

WHITE PAPER

Semantic Intelligence:

The Value of Trusted, Secure Data and Unified Semantics in the Era of AI

By Mike Ferguson
Intelligent Business Strategies
February 2026

Research sponsored by:



Table of Contents

Introduction3

What is Slowing Progress Towards AI Success?4

 Data and metadata hindrances4

 AI Hindrances6

 What kind of issues are caused by these hindrances6

Requirements For Semantic Intelligence in an Artificially Intelligent Enterprise8

 Unified AI powered data and AI governance requirements8

 Knowledge bases12

 Enterprise semantic layer, knowledge graph and context12

How is Collate Stepping Up to Meet the Requirements?13

 The Collate Semantic Intelligence Platform.....13

 The Collate Semantic Intelligence Platform deployment options14

 Platform components and capabilities14

Conclusion21

INTRODUCTION

Discussions around business transformation seem to be dominated by AI

In almost every vertical industry today, discussions around business transformation seem to be dominated by the potential use of AI. More specifically, AI agents are now the focus as companies look to accelerate and redesign their business processes. Not surprisingly, Chief Operating Officers and Chief Financial Officers have leapt on this, initially funding the use of AI to automate tasks, improve operational business process efficiency and reduce costs. Innovation is also high in priority.

AI has made it easier for many more business users to use self-service analytics

In analytics, AI is also at work. Conversational self-service analytics is now available to everyday business users who can interact with software using voice and natural language to query and analyse data. AI Agents can also be used to automatically analyse data on a user's behalf to help quickly identify opportunities and avoid problems by enabling them to make more timely and effective decisions. AI is also helping to democratise data science with 'citizen data scientists' now able to use AI-powered no-code tools along with ML-Ops and Agent-Ops to simplify and accelerate development and deployment of ML models and AI Agents.

AI needs access to contextually relevant, high-quality data, which makes data governance critically important

There is no doubt that companies are already seeing measurable results from AI and that its promise is huge. However, they also know that to be successful with AI, that agents need access to contextually relevant, secure, high-quality data. As a result, the AI boom has put the spotlight back on data governance with demand for quality, secure data being very high. Only, one thing is of even higher priority and that is the need for AI agents to also understand the meaning of that data, as well.

Companies need to adopt a unified approach to data governance and create AI ready data products

For these reasons, companies need to use AI to accelerate and automate data management tasks and adopt an AI-powered unified approach to governing data, preferably on a continuous basis. They also need to accelerate the development of a high quality, secure, business and AI-ready, reusable data products to build a data foundation for AI.

AI understanding of data meaning and data relationships is even higher priority and critical to success

In addition, AI agents need to be provided with and understand the meaning of all relevant data, and data relationships across the enterprise as context to ensure they avoid hallucinations and can be trusted.

On the surface, this looks like a massive challenge. But it can be done. To understand how, this paper looks at what on the data and AI side is hindering progress to achieving success with AI, what issues these hindrances are causing, what the requirements are to overcome them, and what one vendor, Collate, is doing to help companies succeed.

WHAT IS SLOWING PROGRESS TOWARDS AI SUCCESS?

There are several data, metadata and AI related things slowing down companies' progress towards adopting and using data and AI to drive business value. Some of the major ones are described below.

DATA AND METADATA HINDRANCES

The Growing Complexity of a Distributed Data Estate

The first difficulty is the complexity caused by a distributed data estate. Today, most enterprises have data scattered across many different data stores as shown in Figure 1 which makes data hard to find and govern.

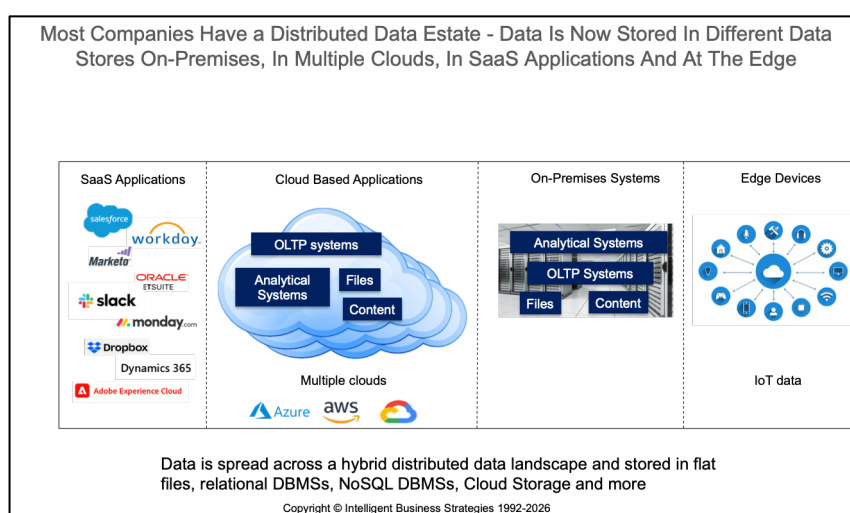


Figure 1

Data Redundancy

In addition to all the different types of data stores that exist in this distributed data estate, there is also the problem of data redundancy with the same or similar data existing across multiple different data stores. The extent of this for most companies has historically been unknown and the move to the cloud (or multiple clouds) has increased data redundancy as people copy data between on-premises and cloud-based systems.

More Data and More Challenging Data Types

Also, over the last several years, the number of data sources has continued to increase both within the enterprise and from external data providers. This has been driven by the thirst to understand more about:

- Customers including customer interactions, on-line behaviour, social relationships, etc.
- Suppliers to improve performance and resilience in supply chains
- Business operations through instrumentation in processes
- Unstructured data to add insights to what companies already know

This adds to the overall complexity and possible redundancy if that data is needed in SaaS applications, as well as operational and analytical systems residing in multiple clouds.

Most organisations now have data scattered across a distributed data estate

Data redundancy has increased with the move to the cloud

The demand for more data continues to grow

Demand for unstructured data is growing rapidly as people and AI look to leverage knowledge to make better decisions

Metadata Complexity

Metadata complexity is increasing

In addition to the complexity around data, there is also growing complexity around metadata. Metadata complexity is caused by companies buying different tools that generate and store metadata in different proprietary metadata repositories. It is also caused by the need to share metadata across multiple different tools. For example, if a company has purchased a data catalog and separate tools for data quality, lineage, data access security, etc., from different suppliers, with each tool having its own proprietary metadata repository and metamodel, then time to value can slow and progress may become more difficult and expensive. This is shown in Figure 2 where a data catalog and multiple independent tools are being used to govern data quality, data access security, etc, across a distributed data estate. This approach often results in source data discovery scans and source data classification being done repeatedly by each tool. Also, some policies may be defined and stored in a data catalog and others in different tools. Synchronisation of metadata across tools to share metadata is also a challenge because there is no guaranteed interoperability between tools to share metadata in a consistent way.

