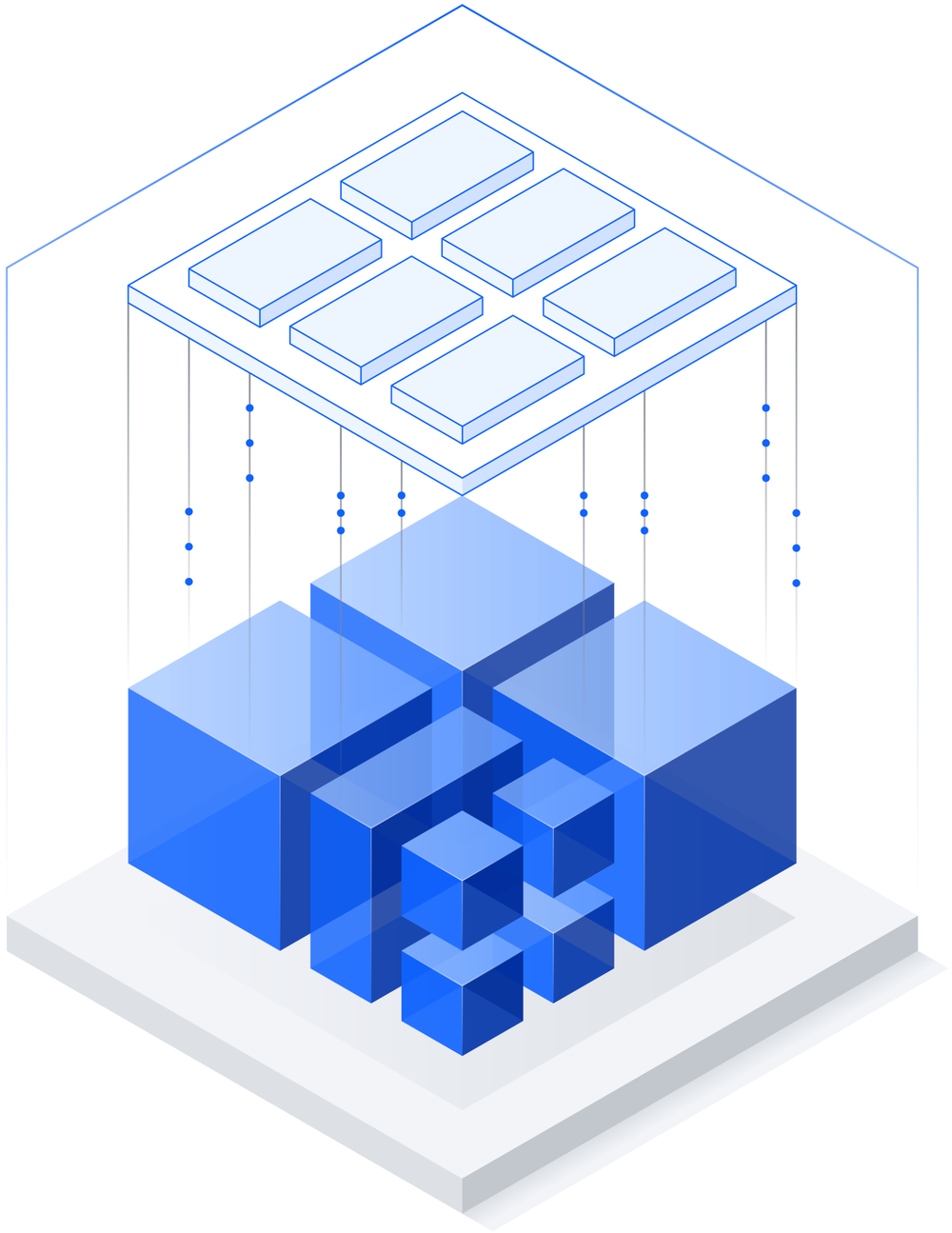
Data Intelligence

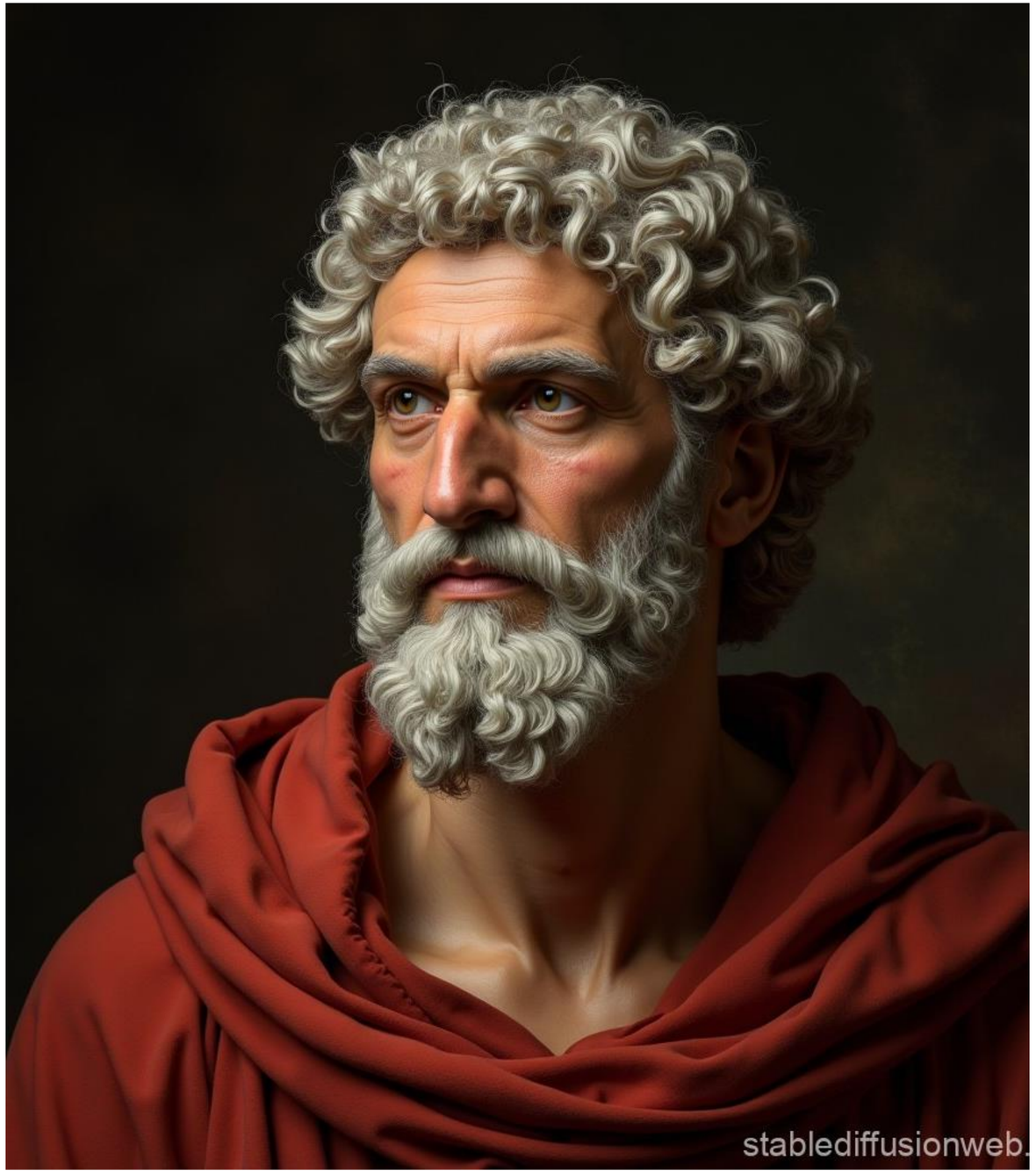
Transforming Data into Business Value



Hizkia Febianto
Client Engineering – Data and AI
hizkia.febianto@ibm.com

IBM

Knowing yourself is the beginning of all wisdom.
-Aristotle



Market Understanding and Trends

AI+

Data for AI & AI for Data

Enterprises are shifting mindset from “adding AI” to “starting with AI”

- Accelerated decision making
- Renewed interest in unstructured data for RAG tuning
- Employing AI to improve data curation and interaction

Modernization of governance practices

Decentralization of governance to support diverse, distributed hybrid cloud environments

- Modernization to cloud
- Mixed new & old technology stacks and formats
- Integrate data & AI using active metadata
- Convergence of integration with low-code tools

New data consumers

New applications of data and new data consumers that require controlled, managed access

- Expanding definition of data products
- Bidirectional contracts for data products (e.g. quality, authorized use, availability SLAs)
- Foster data asset reuse to limit data sprawl

Enforced regulations

Consumer protections for privacy, AI, quality & transparency enforced with hefty fines for non-compliance

- COVID enforcement pauses have resumed
- Regulators rushing to define AI transparency requirements
- Executives want consistent enforcement of policies across data estate

