requests	1 Request
Product Data	Data related to products manufactured by the company.	Product	0 Requests	0 Requests
Returned Product Analysis	Data gathered together for analysing why items are return...	Product	0 Requests	0 Requests
Product Suppliers	Information relating to individual suppliers of each product...	Product	0 Requests	0 Requests
Product History Data	Data related to the evolution of products within the organiz...	Product	0 Requests	0 Requests
Product & Supplier Data Collection	Data related not just to products but who supplied them an...	Product	0 Requests	0 Requests
NAICS Codes (Industry)	The North American Industry Classification System or NAI...	Reference Data	0 Requests	0 Requests
Industry Codes Collection	NAICS and SIC codes with mapping where possible	Reference Data	0 Requests	0 Requests

1 - 10 of 56 data collections — 1 of 6 — Rows: 10

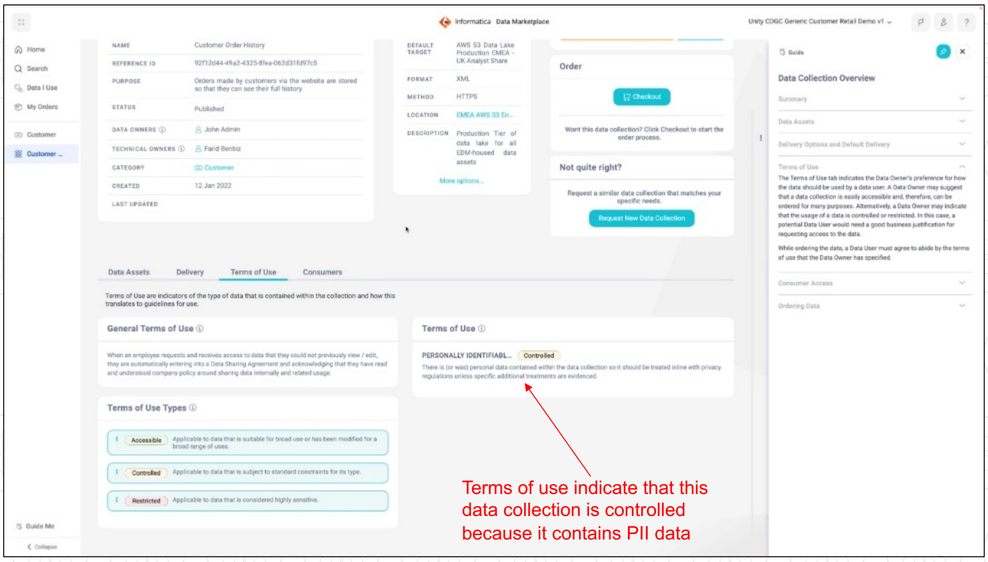
AI model plus its training datasets and tables from different sources

Source: Informatica

Copyright © Intelligent Business Strategies 1992-2025

72

Informatica IDMC Data Marketplace – Data Product Terms Of Use (Data Contract)



Informatica Data Marketplace
Unity CDGC Generic Customer Retail Demo v1
Home
Search
Data I Use
My Orders
Customer
Customer ...
Guide Me
Collapse

NAME	Customer Order History
REFERENCE ID	92f12d44-49a2-4325-8fea-062d31fd97c5
PURPOSE	Orders made by customers via the website are stored so that they can see their full history.
STATUS	Published
DATA OWNERS	John Admin
TECHNICAL OWNERS	Farid Benbiz
CATEGORY	Customer
CREATED	12 Jan 2022
LAST UPDATED	

DEFAULT TARGET	AWS S3 Data Lake Production EMEA - UK Analyst Share
FORMAT	XML
METHOD	HTTPS
LOCATION	EMEA AWS S3 En...
DESCRIPTION	Production Tier of data lake for all EDM-housed data assets

More options...

Order

Checkout

Want this data collection? Click Checkout to start the order process.

Not quite right?

Request a similar data collection that matches your specific needs.