The use of multiple stand-alone data governance and data management tools and also no standard to exchange metadata, is causing metadata silos, metadata complexity and rework

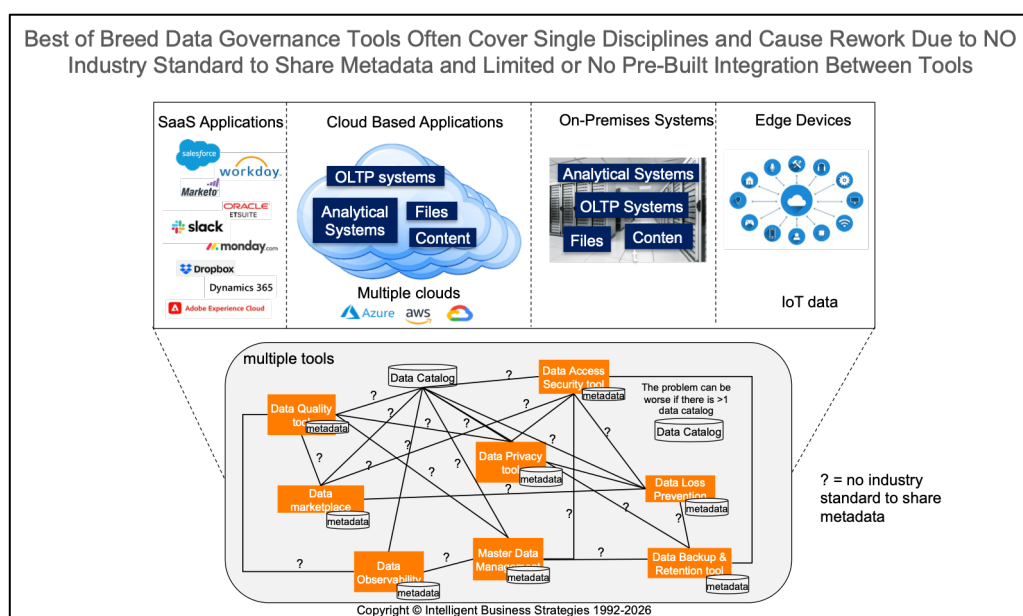


Figure 2

Data Governance is Fragmented

As a result, data governance is fragmented

Building on Figure 1 and Figure 2, because of multiple different tools, many organisations also have a fragmented approach to data governance. Many different people are typically involved in setting and maintaining policies and rules across different tools and systems to deal with data quality and privacy issues and to govern data access and maintenance entitlements. Data governance, therefore, is not automated, prone to inconsistency, and not fully trusted. Enterprises are reliant on different administrators to govern different data stores using different tools that do not integrate.

AI-powered conversational data engineering may have democratised this capability, but many people are still struggling to find data and understand its meaning

Citizen Data Engineers Have Difficulty Finding Data and Understanding Semantics

The emergence of AI co-pilots in data engineering tools has enabled no-code / low-code conversational data engineering using natural language to guide the generation of data engineering pipelines and code. This has simplified data integration pipeline development and enabled democratisation of data engineering to 'citizen' data engineers. However, many are often still unaware of where the data they need is located. They may also be unaware of source data quality and any sensitive data they are obligated to protect. In addition, many still struggle with data meaning.

AI HINDRANCES

Sporadic approaches to making data AI ready is resulting in a proliferation of RAG pipelines and vector databases

A wide range of agent building tools and 'semantic layers' are emerging within the enterprise

There is no common understanding of data concepts and semantics for all AI agents to use

With respect to AI, companies have rapidly started implementing but with governance often coming later. So far, we are seeing:

- Many different sporadic approaches being taken to making data AI ready. These may be database specific (e.g., adopting vector data types and functions in a specific DBMS), many different RAG pipelines with many different vector databases and they may or may not be connected to the building of data products
- AI Agents are being developed using a wide range of Agent building tools with no guarantee of common understanding of data concepts and data semantics being in place across those tools
- Multiple 'semantic layers' are being introduced by many different application and software tool vendors to provide context for pre-built AI agents which can lead to inconsistent understanding among agents and the potential for hallucinations. This includes semantic layers in BI tools and other tools where business logic is defined. However, most companies want shared understanding of meaning across people and AI for consistent decision making

All of this is going on, while the pace continues to quicken in enterprises without common approaches being in place.

WHAT KIND OF ISSUES ARE CAUSED BY THESE HINDRANCES

Fragmented data governance can lead to inconsistent, conflicting and overlapping policies being created in multiple different tools and databases

Policies may also be missing which increases risk

There are several issues that arise because of the hindrances discussed.

Data Governance Issues

With respect to data governance, the fragmented approaches discussed are not sufficient for the AI era where the need for well governed data underpins AI success. For a start, with fragmentation, the definition of policies and rules is almost entirely done manually by different people using different tools. These policies and rules reside in multiple different places and therefore run the risk of being potentially inconsistent (policy drift) and may conflict across systems. Also, the scope of policy enforcement may not cover all necessary data stores and so lead to potentially incomplete governance actions. In addition, for a specific discipline (e.g. data access security), incomplete or inconsistent

governance caused by policies existing in some tools and systems and not in others raises the risk of things like data breaches occurring.

Problem detection is likely to be slow, overly complex and will not scale

Furthermore, with fragmentation, problem detection is likely to be slow, it will not scale, and there is less ability to continuously and automatically detect policy violations, pipeline failures and data anomalies across a distributed data estate. It is also unlikely that it is possible to consistently automate policy recommendations and data governance actions to continuously improve data quality, privacy, access security, etc. Even if automation can be achieved, there is a risk that different independent tools could make recommendations and take automated actions that may not be consistent because policies conflict or are incomplete.