Top data barriers to AI implementation

35%

Data accessibility challenges

30%

Data volume and/or complexity

26%

Data scope or quality problems

80%

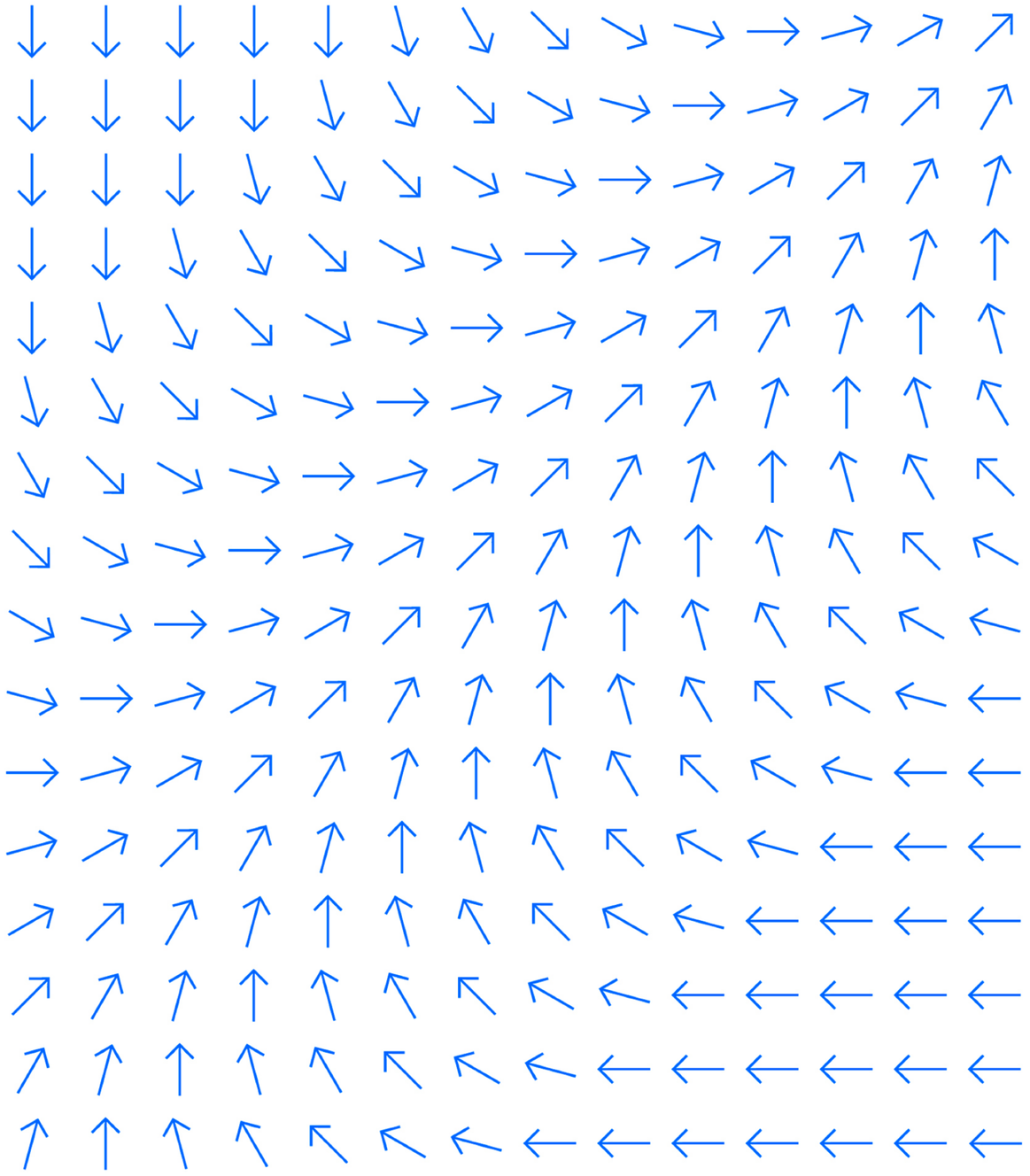
New data is unstructured

Source: Gartner, Cool Vendors in Data-Centric AI, Svetlana Sicular, Arun Chandrasekaran, Sumit Agarwal, Oleksandr Matvitskyy, Anthony Mullen, 14 December 2022; MIT Sloan, Tapping the power of unstructured data, Tam Harbert, Feb 2021

How do you solve for this?

“The foundation of enterprise intelligence is data intelligence. Intelligence about data supports and informs every data-driven learning, analytics, decision, action and outcome.”

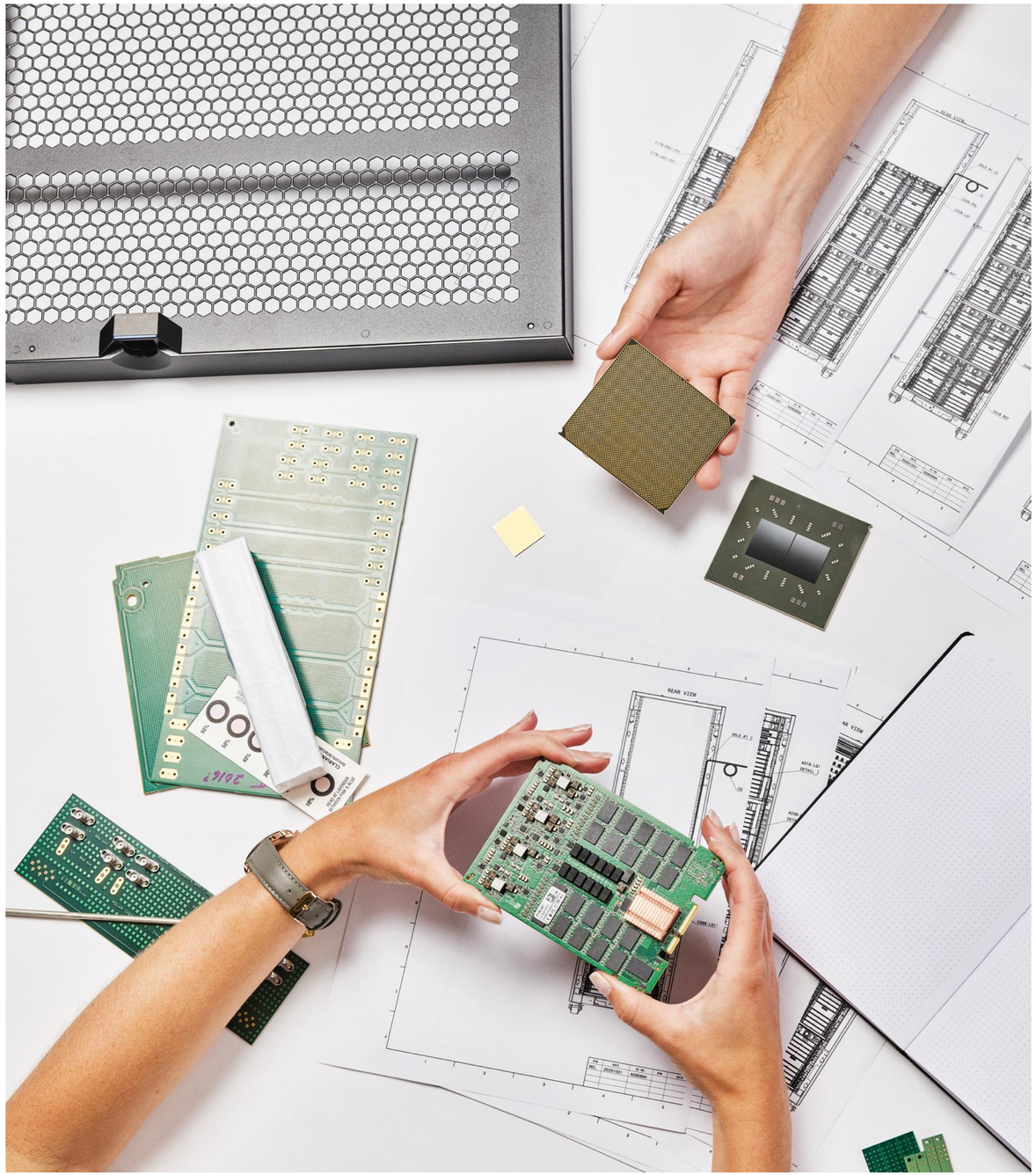
Stewart Bond
Vice President
Data Intelligence and Integration Software, IDC



Data Intelligence in the Future of Intelligence, IDC (Stewart Bond), March 2023.

Transforming data into business value

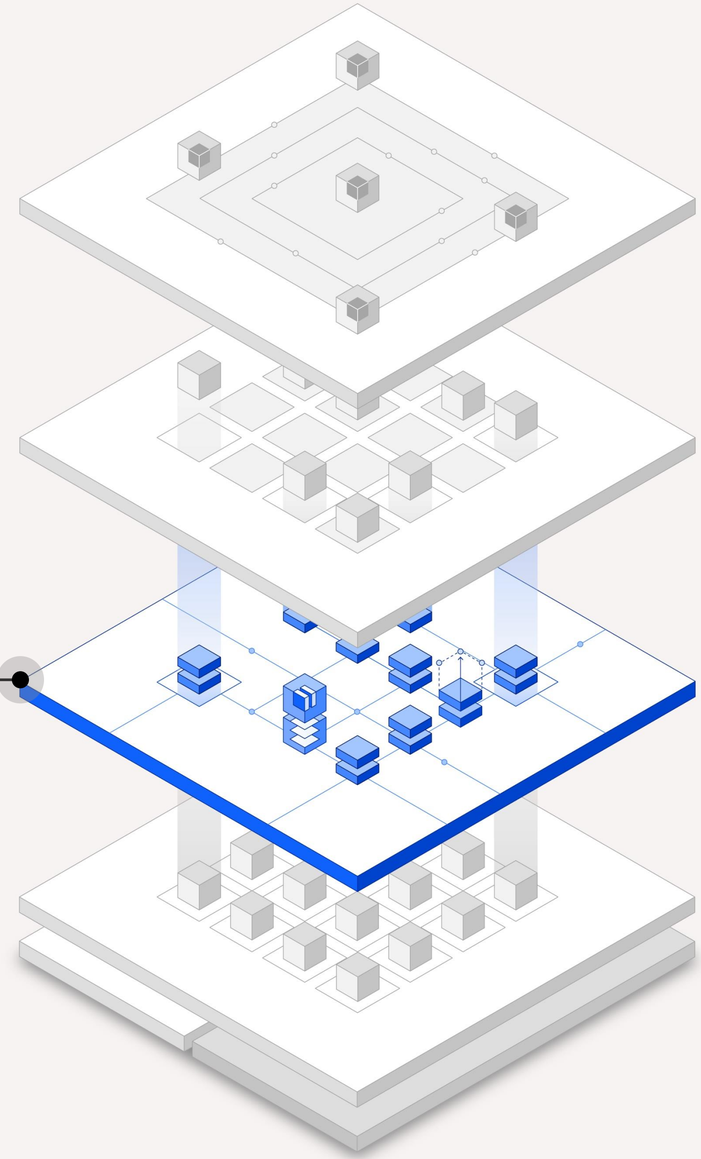
IBM Data Intelligence enhances productivity for data professionals and streamlines the process of discovering, curating and sharing trusted data assets across the organization, accelerating time to value.



Data cataloging and governance

The world’s most scalable data curation is powered by trusted LLMs from IBM Research and prebuilt accelerators to fast-track governance and compliance programs.