Request New Data Collection

Data Assets | Delivery | Terms of Use | Consumers

Terms of Use are indicators of the type of data that is contained within the collection and how this translates to guidelines for use.

General Terms of Use

When an employee requests and receives access to data that they could not previously view / edit, they are automatically entering into a Data Sharing Agreement and acknowledging that they have read and understood company policy around sharing data internally and related usage.

Terms of Use Types

- Accessible — Applicable to data that is suitable for broad use or has been modified for a broad range of uses.
- Controlled — Applicable to data that is subject to standard constraints for its type.
- Restricted — Applicable to data that is considered highly sensitive.

Terms of Use

PERSONALLY IDENTIFIABL... Controlled

There is (or was) personal data contained within the data collection so it should be treated inline with privacy regulations unless specific additional treatments are evidenced.

Guide

Data Collection Overview

Summary

Data Assets

Delivery Options and Default Delivery

Terms of Use

The Terms of Use tab indicates the Data Owner's preference for how the data should be used by a data user. A Data Owner may suggest that a data collection is easily accessible and, therefore, can be ordered for many purposes. Alternatively, a Data Owner may indicate that the usage of a data is controlled or restricted. In this case, a potential Data User would need a good business justification for requesting access to the data.

While ordering the data, a Data User must agree to abide by the terms of use that the Data Owner has specified.

Consumer Access

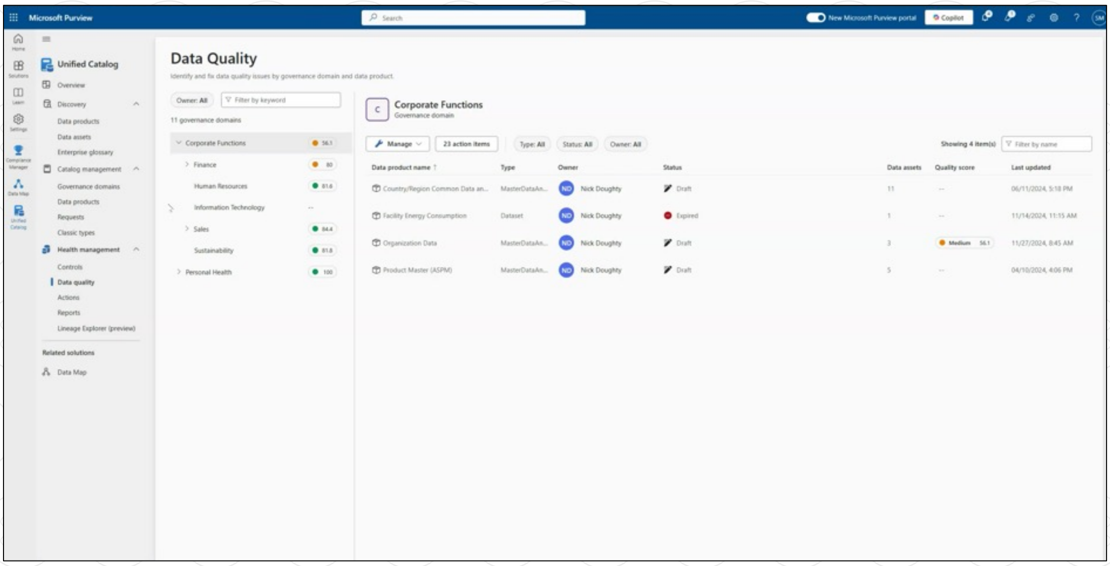
Ordering Data

Terms of use indicate that this data collection is controlled because it contains PII data

Source: Informatica

Some Data Catalogs Will Now Let You View Data Quality Scores By Business Domain and By Data Products – E.g. Microsoft Purview

In this case there is a hierarchy showing Corporate Functions and then business domains



Data Quality

Identify and fix data quality issues by governance domain and data product.