Data Engineering Issues

Citizen data engineering, with no easy way to locate data and understand what it means, can lead to chaos, especially if they are using multiple different tools

With respect to AI powered, conversational citizen data engineering, there is no question that this is resulting in more citizen data engineers emerging and more rapid production of data engineering pipelines. However, if they have no easy way of understanding what data is available, where it is, what other related data may be of interest and what that data means, then it could easily lead to the creation of inconsistent and overlapping data products that may not be reusable or AI ready. We could repeat the mistakes of over a decade ago, when self-service data preparation was introduced and we ended up with a 'wild west' approach to data engineering that resulted in data inconsistencies and unnecessary repetition of effort.

AI Issues

Having multiple semantic layers could cause potential inconsistency in understanding data meaning and context across multiple AI agents

With AI, the danger of multiple siloed semantic layers is that they could cause potential inconsistency in data meaning, understanding and context across multiple different AI agents. In addition, there is no guarantee that all semantic layers will know about all relevant information that might be needed by AI-Agents. Also, a lack of a persistent semantic memory for AI and human workflows can cause repeated rediscovery, revalidation, and rework.

All this breeds distrust in data, potentially unreliable results, and slows down AI initiatives.

REQUIREMENTS FOR SEMANTIC INTELLIGENCE IN AN ARTIFICIALLY INTELLIGENT ENTERPRISE

There are several data governance and semantics requirements that need to be met if we are to succeed with AI

Having understood the hindrances inhibiting progress and the issues they cause, this next section of the paper looks at requirements needed to overcome these issues and deliver well governed high-quality data and semantic intelligence to support AI across the enterprise. Note that Semantic Intelligence is enabling of an enterprise semantic layer that includes an ontology (a model) of an organisation's data concepts (entities), their properties and the relationships between them that can be used to organise structured and unstructured data in a graph to provide consistent meaning and context to both people and AI.

As there are several requirements associated with data governance, consistent semantics and context, they are categorised for easier reading.

UNIFIED AI POWERED DATA AND AI GOVERNANCE REQUIREMENTS

Companies need a common platform that offers unified data governance and semantics

To address the need to govern data and AI it should be possible for companies to use a common AI-assisted metadata platform that enables data cataloguing and supports unified data and AI governance. By unified data and AI governance we mean the governance of data quality, data privacy, data access security, data retention, data sharing and data usage as well as AI governance. This includes data observability to continuously monitor data to identify data anomalies and incidents. It also includes common semantics to enable consistent understanding of data meaning. Given this definition, the platform should include support for:

Business Glossary Creation and Maintenance

- Import and manual definition of glossary terms and descriptions
- AI-assisted generation of glossary terms and descriptions from data structures and usage patterns
- Governing glossary changes using term states and approval workflows

Enterprise Data Discovery and Cataloguing

- Connectivity to a wide range of data sources, packaged application metadata APIs and tools
- AI-powered automatic discovery and cataloguing of structured, semi-structured, and unstructured data across multiple clouds, on-premises systems, SaaS applications and the edge
- Continuous scanning of your distributed data estate for new, changed, or deprecated data assets
- Identification of "dark data" and unused assets

AI-assisted business glossary creation and maintenance processes

AI-powered automatic data discovery and cataloguing of structured and unstructured data

Automated mapping of data to glossary terms and generation of asset descriptions to improve understanding of business meaning and relationships and to build a semantic graph

- Automatic detection of schema drift and changing data structures

Semantic Understanding of Data

- Automatically infer meaning from and map discovered data assets to business glossary terms using AI to understand data meaning, intent and relationships within data
- AI-powered generation of business descriptions of discovered data assets to enrich metadata semantics and improve understanding of data meaning for both humans and AI agents
- Identify relationships between datasets across your data estate
- Build a persistent semantic graph linking data assets, business concepts, terms, classifications, policies, lineage and usage

Automated detection and labelling of sensitive data to govern data privacy and restrict access

Data Governance Classification - Privacy, Access & Retention

- AI-powered automated detection and labelling of country-specific sensitive data in structured and unstructured data sources to know where that data is located so it can be protected in line with legislative and regulatory obligations
- Ability to support jurisdiction-specific classifications
- User defined classification schemes to label data according to different levels of confidentiality and retention
- Trainable classifiers to enable automated classification of data and content across your distributed data estate using user defined confidentiality and retention labels
- Ability to push down classification labels to underlying data sources

Trainable classifiers so you can label your data using your own confidentiality and retention labels to govern data access and retention

Data Quality Profiling

- AI-powered, automated data quality profiling of structured, semi-structured and unstructured data to identify anomalies in data. These capabilities support continuous measurement and enforcement of data quality as a foundational element of unified data governance.

Automated profiling of data to understand its quality

Lineage Derivation and Data Tiering

- Automated derivation of metadata lineage from data and pipelines to understand data origin and downstream dependencies, to help determine the impact of detected data anomalies and schema changes on dependent artefacts
- Automated analysis of usage, lineage, and impact to assign data criticality level labels (tiers) to discovered data assets
- Lineage and tiering are also used to assess business impact and prioritise response to detected data incidents.

Automated lineage derivation to understand how data is processed and to understand the impact of data anomalies

Automated tiering to label critical data

Natural Language Data Exploration and Insights

- The ability to find data using natural language conversation and AI assisted search
- The ability to provide data insights in the form of interactive reports to help you govern the data in your data estate

Make data easy to find using natural language queries and AI assistance

Data Governance AI Agents

Speed up our ability to govern data using AI Agents to assist us

Enable AI Agents to easily access metadata and semantics

- Configurable out-of-the-box data governance co-pilots and AI-Agents to help teams govern data and get work done faster:
 - Interact with using natural language e.g. to find data, define and test policies, approve actions, report data health
- Enable out-of-the-box and 3rd party AI Agents to easily access metadata and semantics via Model Context Protocol (MCP)
- Tooling to build your own data governance agents

Data Governance Health Metric, Policy and Rule Creation and Maintenance

Health metrics and thresholds to measure and monitor how well we are governing data

Natural language data governance policy and rule creation

Policy grouping to make data governance easier

- Ability to define health metrics and severity thresholds to govern data
- Ability to create and test data privacy, data access security, data retention and data sharing policies and data quality rules in natural language to define how to validate and govern data across your data estate using an AI-assistant and graphical user interface
- Ability to group together data governance policies in data contracts to govern data quality, privacy, access security, retention, sharing and usage of data products