70% reduction in manual labor efforts to run a data governance and compliance program*



IBM Knowledge Catalog

A centralized enterprise metadata repository that automates the curation of data assets at scale, enabling data accessibility and usability across organizations

**Examples presented are illustrative only. Actual results will vary based on client configurations and conditions; therefore, generally expected results cannot be provided.*

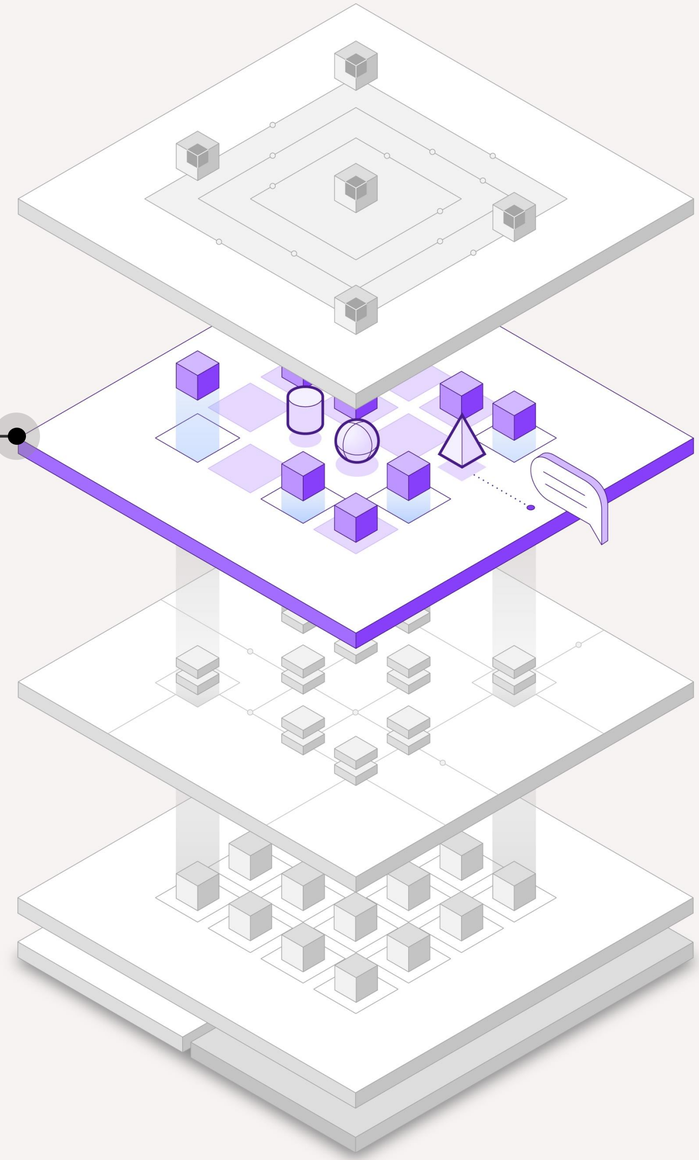
Data lineage

IBM Manta Data Lineage is the preferred lineage solution of the data intelligence market.

95% reduction in time spent debugging root cause analysis in streaming data pipelines*

IBM Manta Data Lineage

Enables organizations to track and understand the origin, movement and transformation of their data, reducing errors and improving accuracy



**Examples presented are illustrative only. Actual results will vary based on client configurations and conditions; therefore, generally expected results cannot be provided.*

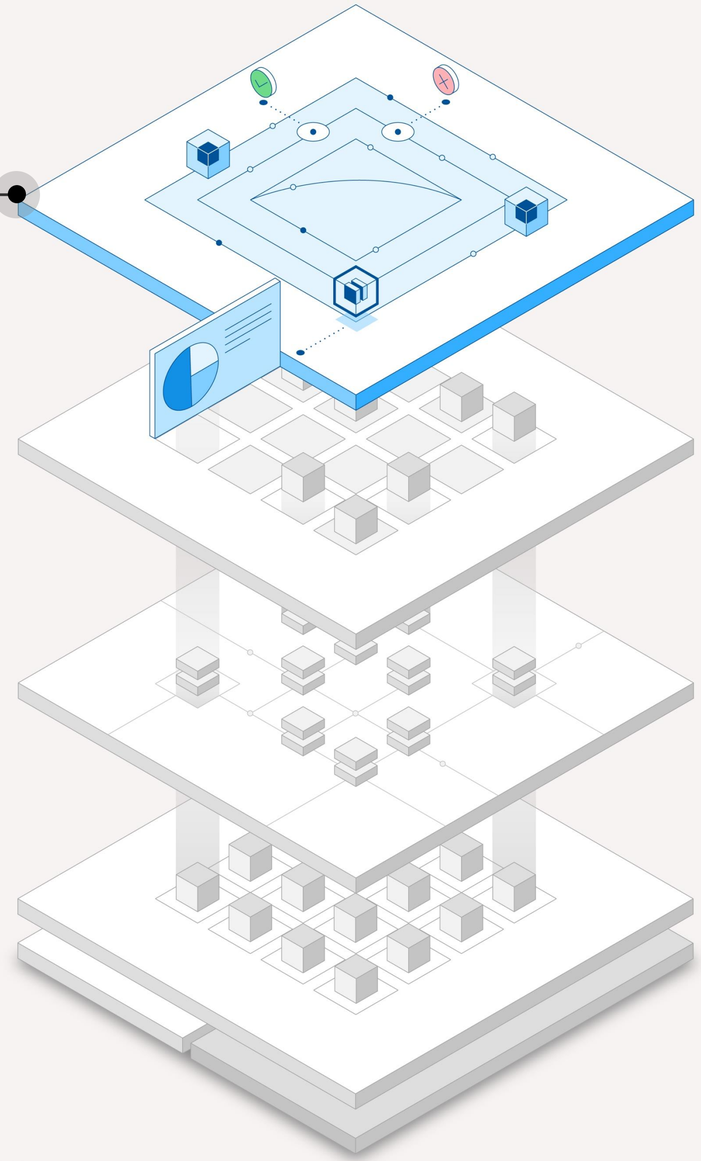
Democratizing data and data products

IBM is the only vendor to support data product sharing across a heterogenous data estate.

90% reduction in time spent implementing data in new use cases*

IBM Data Product Hub

A centralized marketplace for data products, which are collections of curated data or related assets packaged for distribution and consumption

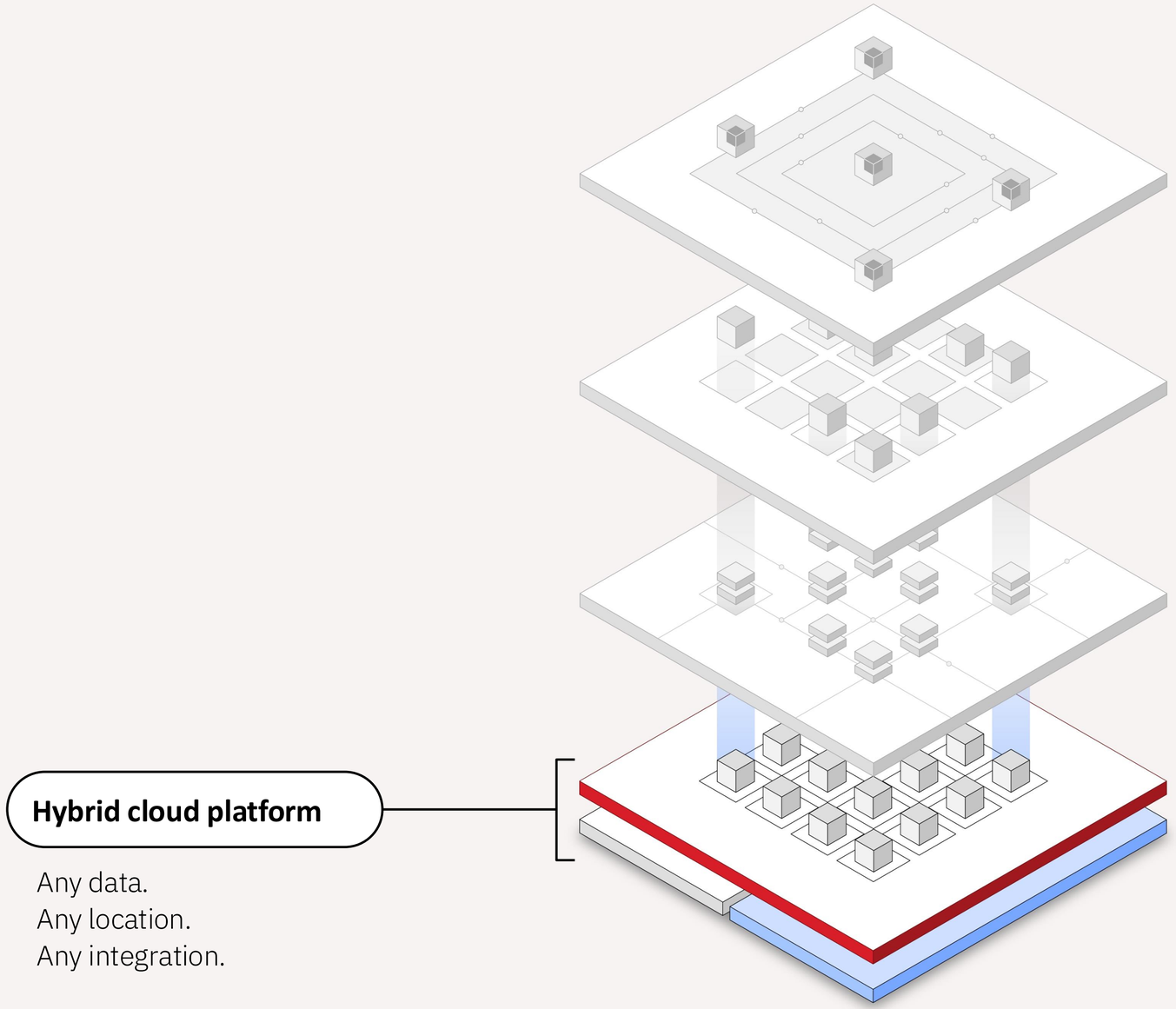


**Examples presented are illustrative only. Actual results will vary based on client configurations and conditions; therefore, generally expected results cannot be provided.*

Hybrid by design

The world’s most scalable data curation is powered by trusted LLMs from IBM Research and prebuilt accelerators to fast-track governance and compliance programs.

Hybrid by design can deliver upwards of 3.3x the return, as compared to a siloed infrastructure approach*



*Examples presented are illustrative only. Actual results will vary based on client configurations and conditions; therefore, generally expected results cannot be provided.

How can IBM help?

IBM offers data governance, data lineage, data quality and data sharing capabilities as part of an extensible data and AI platform

IBM Knowledge Catalog

IBM Manta Data Lineage

IBM Data Product Hub

Data stewards have seen a productivity increase through leveraging data intelligence solutions of up to

110%

Data engineers decrease time spent on optimizing data processes by as much as

90%

Data consumers leverage self-service data solutions to decrease time spent understanding, finding and accessing data by an estimate of

95%

Examples presented are illustrative only. Actual results will vary based on client configurations and conditions; therefore, generally expected results cannot be provided.