Owner: All | Filter by keyword

11 governance domains

Corporate Functions	56.1
Finance	80
Human Resources	81.6
Information Technology	--
Sales	84.4
Sustainability	81.8
Personal Health	100

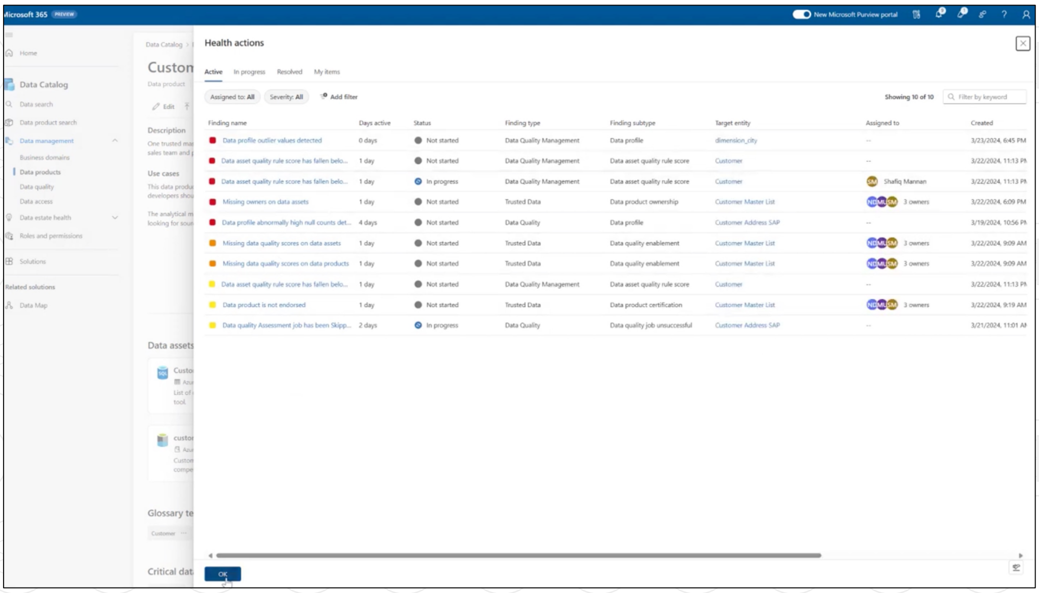
Corporate Functions
Governance domain

Manage | 23 action items | Type: All | Status: All | Owner: All | Showing 4 item(s) | Filter by name

Data product name ↑	Type	Owner	Status	Data assets	Quality score	Last updated
Country/Region Common Data an...	MasterDataAn...	Nick Doughty	Draft	11	--	06/11/2024, 5:18 PM
Facility Energy Consumption	Dataset	Nick Doughty	Expired	1	--	11/14/2024, 11:15 AM
Organization Data	MasterDataAn...	Nick Doughty	Draft	3	Medium 56.1	11/27/2024, 8:45 AM
Product Master (ASPM)	MasterDataAn...	Nick Doughty	Draft	5	--	04/10/2024, 4:06 PM

Source: Microsoft

Microsoft Purview - You Can Also Monitor Health Actions Associated With Data Products