Data Observability, Incident Detection and Triage

Data observability to continuously monitor and detect data governance incidents

AI-assisted triage and prioritisation of incidents to deal with the most severe issues on critical data first

Automatically notify stewards and recommend actions

Trigger automated actions to reduce risk

- Out-of-the-box (or integration with) data observability to continuously monitor data processing and user/agent activity across different data stores and tools to detect data governance incidents and conditions e.g.,
 - Data quality rule violations, data freshness issues, schema drift and sensitive data processing in executing data pipelines
 - Unprotected sensitive data, unauthorised data access, data loss policy violation, data retention expiry, suspicious user behaviour...
 - User and agent activity against control metric thresholds
- Use AI Agents to automatically:
 - Triage incidents by analysing their impact using lineage, health metrics and thresholds
 - Prioritising them using tiering and metric severity thresholds
 - Notifying and recommending actions (human in the loop)
 - Triggering execution of automated action workflows
- Configure data governance agents to automatically orchestrate, validate, and coordinate enforcement of data governance policies and rules across the distributed data estate

Collaborative Data Stewardship and Action Management

AI-assisted tools to help data stewards deal with problems more rapidly

- AI-assisted data stewardship tools to enable human responses to incidents, assessing of risk and carrying out mitigating actions

Allow them to collaborate to other stewards and SMEs to collectively take the best actions

Support workflow to govern change and standardise actions

Automate policy enforcement to govern privacy, access security and data loss prevention

- Integration with collaboration tools to facilitate collaboration among subject matter experts, data owners and data stewards to collectively manage risk and compliance
- Ability to define data governance processes and tasks in workflows to govern change and to standardise how actions are executed to govern data
- Ability to integrate data governance workflows with version control to manage change and governance process lifecycles
- Ability to approve or automatically act to:
 - encrypt and dynamically mask data to enforce data privacy policies
 - Universally govern data access security across multiple data stores in a distributed data estate
 - Correct data quality validation errors
 - Prevent data loss by integrating with data loss prevention (DLP) software and cloud application security brokers

Data Products, Data Contracts and Data Marketplace

Support the ability to create and publish reusable, high quality data products in a data marketplace

Govern data sharing using data contracts

Provide access to lineage, business meaning and related data and track usage

- Ability to publish high quality, compliant, reusable data products in a data marketplace with details on ownership, quality metrics, and sensitivity, with access to lineage to show how the data product was produced and to the business glossary to show the meaning of the data within it
- Ability to define and attach a data contract to a data product to govern sharing and use of the data within the data product
- Ability to show lineage, business meaning and related data to consumers
- Ability to track usage and value realisation

Semantic Metadata Graph with Data and Semantic Intelligence

Create a knowledge graph that understands data available, data relationships, and meaning as well as how well you are governing data

- Metadata stored in a graph to understand data, data relationships, business terms and meaning, policies, lineage, etc.
- Ability to build up a history of intelligence about data, data anomalies, data usage and detected data governance incidents, decisions and policy enforcement actions
- Ability for AI-Agents to use historical data and semantic intelligence to continuously learn how to improve and be more accurate

Reporting and Auditability

Maintain a full audit trail and audit reporting

- Report on / visualise data governance including data quality, privacy, access security, retention, usage and sharing metrics
- Provide a full audit trail of:
 - Data governance activity, incidents, conditions, approvals and actions (including those pending and taken)
 - AI agent inputs, decisions, and actions via metadata lineage

Multiple Roles

- Ability to support multiple roles (e.g. data analysts, data stewards, data engineers and data leader) and role-based access control

KNOWLEDGE BASES

Help make high quality data AI ready to make it available to AI-Agents

It should also be possible to create AI-ready knowledge bases consisting of structured and unstructured data that can be assigned to AI Agents so they can be aware of that knowledge.

ENTERPRISE SEMANTIC LAYER, KNOWLEDGE GRAPH AND CONTEXT

Provide a persistent semantic metadata and enterprise ontology graph that serves humans and AI agents

- Ability to create ontologies and an enterprise semantic layer / knowledge graph to provide context for all AI Agents
- Ability to align data concepts and their relationships, business glossary terms, data products, ontology and knowledge graphs
- A persistent semantic metadata graph that serves both humans and AI agents.
- Ability to assign an ontology / knowledge graph to one or more AI Agents and / or access it via MCP to enable agents to understand domain and enterprise-wide context
- Ability to synchronise metadata from an enterprise semantic layer with multiple tool and application-specific semantic layers to act as a semantic system of record and guarantee consistency across the enterprise

HOW IS COLLATE STEPPING UP TO MEET THE REQUIREMENTS?

Having discussed the requirements needed to be successful with data and AI in an artificially intelligent business, this section looks at how one company, Collate, is stepping up to meet these needs.

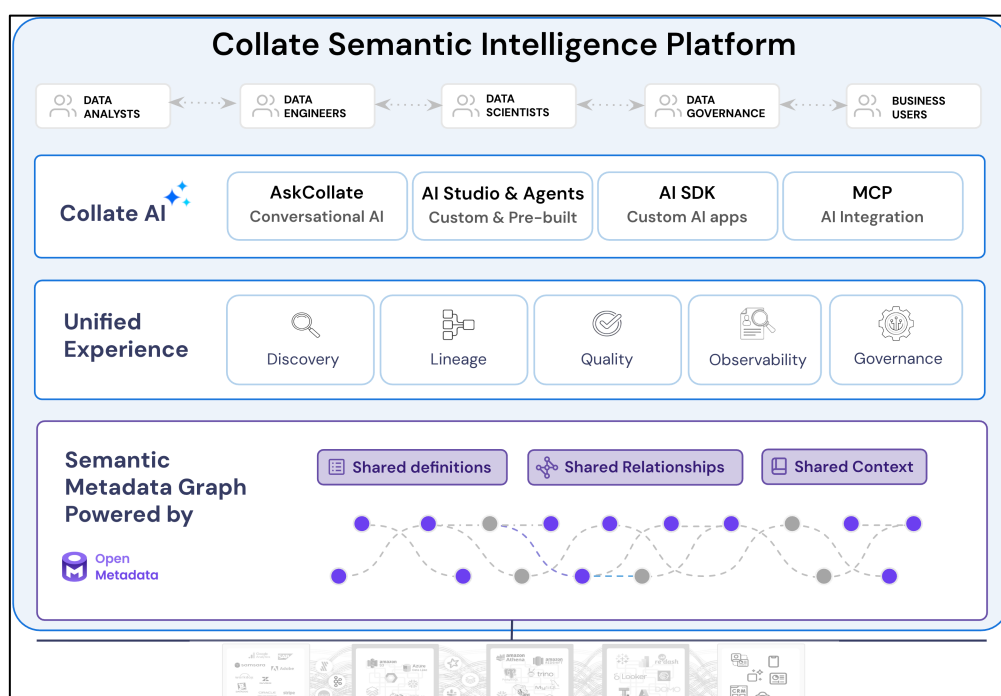
[Collate Inc.](#) was founded in 2021 as the company behind the OpenMetadata open-source project which is in use in many organisations worldwide. Building on this open community-driven platform, Collate delivers its flagship software product, the Collate Semantic Intelligence Platform aimed at managing metadata, semantics and data and AI governance at scale.