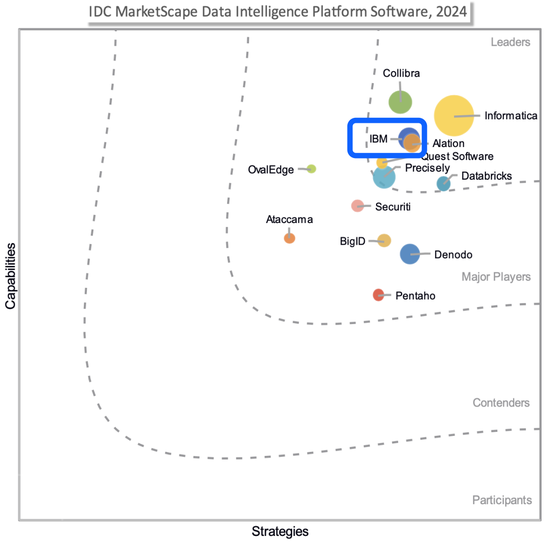
And analysts agree: IBM is a leader in multiple categories

2024 Gartner Magic Quadrant for Data Analytics and Governance Platforms[5]

Figure 1: Magic Quadrant for Data and Analytics Governance Platforms

Gartner

IDC MarketScape: Worldwide Data Intelligence Platform Software 2024[6]



ISG Data Intelligence Buyers Guide[7]

Data Intelligence
Product Experience

Providers	Grade	Performance
Informatica	A-	Leader 69.6%
Oracle	A-	Leader 66.0%
IBM	A-	Leader 65.3%
Microsoft	B++	64.2%
SAP	B++	63.8%
AWS	B++	62.4%
Databricks	B++	60.8%
Google Cloud	B++	60.5%
Qlik	B+	55.7%
Alation	B+	55.6%
Huawei Cloud	B	54.7%
Solace	B	52.3%
Alibaba Cloud	B	52.2%
Collibra	B	51.7%
Cloud Software Group	B-	48.8%
SAS Institute	B-	48.2%
Boomi	B-	46.7%
Reltio	B-	45.8%
Precisely	B-	45.6%
Syniti	C++	43.5%
Ataccama	C++	42.3%
Quest Software	C++	41.5%
Rocket Software	C++	41.1%

ISG
Source: ISG Research
Data Intelligence Buyers Guide
© 2024 Information Services Group, Inc.

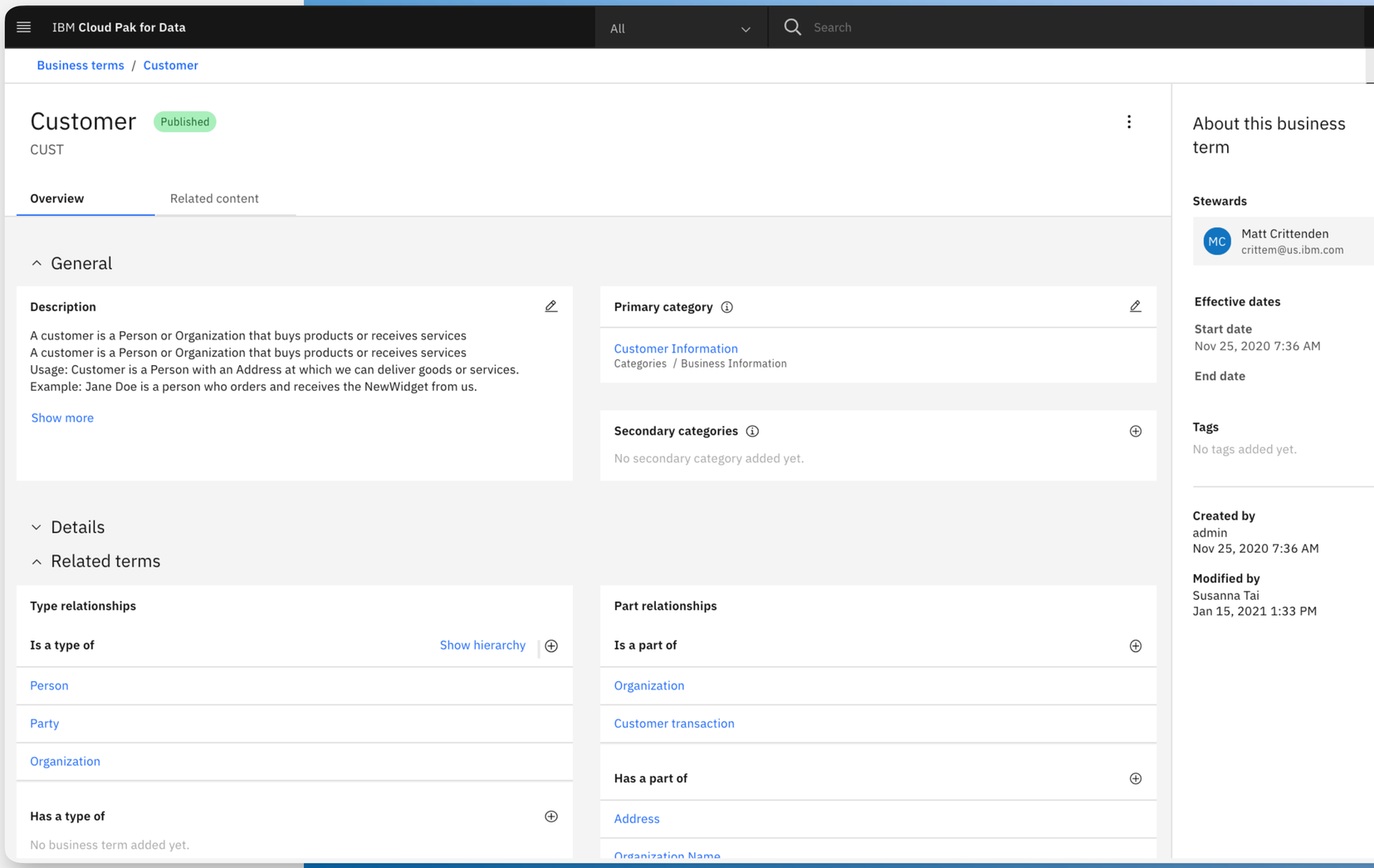
5. 2024 Gartner Magic Quadrant for Data Analytics and Governance Platforms, Gartner, 2024.
6. IDC MarketScape: Worldwide Data Intelligence Platform Software 2024 Vendor Assessment, IDC, November 2024.
7. ISG Data Intelligence Buyers Guide, ISG, 2024.

IBM Knowledge Catalog Key Features

Business Glossary

Govern data through polices, rules business terms, data classes, reference data, and classifications to standardize definitions of business concepts, and to enrich and govern your data

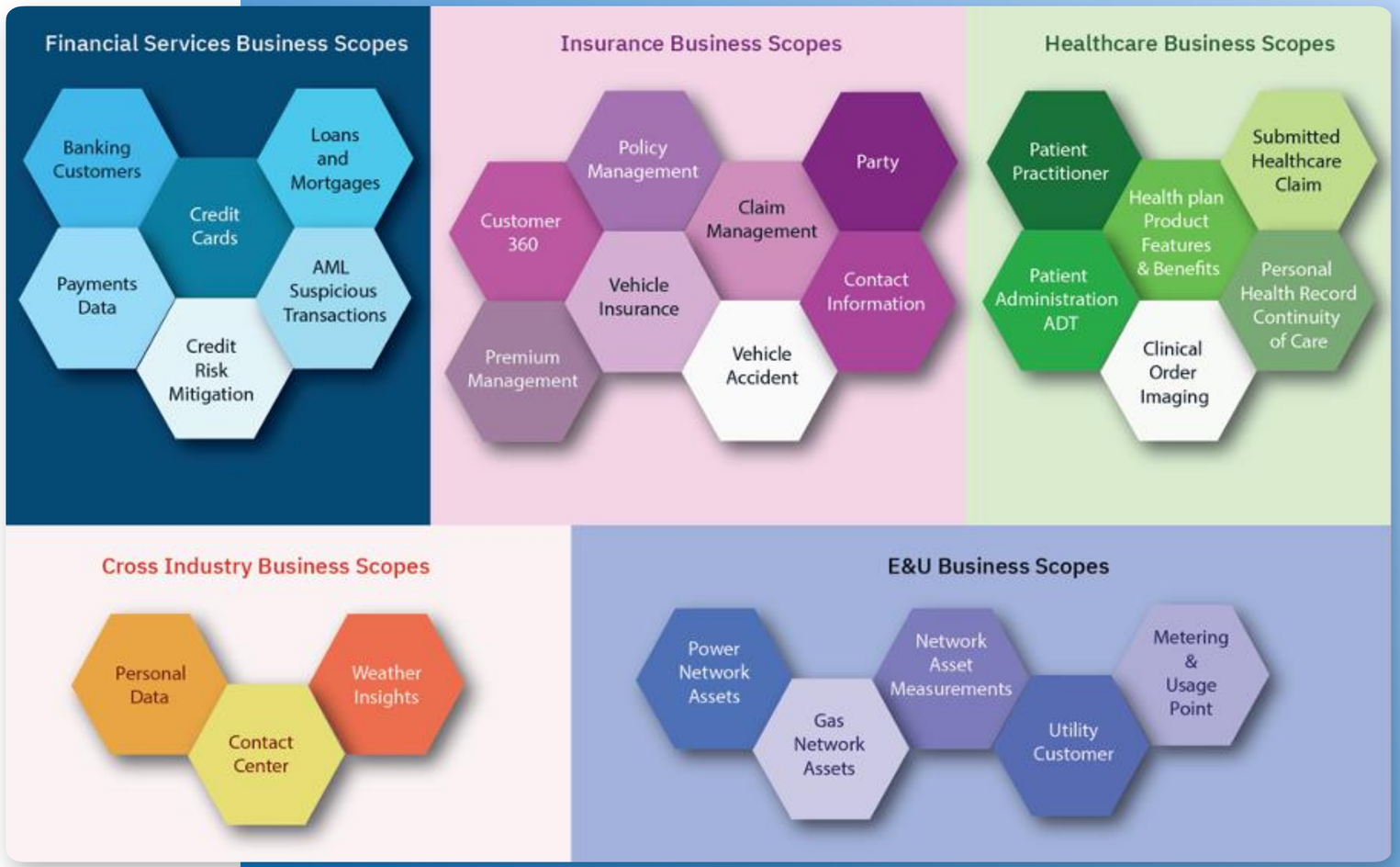
- Extend metamodel with custom properties and relationships
- Promotes consistent terminology and understanding
- Makes it easier to find and understand data



Knowledge Accelerators

Pre-built, industry-specific glossary assets designed to help enterprises quickly implement specific business outcomes

- Set of business scopes for each supported industry that are smaller and more manageable in size
- Align concepts from industry regulations and standards with your business data
- Accelerate regulatory compliance