Health actions

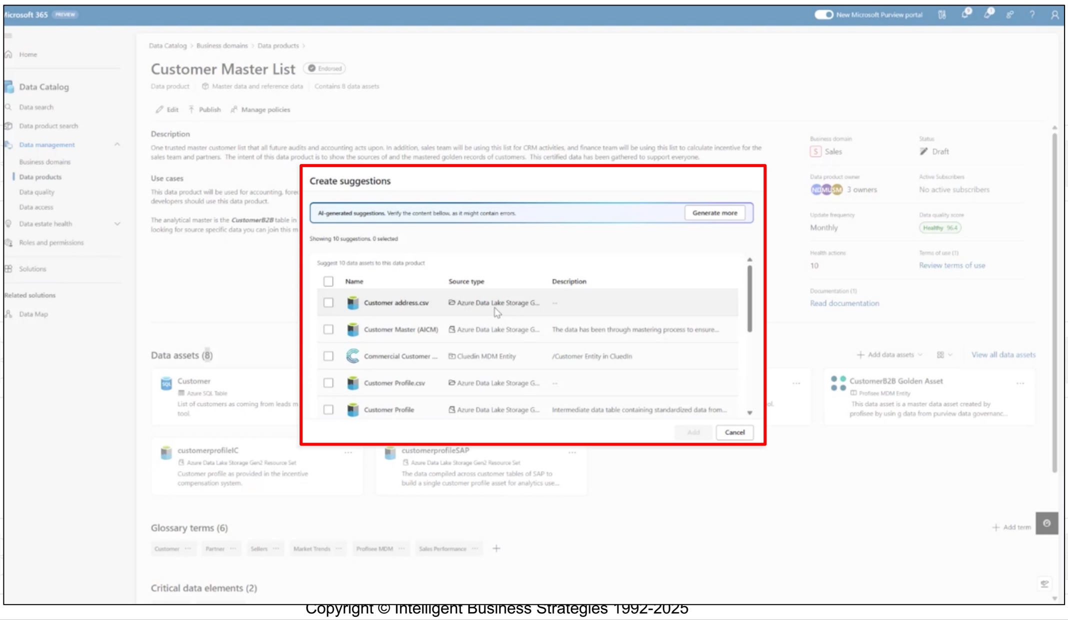
Active | In progress | Resolved | My items

Assigned to: All · Severity: All · Add filter

Showing 10 of 10 · Filter by keyword

Finding name	Days active	Status	Finding type	Finding subtype	Target entity	Assigned to	Created
🟥 Data profile outlier values detected	0 days	● Not started	Data Quality Management	Data profile	dimension_city	--	3/23/2024, 6:45 PM
🟥 Data asset quality rule score has fallen belo...	1 day	● Not started	Data Quality Management	Data asset quality rule score	Customer	--	3/22/2024, 11:13 PM
🟥 Data asset quality rule score has fallen belo...	1 day	● In progress	Data Quality Management	Data asset quality rule score	Customer	Shafiq Mannan	3/22/2024, 11:13 PM
🟥 Missing owners on data assets	1 day	● Not started	Trusted Data	Data product ownership	Customer Master List	3 owners	3/22/2024, 6:09 PM
🟥 Data profile abnormally high null counts det...	4 days	● Not started	Data Quality	Data profile	Customer Address SAP	--	3/19/2024, 10:56 PM
🟧 Missing data quality scores on data assets	1 day	● Not started	Trusted Data	Data quality enablement	Customer Master List	3 owners	3/22/2024, 9:09 AM
🟧 Missing data quality scores on data products	1 day	● Not started	Trusted Data	Data quality enablement	Customer Master List	3 owners	3/22/2024, 9:09 AM
🟨 Data asset quality rule score has fallen belo...	1 day	● Not started	Data Quality Management	Data asset quality rule score	Customer	--	3/22/2024, 11:13 PM
🟨 Data product is not endorsed	1 day	● Not started	Trusted Data	Data product certification	Customer Master List	3 owners	3/22/2024, 9:19 AM
🟨 Data quality Assessment job has been Skipp...	2 days	● In progress	Data Quality	Data quality job unsuccessful	Customer Address SAP	--	3/21/2024, 11:01 AM

Microsoft Purview - Generative AI Can Suggest New Assets to Add To a Data Products



Data Catalog > Business domains > Data products >

Customer Master List Endorsed

Data product | Master data and reference data | Contains 8 data assets

Edit | Publish | Manage policies

Description

One trusted master customer list that all future audits and accounting acts upon. In addition, sales team will be using this list for CRM activities, and finance team will be using this list to calculate incentive for the sales team and partners. The intent of this data product is to show the sources of and the mastered golden records of customers. This certified data has been gathered to support everyone.