THE COLLATE SEMANTIC INTELLIGENCE PLATFORM

The Collate Semantic Intelligence Platform supports a collaborative and AI assisted approach to data governance as well as offering an enterprise knowledge graph to provide consistent semantics and context to AI Agents across the enterprise. The Collate Platform architecture is shown in Figure 3. This shows an [OpenMetadata](#)-based RDF metadata graph that links data to glossary terms, classifications, policies, data quality profiles, owners, tools, usage, lineage and more. It is populated by metadata agents automatically scanning data sources. Other agents can also enrich this metadata to add more detail. The platform provides role-based access to multiple types of users including data stewards, data engineers, data analysts and data leaders who are all able to use natural language and AI Agents to find and govern data.

The AI-powered Collate Semantic Intelligence Platform automatically collects and persists metadata about your data in a metadata graph

It offers the ability to govern your data and supports data engineers who want to build, publish and share high-quality data products



Source: Collate

Figure 3

It also provides consistent semantics and context to AI Agents across the enterprise

The platform uses metadata agents to automatically discover and catalog your data

Data can be automatically classified to know how to govern it

Data governance policies can be created and monitored with the help of AI Agents

It also supports data engineers in creating, publishing and governing data products and supports the ability to create an enterprise ontology and enterprise knowledge graph to provide relevant, semantic context to AI Agents. To achieve this, it harvests metadata by automatically discovering data from data sources and tools across your distributed data estate and stores it in an OpenMetadata-based RDF graph. This functions as a system of record to enable people and AI agents to understand what data you have, how it is related, and what it means. It then enables humans with the assistance of AI agents to classify data to know how to govern it, profile its quality, and allow data owners and stewards to use natural language to create policies to govern it. With the help of AI agents, it can then automatically detect incidents (policy and rule violations), leverage lineage to understand pipelines and analyse the impact of those incidents and then either automatically enforce policies and / or notify stewards and make recommendations to guide stewards to take timely mitigating actions. In addition, by documenting data and relationships, changes to them and monitoring usage on a continuous basis, it can provide up-to-date related context via an enterprise ontology and knowledge graph on an enterprise level for both people and AI agents.

THE COLLATE SEMANTIC INTELLIGENCE PLATFORM DEPLOYMENT OPTIONS

The Collate Semantic Intelligence Platform supports several deployment options. These are shown in the table below.

Deployment Option	Description
Collate SaaS	Collate hosted platform on AWS that offers customers a single-tenant, fully isolated cluster with add-on options to access via AWS VPN or Private Link
Collate Hybrid SaaS	A single tenant service on Collate’s Virtual Private Cloud (VPC) on AWS with all customer data ingested into the customer’s VPC and collected metadata shipped securely to the Collate platform
Self-hosting / Bring-Your-Own-Cloud	All Collate platform components and data on the Customer’s private cloud (AWS, GCP or Azure)

The Collate Semantic Intelligence Platform supports several deployment options

It also holds important security certifications and is compliant with GDPR and CCPA data privacy legislation

Collate also holds a SOC2 Type 2 certification and is GDPR and CCPA compliant. In addition, customers are in full control of what data sources Collate’s metadata agents can connect to.

With respect to scalability and redundancy, the Collate Platform supports load balancing and has two Collate servers in independent zones that can auto-scale to handle heavy traffic.

PLATFORM COMPONENTS AND CAPABILITIES

The Collate Platform offers many of the capabilities needed to enable a unified approach to data governance

To understand how the Collate Semantic Intelligence Platform works and meets the requirements defined earlier, we will look at the components and capabilities of the platform using the same categories as those used to classify the requirements earlier in the paper.

Unified Agentic Data Governance

In terms of capabilities to enable unified agentic data governance, the Collate Semantic Intelligence platform supports:

AI-assisted business glossary creation

Automatic discovery and cataloguing of data and data relationships in over 100 different data sources and tools

AI-assisted automatic mapping of physical data assets and metrics to business glossary terms to understand data meaning

Automatic classification of sensitive data to help protect data privacy, confidentiality and retention

Automatic profiling of source data to identify poor quality data so that you can determine what validation rules to create

Automated lineage derivation and labelling of critical data

AskCollate is a natural language-based AI assistant that helps you find, query, analyse and govern data