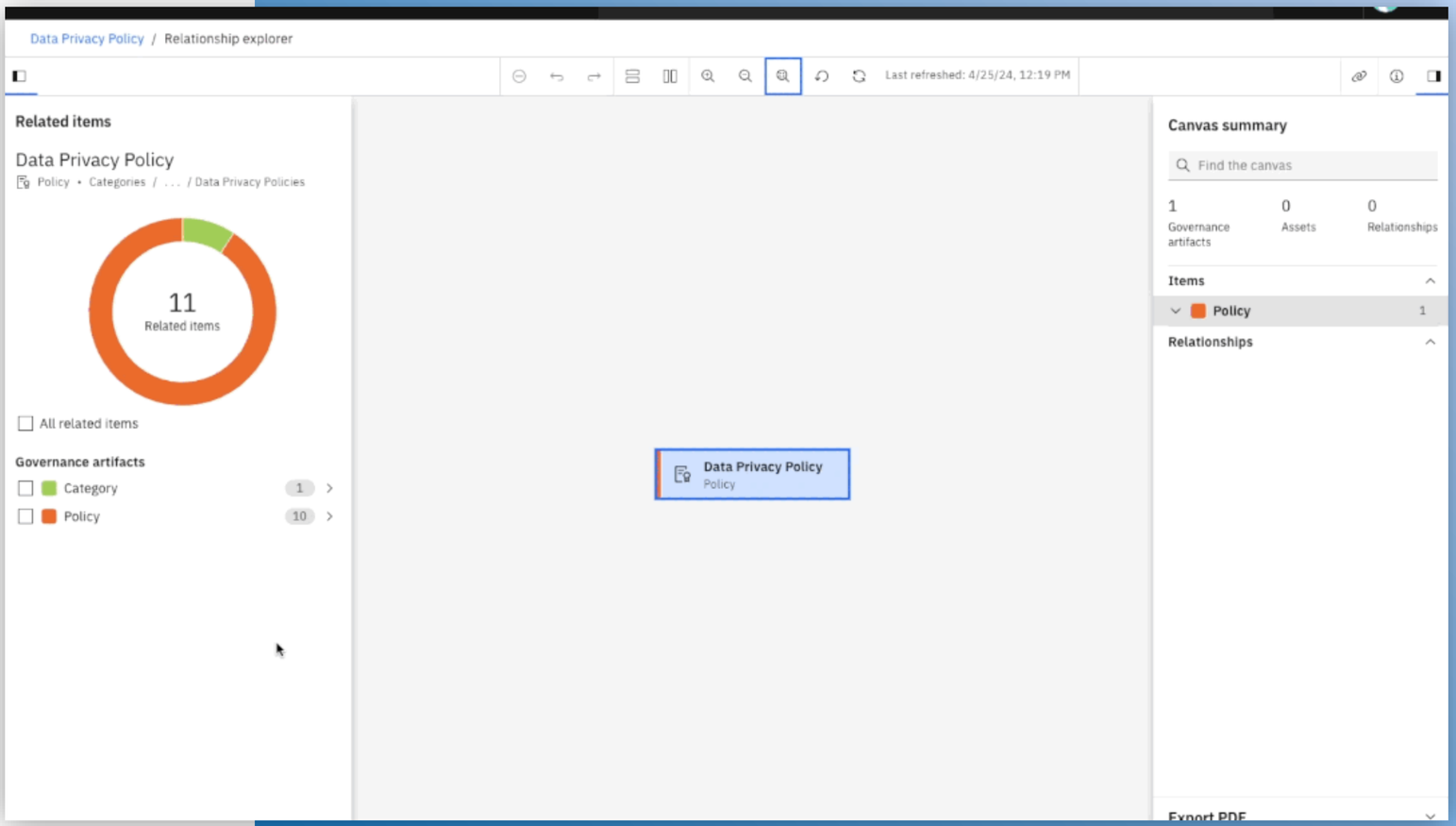
Relationship Explorer

New in June 2024! (animated)

Explore your data landscape and relationships through visualization

Relationship explorer is a visualization tool of the knowledge graph relationships. It enables users to graphically navigate between business and technical assets through their relationships.

- Data consumers can find relevant data assets and their relationships with business concepts and other data assets. They can understand the impact and importance of data governance.
- Data stewards can explore asset relationships for impact analysis, audit metadata and investigate data governance.



"When you're crafting a business term and putting it together, at the end of the day, it's not in isolation. It's there for a reason. And I'm seeing "the reason". I'm loving what you guys are showing."

How it's Unique

- Export to PDF: Share outside of IKC.
- Save Graph: Share with other IKC users.
- Refresh Graph View: Keep the graph view in sync with backend changes.
- Find Canvas: Easily locate relevant assets and metadata.
- Highlight Nodes and Relationships: Immediately see their location on the graph view. Summary of Related Items: Easily add or remove nodes individually or in bulk to the graph view.

Automated Metadata Discovery and Enrichment

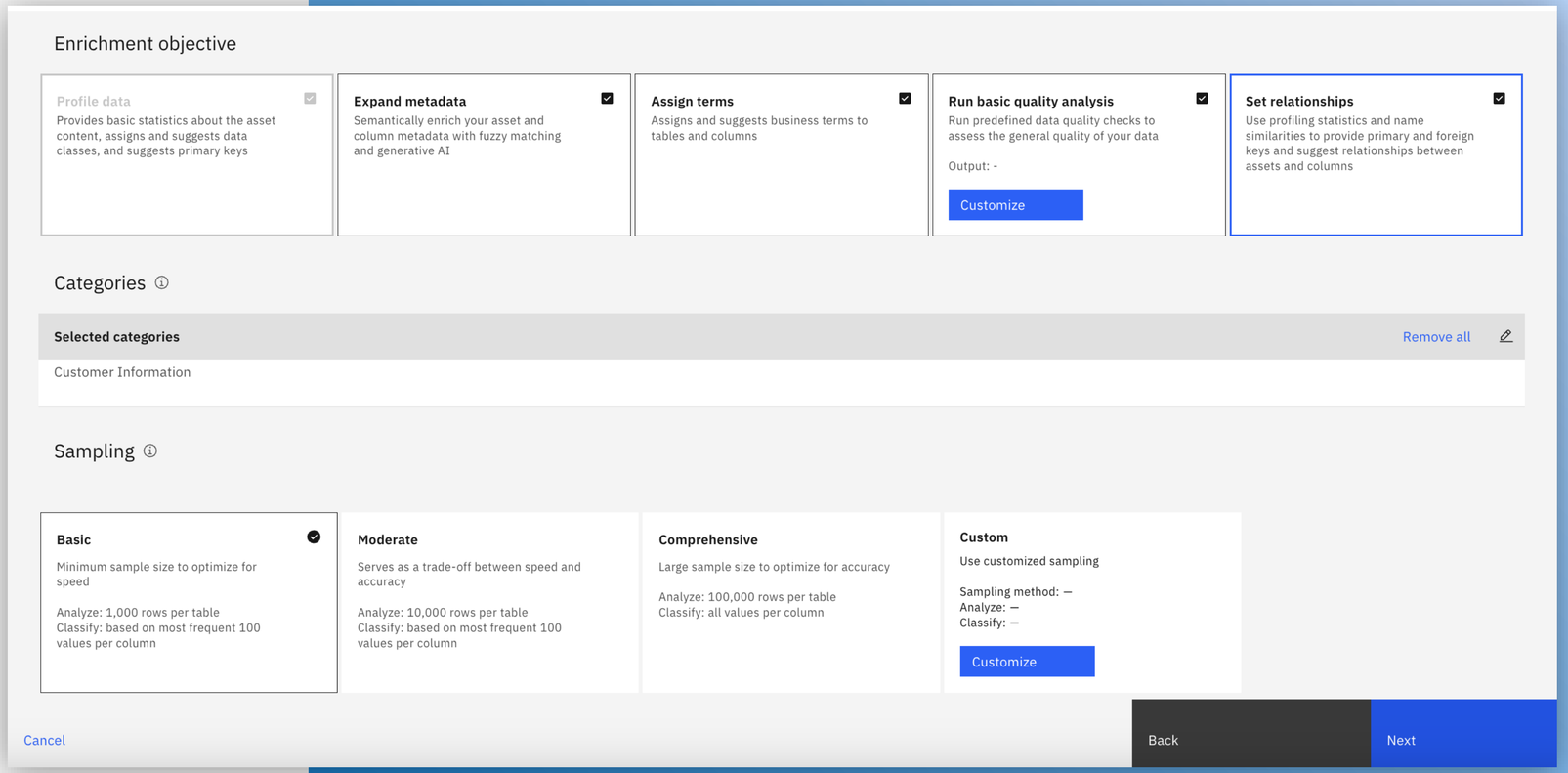
Kickstart governance and scale business understanding with ML-powered ontology mappings when onboarding assets

Automated enrichment includes:

- Statistical analysis through profiling of structured data
- Data patterns detection (data class assignments)
- Data quality checks and scores
- Relationship analysis
- ML-based business term assignments

Automate data ingestion to detect changes over time and ensure currency of assets

Reduce bias through random or other sampling options



Metadata Enrichment, now with LLMs

New in June 2024!