Business domain: Sales · Status: Draft · Data product owner: 3 owners · Active Subscribers: No active subscribers · Update frequency: Monthly · Data quality score: Healthy 96.4 · Health actions: 10 · Terms of use (1): Review terms of use · Documentation (1): Read documentation

Create suggestions

AI-generated suggestions. Verify the content bellow, as it might contain errors. [Generate more]

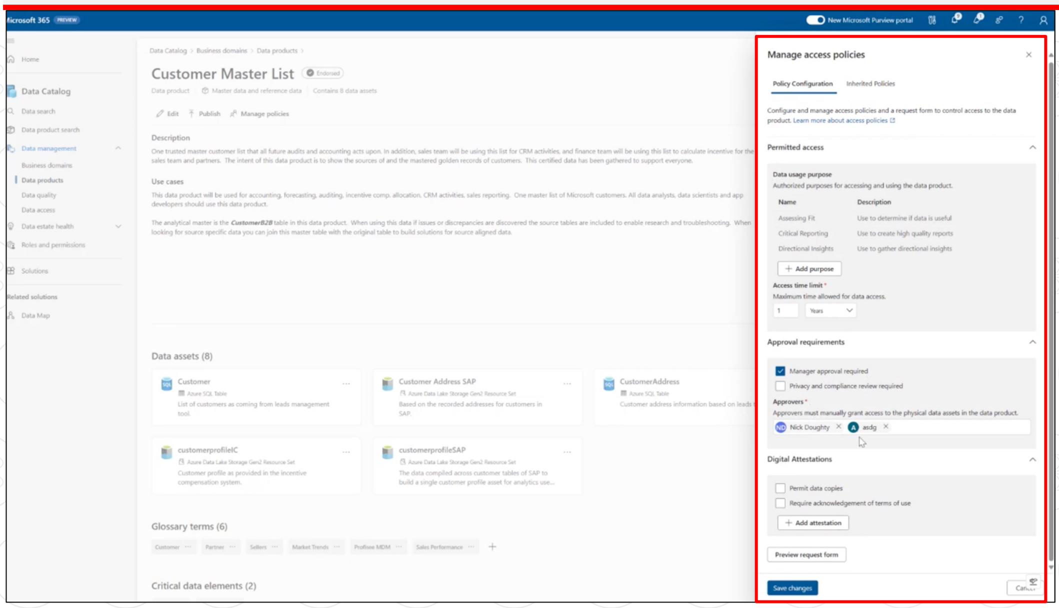
Showing 10 suggestions, 0 selected

Suggest 10 data assets to this data product

☐	Name	Source type	Description
☐	Customer address.csv	Azure Data Lake Storage G...	--
☐	Customer Master (AICM)	Azure Data Lake Storage G...	The data has been through mastering process to ensure...
☐	Commercial Customer ...	Cluedin MDM Entity	/Customer Entity in CluedIn
☐	Customer Profile.csv	Azure Data Lake Storage G...	--
☐	Customer Profile	Azure Data Lake Storage G...	Intermediate data table containing standardized data from...

Data assets (8) · Glossary terms (6) · Critical data elements (2)

Microsoft Purview - Data Product Owners Can Also Set Policies Including the Duration of the Data Access Policy



Customer Master List Endorsed

Data product | Master data and reference data | Contains 8 data assets

Description

One trusted master customer list that all future audits and accounting acts upon. In addition, sales team will be using this list for CRM activities, and finance team will be using this list to calculate incentive for the sales team and partners. The intent of this data product is to show the sources of and the mastered golden records of customers. This certified data has been gathered to support everyone.

Use cases

This data product will be used for accounting, forecasting, auditing, incentive comp. allocation, CRM activities, sales reporting. One master list of Microsoft customers. All data analysts, data scientists and app developers should use this data product.

The analytical master is the ***CustomerB2B*** table in this data product. When using this data if issues or discrepancies are discovered the source tables are included to enable research and troubleshooting. When looking for source specific data you can join this master table with the original table to build solutions for source aligned data.