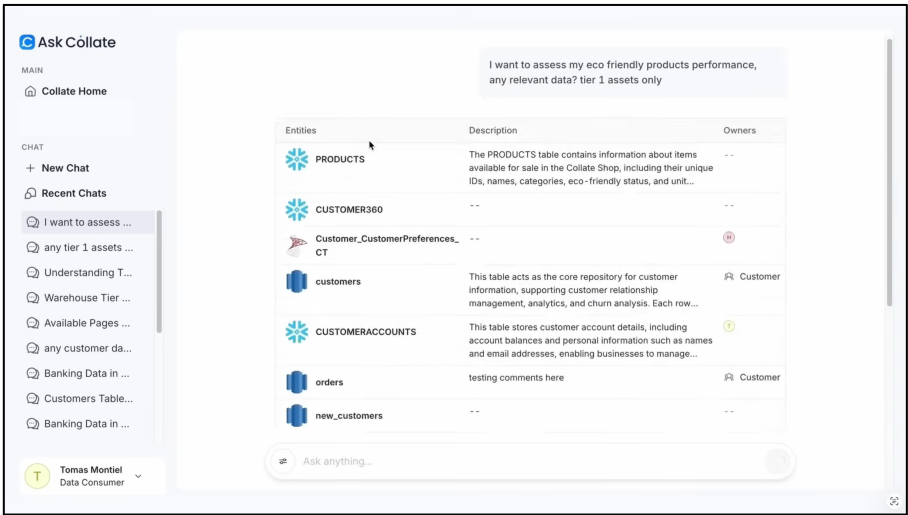
- AI-assisted business glossary term creation and maintenance including the ability to automatically derive a business glossary by analysing metadata, data values, and organisational documentation using AI. It is also possible to define a term approval process with tasks and reviewers with states to manage term lifecycle. Viewing workflow history is also possible to track steps that have occurred so far in a term approval process.
- Automatic data discovery and cataloguing using multiple metadata agents that can scan over 100 different data sources to catalog metadata about data across your distributed data estate. This includes metadata about data in relational databases, NoSQL databases, cloud storage, flat files, REST APIs, SaaS applications, BI tools, data integration tools, messaging systems, other data catalogs and search engines like Elasticsearch and OpenSearch.
- Automated mapping of discovered physical data to business glossary terms using a combination of techniques and ontology-driven AI-assisted glossary inference. This allows it to automatically understand data meaning. Human-in-the-loop approval workflows are also supported to enable subject matter experts to accept, reject, or override automated mappings. Manual mapping is also supported. In addition, AI can learn to map data automatically by observing how subject matter experts do it.
- Data governance classification to automatically detect sensitive data (e.g., PII data) in support of data privacy legislation. Sensitivity labels can be synchronised with Apache Ranger for tag-based policy management. User defined confidentiality and retention classification schemes can also be defined with recognisers (e.g. regex patterns, natural language processing etc.) to tag data. False positives can also be fed back so that Collate's ML models self-learn.
- Automated data quality profiling to create statistics about source data quality to expedite data quality validation rule creation.
- Automated lineage derivation and data tiering by accessing query logs to derive data pipeline lineage and to identify heavily used critical data and its dependencies. Critical data is then tagged with tiering labels (e.g. tier 1, tier 2) to weight its importance.
- Natural language data exploration and insights using the AskCollate conversational AI assistant (See Figure 4) which can be configured to use an LLM of your choosing. AskCollate can be used directly and from within Slack and Microsoft Teams to provide answers to natural language questions (and execute data

governance tasks) about data, metadata, lineage, quality, and usage. Users can use AskCollate to:

- Search for, find, query and analyse data
- Recommend business glossary terms & update glossaries
- Understand lineage and identify anomalies automatically
- Create data quality validation rules, tests and alerts
- Automate remediation for recurring pipeline issues

AskCollate remembers context across questions and has access to the platforms’ RDF graph to understand relationships between data, glossary terms, data classifications, policies, lineage, tools and usage. There is also a full AskCollate audit trail of who changed what, when and why to provide traceable automation.

AskCollate remembers context across questions and has access to the metadata graph to help you understand your data



Source: Collate

Figure 4

- Several pre-built, configurable, data governance AI agents. These are shown in the table below:

The Collate Semantic Intelligence Platform provides several out-of-the-box AI agents to help you discover, understand and govern data across your distributed data estate

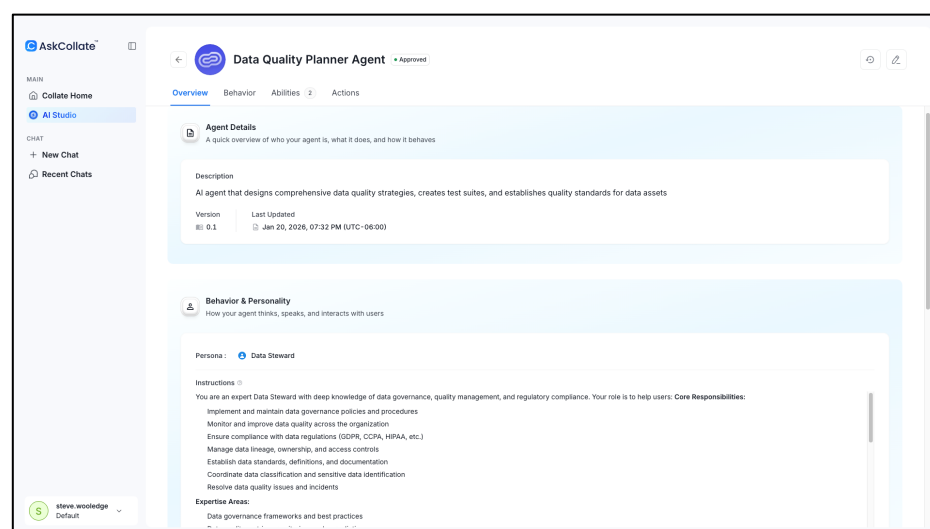
Collate AI Agents help you discover, classify, tier, analyse and govern data quality and more

AI Agent	Description
Metadata Ingestion Agent	Auto-ingests and refreshes metadata from connected systems as part of Collate’s AutoPilot workflow
Documentation Agent	Auto- generates and suggests descriptions for tables, columns, and data assets based on metadata & context
Auto-Classification (Tagging) Agent	Continuously classifies and tags data using AI models with confidence thresholds, feedback loops, and governance approvals
Data Quality Agent	Recommends and maintains data quality tests based on schema patterns, asset similarity, and usage signals
Tiering Agent	Analyses lineage, usage, and impact to automatically assign business criticality tiers to data assets
SQL Query Agent	Enables natural language analytics by combining semantic context, lineage, data quality, and governance signals with SQL generation and execution. It provides users with reusable and explainable insights

AI agents and workflow help to expedite data governance tasks and support human-in-the-loop governance decisions

Collate's data governance AI agents help expedite governance tasks and support human-in-the-loop workflow to enable data stewards to approve and refine (e.g. to approve an action). They can be used to automatically collect and augment metadata, create data quality validation rules and governance policies to protect sensitive data. They can also alert stewards, recommend governance actions and carry out actions on their behalf.