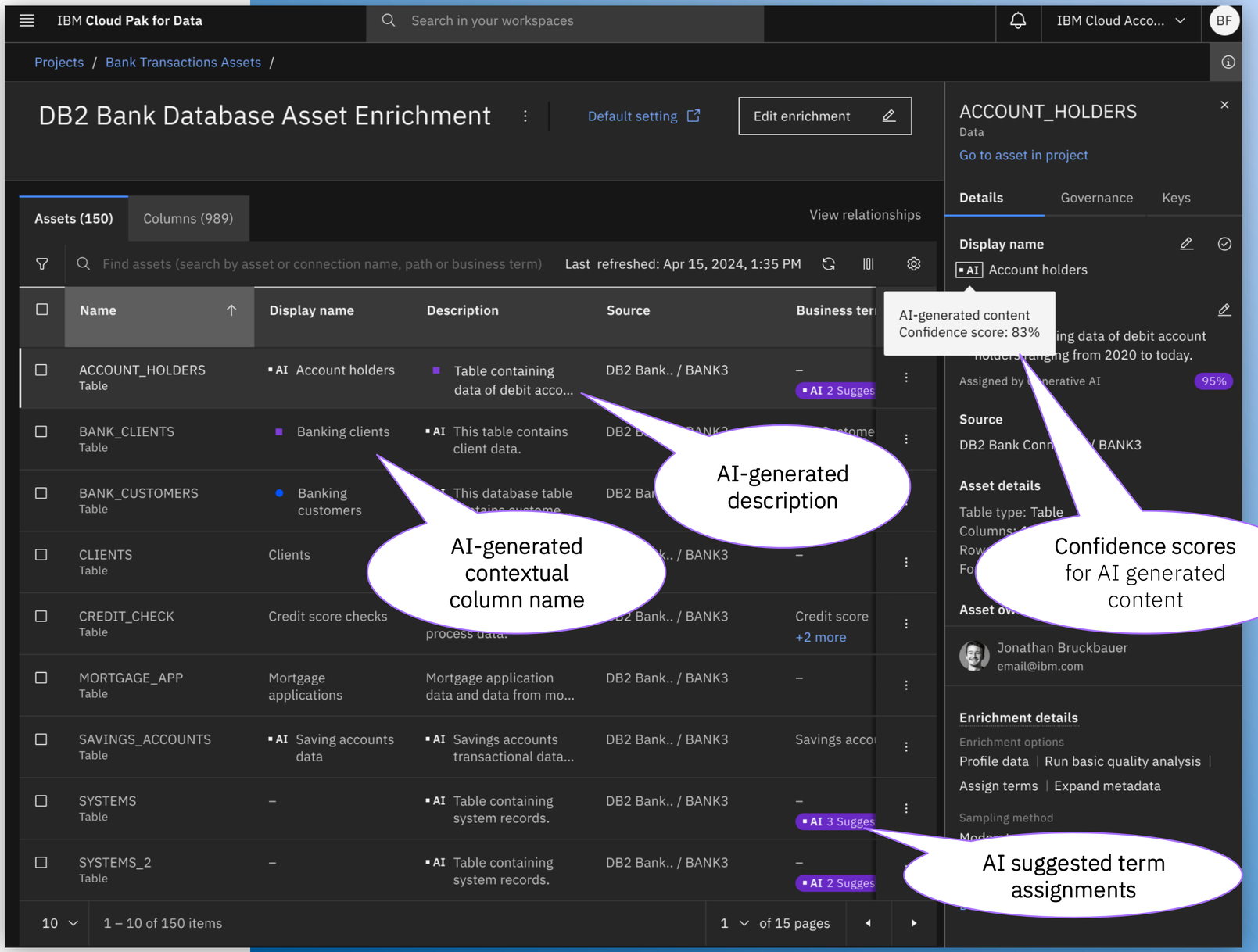
What is it?

- **New -** Automatically generate meaningful column names and descriptions with context
- **Enhanced** - Assign terms based on semantic meaning and context
- **Early results** - Internal testing showed double the number of correct column mappings (reduction in false positives), and improved precision of term assignments with LLMs

Why?

Accelerate data curation through increased accuracy and precision of auto term assignments using AI and LLMs.

Provide Gen AI capabilities to other Data & AI products.



Data Quality

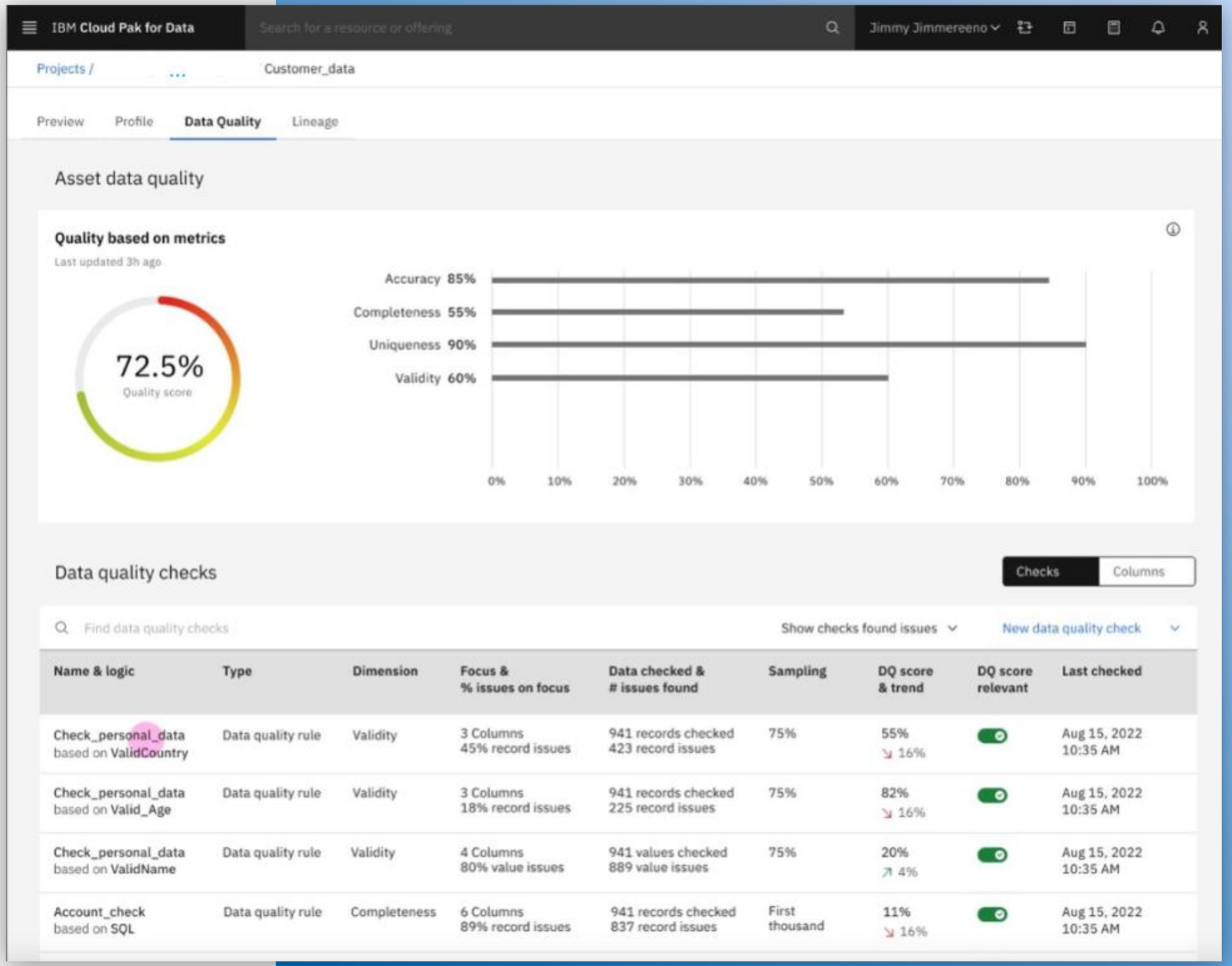
Measure, monitor and maintain quality of your data to ensure that data used in AI and analytics use cases is of high quality and can be trusted

-Automated data quality process with AI, active metadata and knowledge graph

-Score, measure, understand and address data quality issues

-Execute data and quality rules in batch or real-time

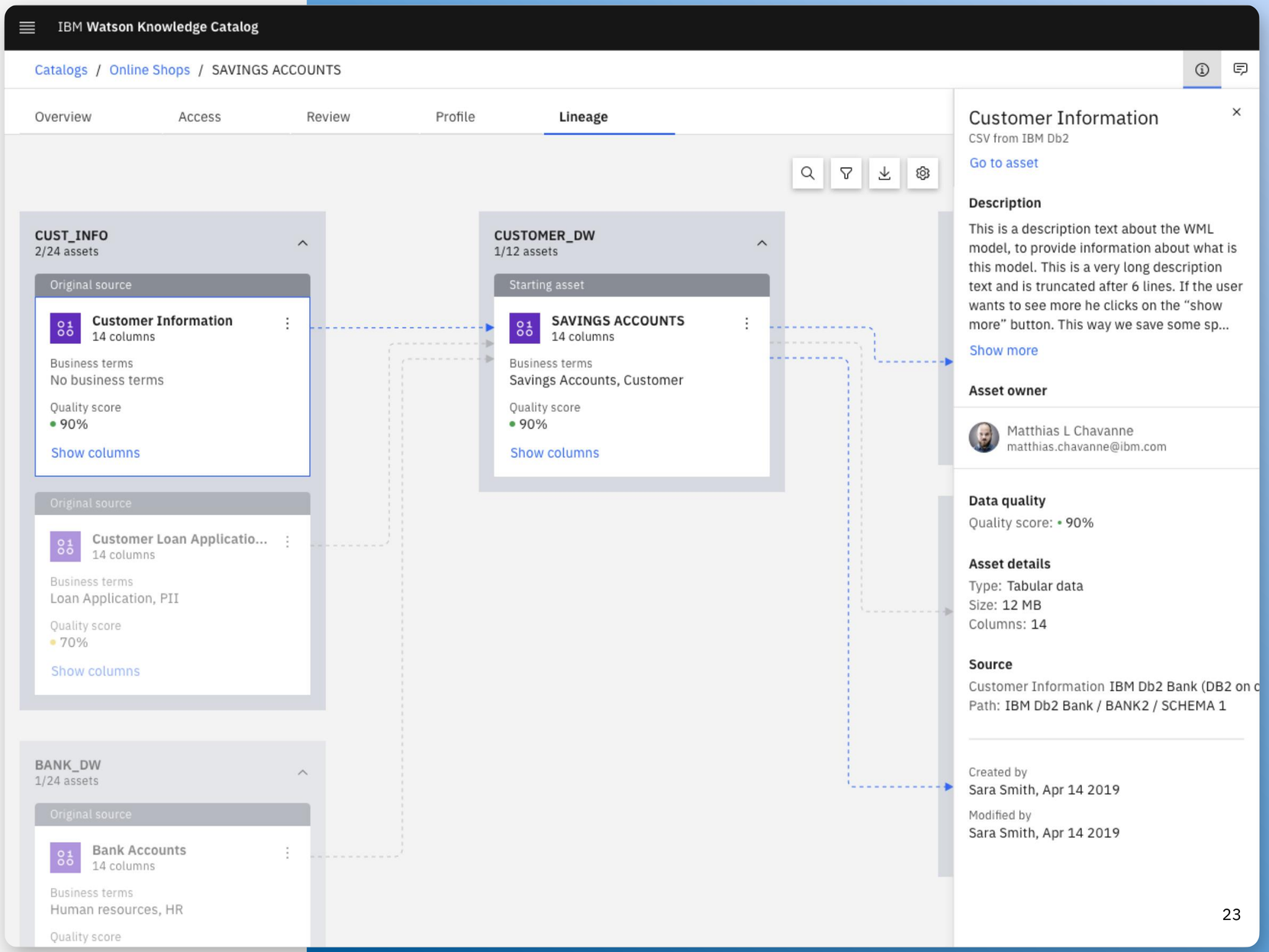
-Set up monitoring of data quality goals for critical data elements so that issues can be addressed in timely manner