Data assets (8)

- Customer — Azure SQL Table — List of customers as coming from leads management tool.
- Customer Address SAP — Azure Data Lake Storage Gen2 Resource Set — Based on the recorded addresses for customers in SAP.
- CustomerAddress — Azure SQL Table — Customer address information based on leads
- customerprofileIC — Azure Data Lake Storage Gen2 Resource Set — Customer profile as provided in the incentive compensation system.
- customerprofileSAP — Azure Data Lake Storage Gen2 Resource Set — The data compiled across customer tables of SAP to build a single customer profile asset for analytics use...

Glossary terms (6)

Customer · Partner · Sellers · Market Trends · Profisee MDM · Sales Performance

Critical data elements (2)

Manage access policies

Policy Configuration | Inherited Policies

Configure and manage access policies and a request form to control access to the data product. Learn more about access policies

Permitted access

Data usage purpose
Authorized purposes for accessing and using the data product.

Name	Description
Assessing Fit	Use to determine if data is useful
Critical Reporting	Use to create high quality reports
Directional Insights	Use to gather directional insights

+ Add purpose

Access time limit *
Maximum time allowed for data access.
1 Years

Approval requirements

☑ Manager approval required
☐ Privacy and compliance review required

Approvers *
Approvers must manually grant access to the physical data assets in the data product.
Nick Doughty, asdg

Digital Attestations

☐ Permit data copies
☐ Require acknowledgement of terms of use

+ Add attestation

Preview request form

Save changes | Cancel

Copyright © Intelligent Business Strategies 1992-2025

77

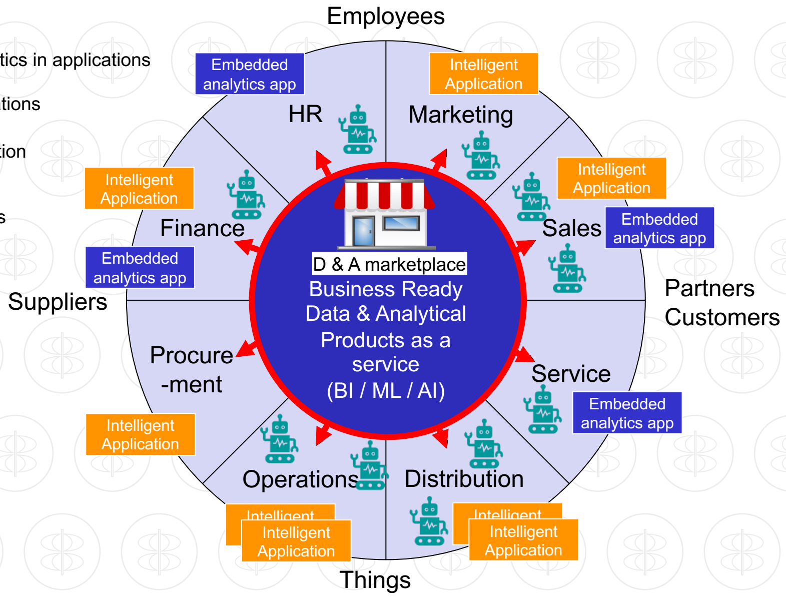
The Last Mile - BI, ML And AI Are 'Wired Into' Every Process At All Levels To Enable Timely, Guided And Automated Decisions To Maximise Value – Create A Closed Loop