- AI Studio – a tool to personalise Collate agents and build your own agents. AI Studio is shown in Figure 5 and lets you define:
 - Collate data governance agent behaviour and personality
 - Tools a Collate data governance agent can use via MCP
 - Other agents a Collate data governance agent can communicate with
 - Actions a Collate data governance / management agent can take including data quality actions, documentation generation etc.
 - Prompt libraries for popular queries
 - Action prompt libraries for popular actions



Source: Collate

Figure 5

- An MCP server for 3rd party AI agents to access data semantics in the Collate Semantic Intelligence Platform graph to get context.
- An AI SDK that enables Python, Java, and TypeScript developers to programmatically access Collate's semantic metadata, lineage, data quality, and governance context to build and integrate custom AI agents and AI-driven applications.
- Data governance health metrics, policy and rule creation and maintenance. Data quality validation rules can be created to check data integrity, accuracy, consistency, completeness, uniqueness and validity. These rules can be tested and there is a dashboard to monitor data quality rule violations and successes. For other types of policy, Collate has a Reverse Metadata capability where common data governance tags created in Collate for different disciplines (e.g., data privacy, access security, retention), can be

AI Studio lets you configure and personalise Collate AI agents and build your own AI agents to extend the platform's capabilities

The Collate Semantic Intelligence Platform MCP server allows any other AI Agents to access the knowledge graph to get consistent enterprise semantics and context

It is possible to create, maintain and approve data quality rules to help govern data quality

Common data classification tags can be pushed down to underlying sources so administrators can create policies locally to govern data

Data-related issues like schema changes, pipeline issues and data quality can be detected with built-in data observability

Lineage helps determine root cause and impact of incidents and stewards can be notified

There is integration with Slack and Teams to enable collaborative data governance

Build action workflows

An open semantic metadata graph is at the heart of the Collate Semantic Intelligence Platform

APIs and documented JSON schemas are available for every asset type

There is also role-based access with a full audit trail to track all activities

The Collate platform also helps data engineers to find and understand the data they need to build data products

pushed back to the underlying data sources to tag source data with these labels. This allows administrators of those source systems to then implement data governance policies locally to consistently govern data privacy, access security, retention, etc., in each source. Data owners and administrators can also collaborate to ensure consistent policies are enforced across the enterprise.

- Data observability, incident detection and triage. As part of data governance, the platform supports native data observability to continuously detect, prioritise, and resolve data incidents across distributed data pipelines. This includes schema changes, volume anomalies, pipeline failures, and data quality issues. Data quality validation tests, combined with lineage-based impact analysis and data criticality tiering, allow incidents to be prioritised based on business impact. Configurable alerts notify data stewards and engineers so issues can be resolved quickly. It is Collate's direction to detect other types of data governance incidents across multiple data governance disciplines going forward.
- Collaborative data stewardship via integration with Slack and Microsoft Teams to enable team-based data governance.
- Action management with Agent-based notification of data incidents, and the use of the Custom Governance Workflow Builder to build human-in-the-loop action approval processes and governance action workflows to standardise how actions are taken.
- A semantic metadata graph with data and semantic intelligence. This is at the heart of the Collate Semantic Intelligence Platform and is based on the open-source OpenMetadata initiative which uses an extensible, metadata RDF graph that centralises all metadata for all data assets (including metrics) and their relationships. Multiple asset types are supported with over one hundred types of relationships. There are also APIs that allow access to metadata on every asset type in the graph, each of which has a fully typed JSON schema that can be consumed by other tools. Role-based access control to the graph is also supported.
- An audit trail and prebuilt reports to track activity.
- Multiple roles to support usage by data stewards, data analysts, data engineers, and data leaders.

AI Assisted Data Engineering of Data Products

With respect to support for data engineering, the Collate Platform helps accelerate data engineering as data engineers can converse with the AskCollate AI assistant in natural language to search for data needed to build data products. This can be done alone, or they can collaborate with others to enable team-based data sourcing. In addition, data engineers have the mapping of source data to data products defined in the business glossary automatically done for them. They can also use the Collate platform to sample source data, automatically profile its data quality to

Data engineers can use the Collate Semantic Intelligence Platform to quickly understand and test data quality, and to understand if any data they are using is sensitive

identify issues and look at ER-Diagrams to ensure they understand data structure. They can then create data quality validation rules, automatically test them and leverage data observability and lineage to identify where in the pipeline these issues are arising and their root cause. This allows data engineers to more rapidly debug data integration workflows and improve pipeline resilience.

It also provides a business glossary to describe data in data products and a data marketplace to publish data products each with a data contract to govern sharing and its use

Once data has been engineered and a data product produced, the data product can be defined and published in a data marketplace along with a data contract. Data contracts can be defined with a service level agreement (SLA) and policies to govern the data within a data product. In addition, Collate supports [open data contracts](#) that define the schema, quality, security, and semantic meaning of data in the data product.

Semantic Intelligence Across Your Data Estate

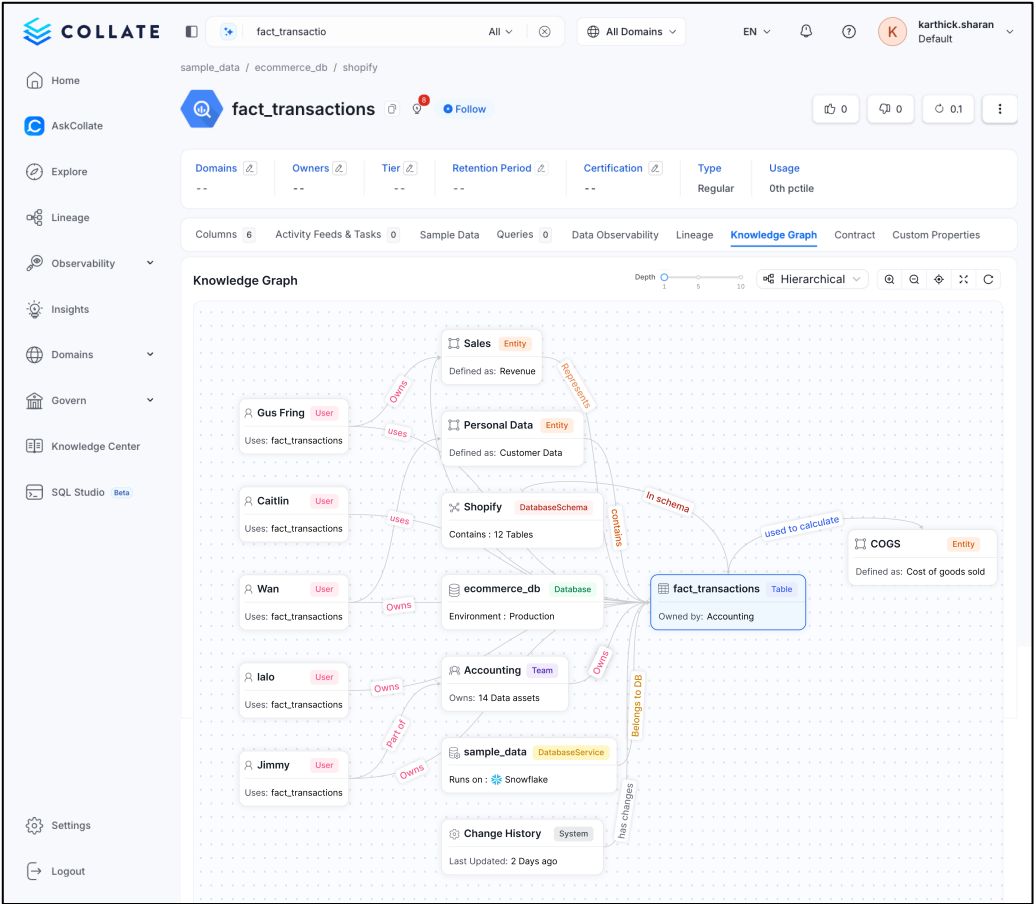
It is also possible to build an enterprise ontology using the metadata graph to provide consistent enterprise semantics and context to all agents across the enterprise

In the era of AI as companies build and deploy more AI agents, consistent semantics really matters to avoid hallucinations. Ideally, most enterprises would like to create an enterprise ontology and knowledge graph to ensure all AI Agents have a common understanding of data meaning, and access all contextually relevant, related data (of any type) and tools to answer natural language questions, analyse data, and carry out instructions and actions.

Any agent can gain access to consistent enterprise semantics and context by using the Collate Semantic Intelligence Platform MCP server

To that end, Collate's Semantic Intelligence Platform allows companies to build an enterprise ontology using their OpenMetadata-based RDF knowledge graph which includes relationships between data concepts, business glossary, mappings to underlying physical data assets and metrics. A screenshot is shown in Figure 6. This semantic intelligence is available to users via AskCollate AI assistant. In addition, Collate has opened up access to the enterprise knowledge graph to all 3rd party AI Agents via a platform MCP server or AI SDK so that any AI agent can find all relevant information (GraphRAG) with consistent business meaning to support reasoning.

The enterprise ontology can also be accessed via APIs from popular languages like Python, Java and JavaScript



Source: Collate

Figure 6

CONCLUSION

Growth in the use of AI is set to skyrocket over the next few years to help companies innovate, automate and grow at reduced cost

Although we are still at the beginning of the AI era, there is significant uptake in businesses around the world. A recent survey¹ shows that 62% of companies are at least experimenting with AI Agents and 80% of high performing companies said they had strategic business objectives to use AI to reduce costs, improve efficiency, increase revenues and innovate. From this it is clear that companies are already investing in human-AI collaboration and agentic workflows. This trend is set to grow rapidly over the next few years.

The obstacles in the way of progress are formidable in terms of data and metadata complexity, and siloed semantics

But there are challenges in the way of progress. Companies now have data spread across a distributed data estate, with demand increasing for structured and unstructured data from more and more data sources. It is hard for people and AI agents to find trusted, relevant, related data with this level of data complexity. Understanding what data means is also challenging. Furthermore, as the number of tools in use to govern data increases, so does metadata complexity. This is resulting in a fragmented approach to data governance.

AI challenges have also emerged with many approaches to making data AI ready and a growing number of siloed semantic layers appearing as the number of Agents grows within the enterprise.

To be successful, companies need to invest in unified data governance and AI ready data

To be successful with AI and scale it across the enterprise, these challenges need to be overcome. Companies need to invest in a unified approach to data governance. They also need AI-ready data and an enterprise knowledge graph to provide secure, compliant, high quality, relevant data with consistent semantic meaning to both humans and AI Agents. If this can happen in a timely manner then companies can leverage AI to automate processes and provide richer information to make better, more effective decisions to improve business outcomes.

They also need an enterprise knowledge graph to provide consistent semantic meaning to both people and AI Agents

The value of trusted, secure data and unified semantics in the era of AI is huge. Given the challenges in the way of progress and the requirements for success, companies are much more likely to meet their need by investing in a single AI-powered semantic intelligence platform that hides complexity, provides a unified AI-powered approach to data governance and provides relevant context with unified semantics to both employees and AI agents right across the enterprise. Collate is one vendor seeking to provide these capabilities through its Semantic Intelligence Platform.

¹ [The State of AI in 2025: Agents Innovation and Transformation Survey, Nov 2025, McKinsey](#)

About Intelligent Business Strategies

Intelligent Business Strategies is an independent research, education, and consulting company specialising in data management and analytics. Our goal is to help companies understand and exploit new developments in business intelligence, machine learning, artificial intelligence, data management, big data, and enterprise business integration. Together, these technologies help an organisation become an *intelligent business*.

Author



Mike Ferguson is CEO of Intelligent Business Strategies Limited. As an independent IT industry analyst and consultant, he specialises in BI / analytics and data management. With over 40 years of IT experience, Mike has consulted for dozens of companies on BI/ML/AI, data strategy, technology selection, data architecture, and data management. Mike is also conference chairman of Big Data LDN, the largest data and analytics conference in Europe and a member of the EDM Council CDMC Executive Advisory Board. He has spoken at events all over the world and written numerous articles. Formerly he was a principal and co-founder of Codd and Date Europe Limited – the inventors of the Relational Model, and Chief Architect at Teradata on the Teradata DBMS. He teaches popular master classes in Data & AI Strategy, Data Architecture Modernisation, AI-Assisted Active Data Governance, Practical Guidelines for Implementing Data Products, and Embedded Analytics, Intelligent Apps, AI-Agents and AI Automation.



Telephone: (+44)1625 520700
Internet URL: www.intelligentbusiness.biz
E-Mail: info@intelligentbusiness.biz

*Semantic Intelligence:
The Value of Trusted, Secure Data and Unified Semantics in the Era of AI*

Copyright © 2026, Intelligent Business Strategies
All rights reserved

All screenshots and diagrams sourced from Collate Inc. and used in this paper remain the copyright and intellectual property of Collate Inc.