Embedded analytics in applications

Intelligent applications

Decision automation

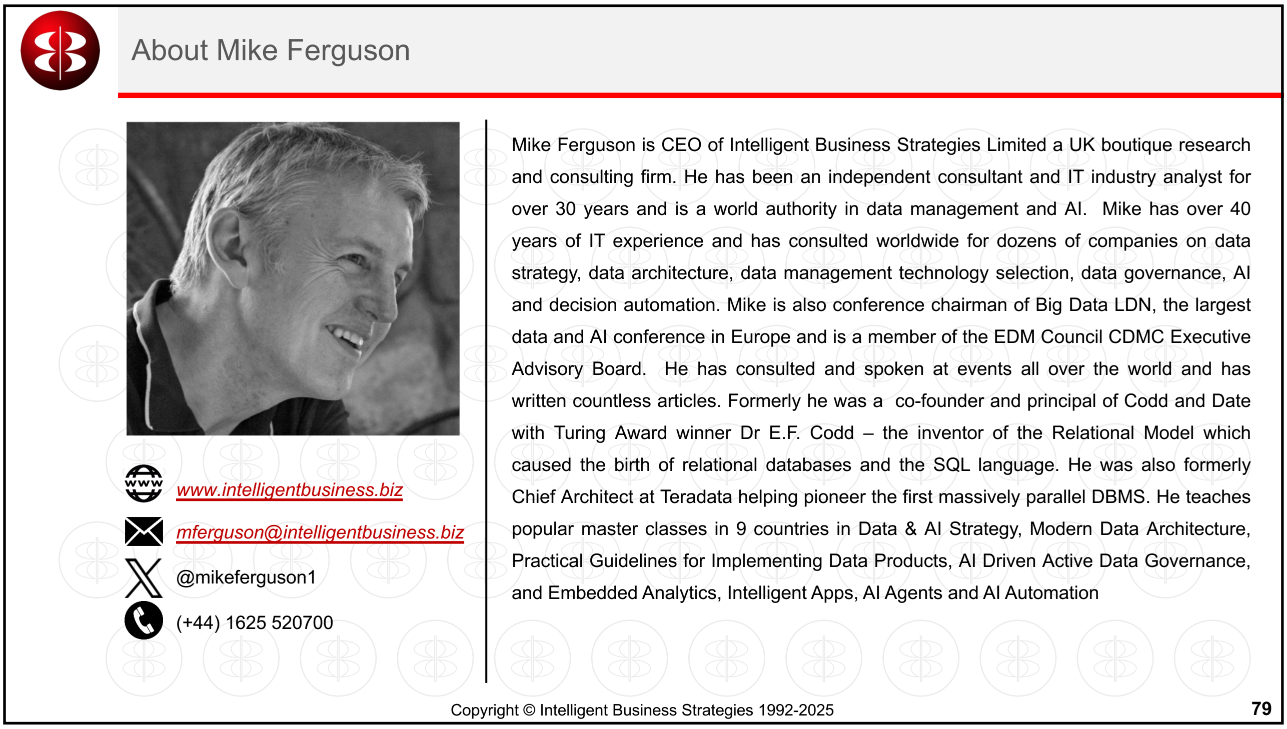
AI Agents & AI-automated tasks



Access to commonly understood, data, BI reports, ML models, AI services across the business is critical to success

Copyright © Intelligent Business Strategies 1992-2025

78



About Mike Ferguson

www.intelligentbusiness.biz

mferguson@intelligentbusiness.biz

@mikeferguson1

(+44) 1625 520700

Mike Ferguson is CEO of Intelligent Business Strategies Limited a UK boutique research and consulting firm. He has been an independent consultant and IT industry analyst for over 30 years and is a world authority in data management and AI. Mike has over 40 years of IT experience and has consulted worldwide for dozens of companies on data strategy, data architecture, data management technology selection, data governance, AI and decision automation. Mike is also conference chairman of Big Data LDN, the largest data and AI conference in Europe and is a member of the EDM Council CDMC Executive Advisory Board. He has consulted and spoken at events all over the world and has written countless articles. Formerly he was a co-founder and principal of Codd and Date with Turing Award winner Dr E.F. Codd – the inventor of the Relational Model which caused the birth of relational databases and the SQL language. He was also formerly Chief Architect at Teradata helping pioneer the first massively parallel DBMS. He teaches popular master classes in 9 countries in Data & AI Strategy, Modern Data Architecture, Practical Guidelines for Implementing Data Products, AI Driven Active Data Governance, and Embedded Analytics, Intelligent Apps, AI Agents and AI Automation