Data Lineage

Easy-to-consume data lineage for business users

- Provides business-user-friendly summary view
- Drill down to view technical details
- Automated discovery of data flows from 3rd party tools
- Code analysis of stored procedures & ETL jobs capture transformations
- Technical, Indirect, and Historical data lineage GUI
- Rules to filter for things needed in data lineage



Data Catalog

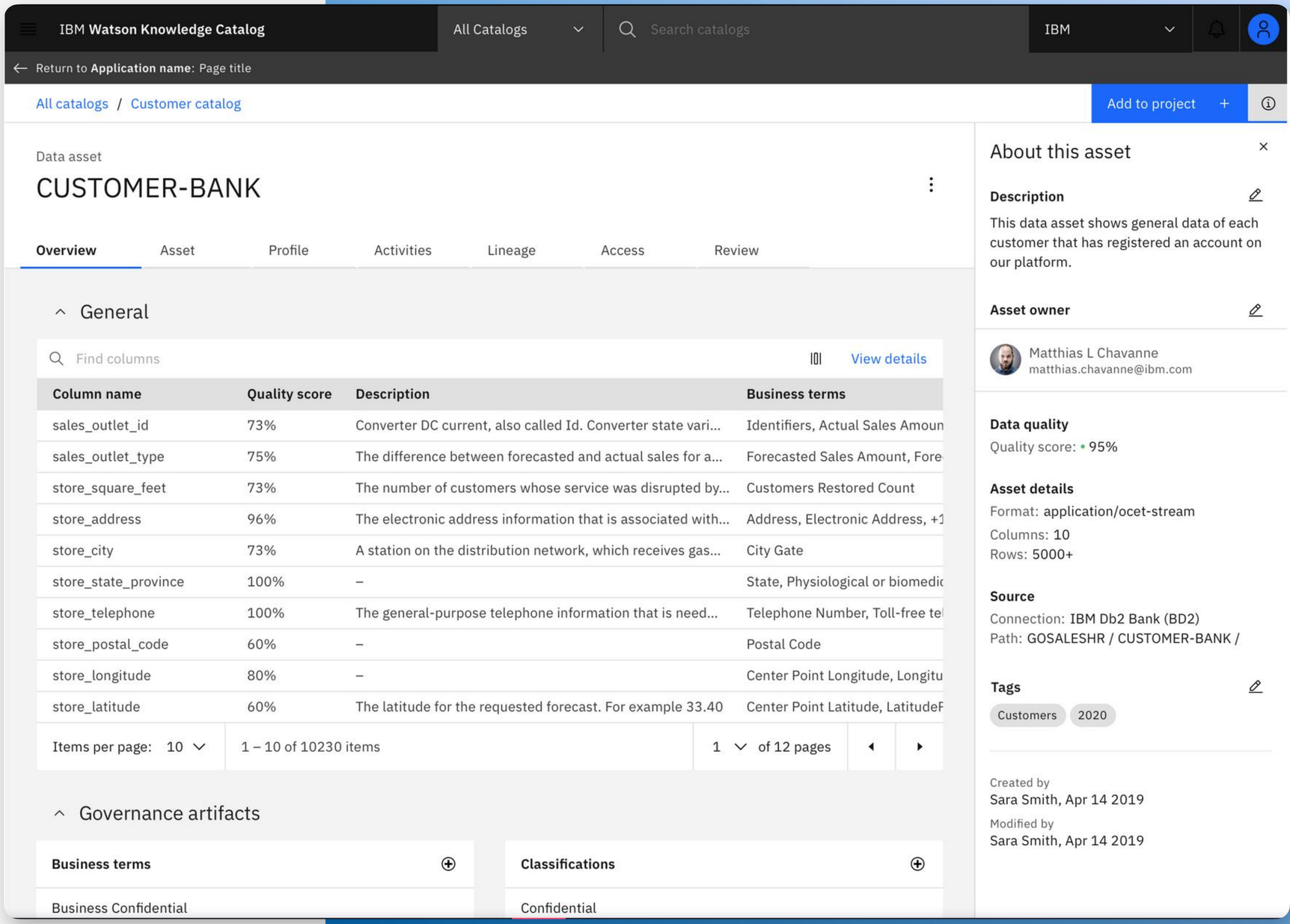
Easily find, understand and share curated and trusted data and other assets in the catalog

-Access connected assets through 70+ supported connectors

-Explore asset metadata and sample data through visualizations and dynamic charts

-Extensible metamodel to inventory any asset type and capture additional details and relationships

-Support domain analytics by ensuring that only the right people can access the right data through access control at the workspace or asset level



Data Privacy

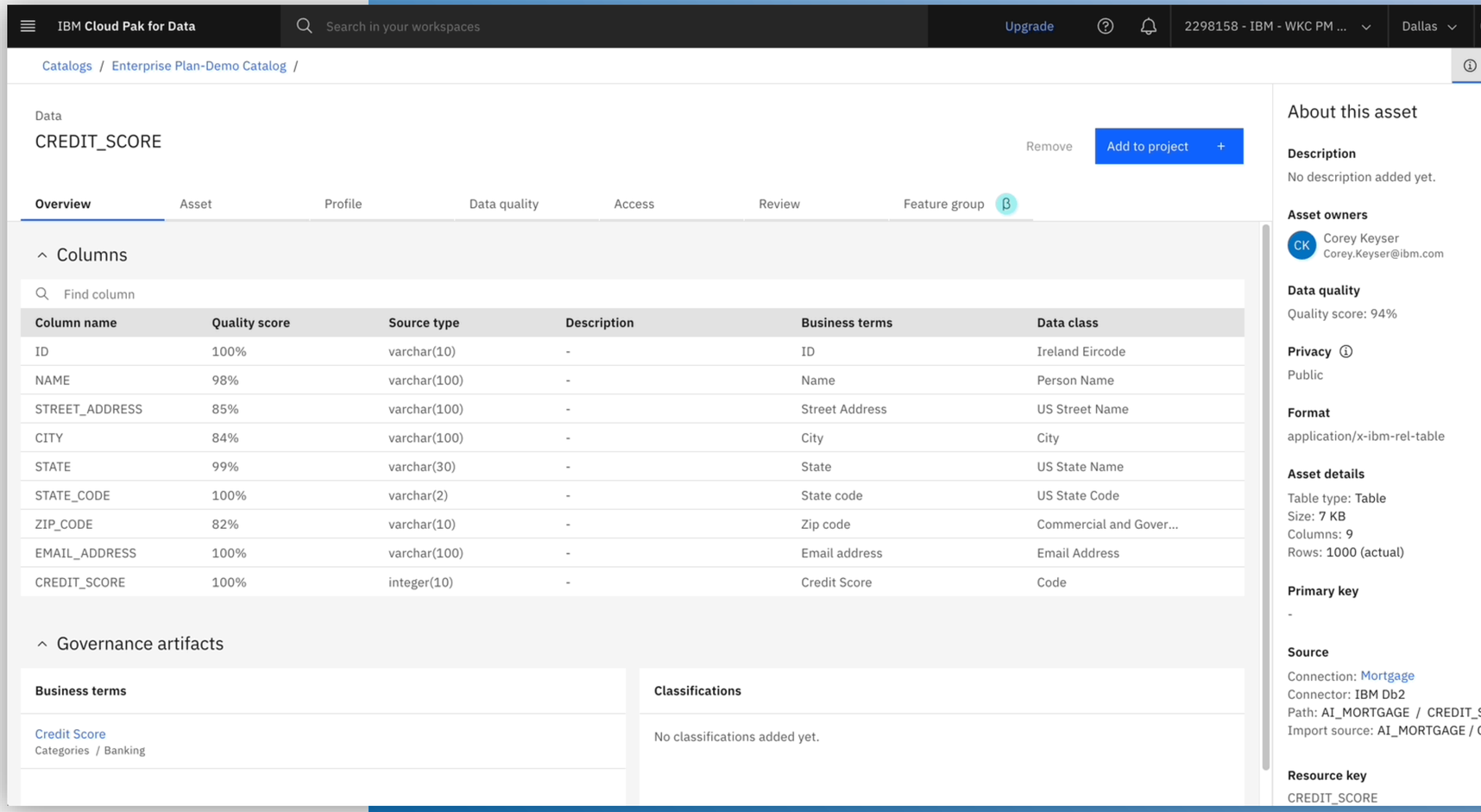
(animated)

Enhance data privacy while preserving format

-Mask data with advanced de-identification techniques such as—format preserving encryption, tokenization, masking and more.

-Trusted data pipelines with de-identification of personal information & confidential information enabling safe data usage in AI, analytics and test data projects when used with integrated deep enforcement solutions through:

- IBM Data Virtualization
- IBM Guardium
- watsonx.data



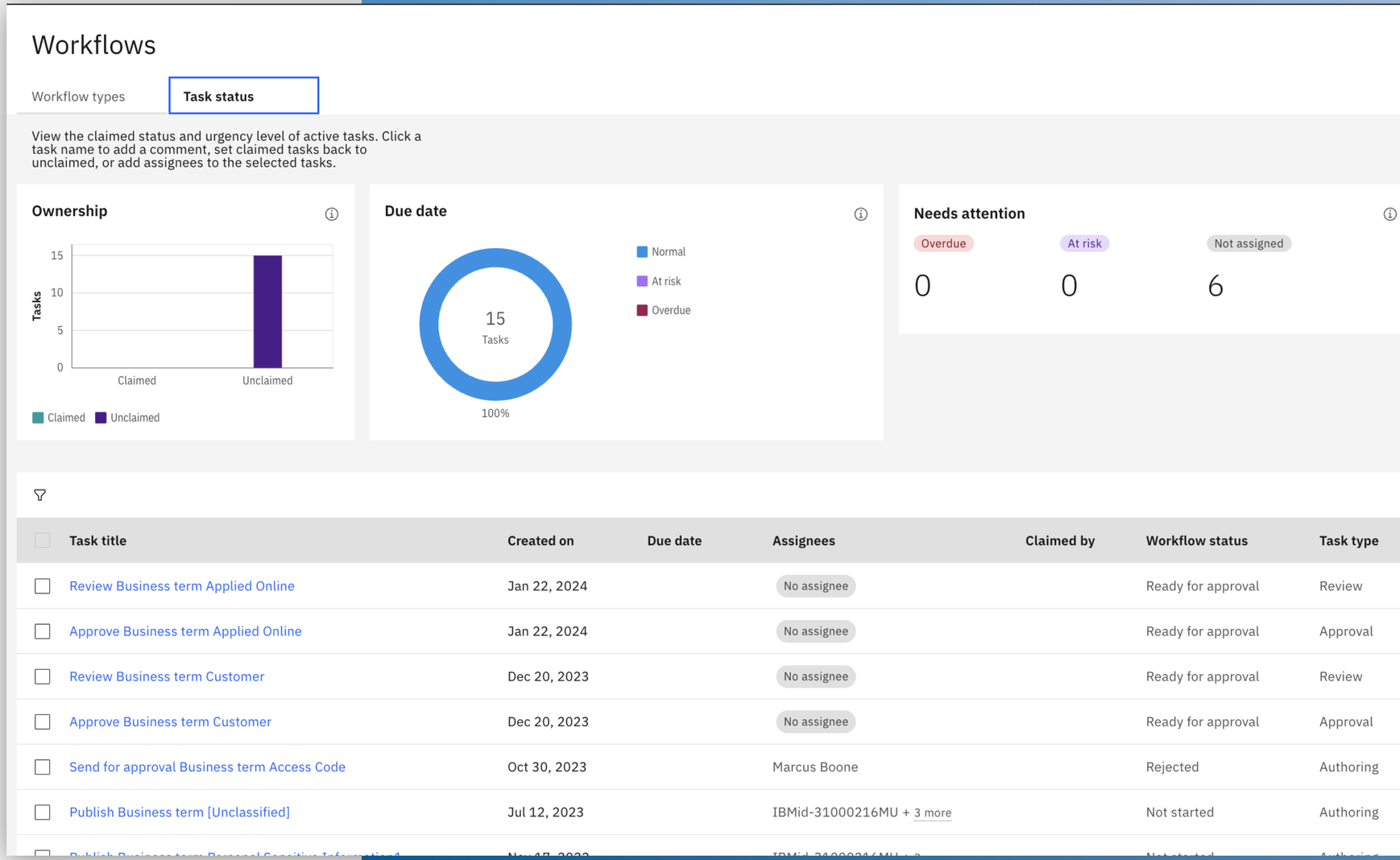
Workflows

Configure workflows to streamline business processes with built-in templates to:

- manage glossary assets
- evaluate potential record matches
- remediate data quality issues

Import custom workflows for data or access requests

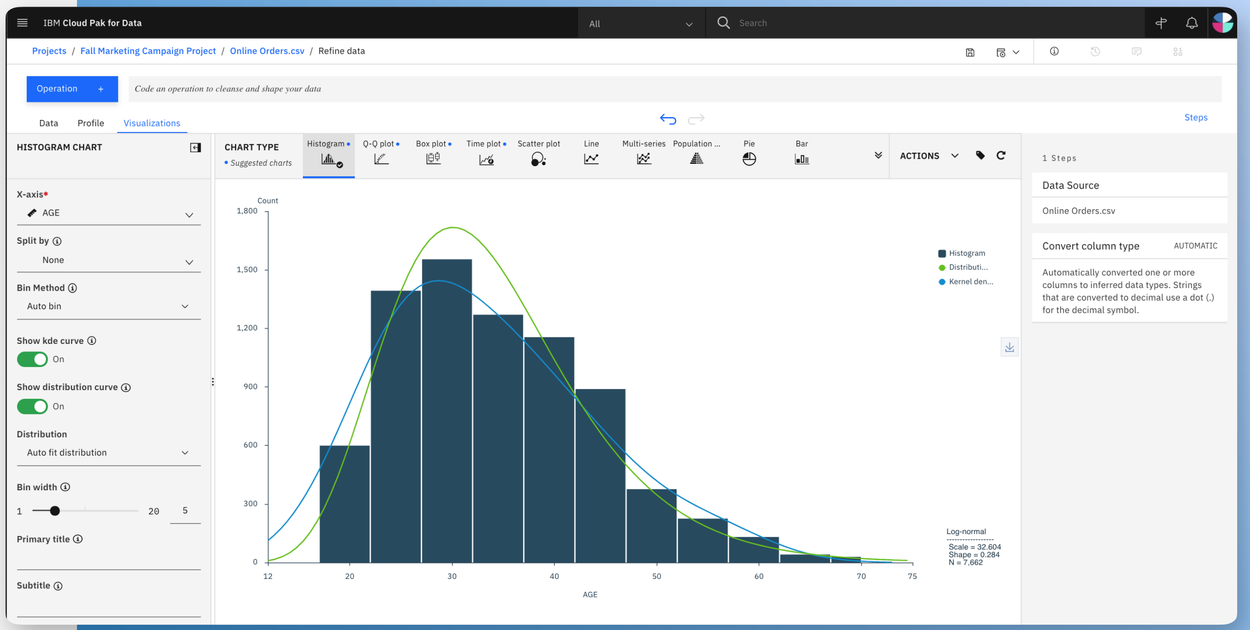
Manage overdue and at-risk tasks through the Task Status Dashboard

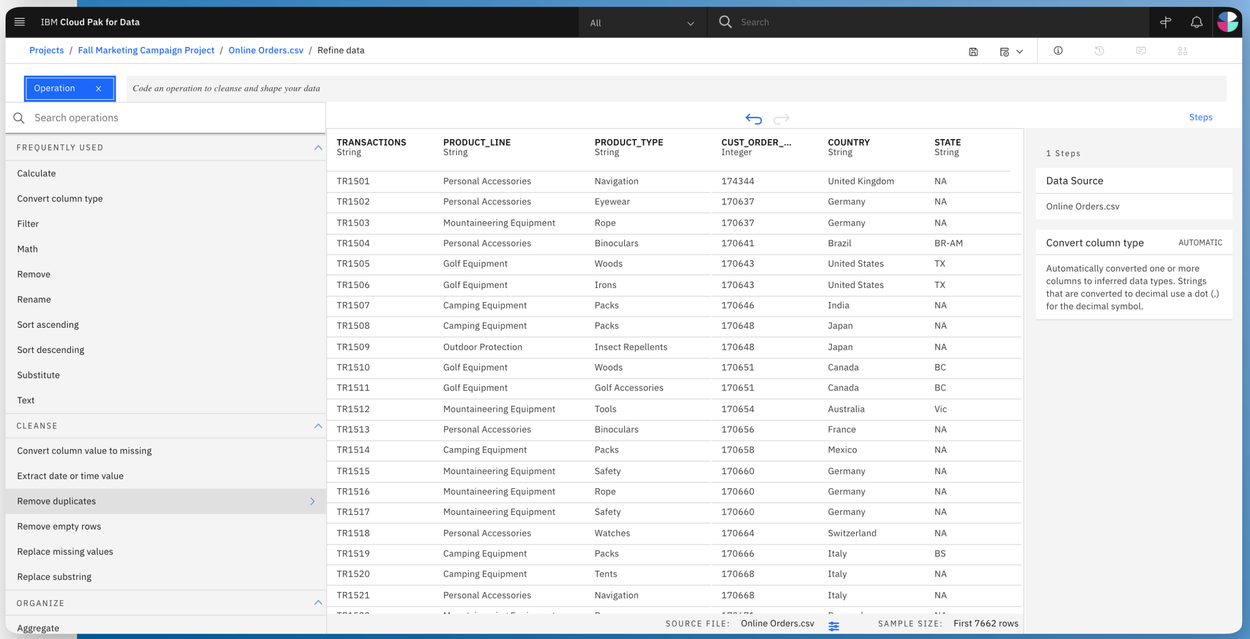


Data Preparation

Prepare raw data for analysis by cleaning, transforming, and enriching it

- Use different built-in charts to visualize, understand and experiment with your data
- Select a dataset and use “click or code” to apply various transformations to create new datasets for analytical models
- User-friendly GUI interface to join datasets from disparate data sources through common connectivity





Thank you

Backup

Intelligent search

Typical use cases
Intelligent search offers users the ability to search for assets and artifacts across all projects, deployment spaces, catalogs and categories.

Key capabilities

Search
Leverage advanced search capabilities such as partial, exact and fuzzy matching with ranked results. Utilize advanced filters to drill down on queries that can be saved for future use.

Accessibility
Intelligent search is persistent and accessible. Users can access, preview and start using assets directly from their search results.

Speed
Leverage knowledge graph to decrease the time spent searching for data.

How generative AI is revolutionizing intelligent search

Search
Intelligent search provides semantic search capabilities to ensure relevant results. Semantic search can surface new insights from the same queries, all while ensuring enterprise-grade trust through LLMs from IBM Research.

Accuracy
By leveraging the graph knowledge, clients can increase the accuracy of their LLMs by **3x** and increase overall search result accuracy by up to **225%.**

Speed
Intelligent search's generative AI capabilities further decrease time spent searching for data.

75%

reduction in time spent searching for data with LLM-based intelligent search

Manta Data Lineage

Typical use cases
Pinpoint the source of streaming data pipeline issues to quickly remediate issues.

Mapping the flow of data from origin to consumption to maximize the value of data governance frameworks.

Seamless cloud or hybrid migrations with complete data lineage, saving time, money, and resources.

Key capabilities

DataOps – Root cause/impact analysis for 95% faster impact analysis and incident resolution.

Data governance and compliance for manual labor cost savings by a large financial institution using Manta for Basel reporting.

Data platform migrations for 30–40% reduction of time and effort for migration initiatives.

How generative AI is revolutionizing Manta

Generative AI creates new challenging regulatory frameworks (e.g., the EU AI Act) that demand stringent documentation on AI model training data provenance, which can easily be accomplished through IBM Manta Data Lineage.

Conversational AI assistants allow nontechnical users to answer lineage-related questions in a natural language interface.

Generative AI reduces the complexity of documenting custom lineage by providing a conversational interface on top of traditional API or CSV custom lineage formats.

95%

reduction in time and effort analyzing scripts and graphing lineage with LLM-based data lineage

IKC Data Catalog

Typical use cases
Only IBM provides Industry Alignment Vocabularies out of the box.

IBM offers AI-powered scalable automation to curate data assets at up to 100 columns per second.

Metadata-driven automation for improved productivity.

Key capabilities

Up to 70% reduction in manual mapping of terms to industry regulations, and terms to assets.

Jump-start creation of business glossary for shared context and understanding.

Improve data steward productivity by assigning terms to data assets and columns with ML, LLM, data class-based assignments or linguistic matching.

Automate classification of data assets for faster identification of, and reporting on, assets with sensitive, personal or critical data.

Automate the application of data protection rules in catalogs through watsonx.data or Guardium data protection based on role, or attribute with up to 80% less manual intervention.

How generative AI is revolutionizing Data Catalog

Documenting context about technical assets so they can be easily found and used can be tedious or even impossible in large organizations with countless data assets.

Generative AI is helping organizations add context to assets by understanding their meaning and extrapolating them into plain language.

This helps data analysts more easily find the assets they’re looking for and helps compliance teams more readily meet obligations.

$2M–$5M

in cost savings with LLM-based metadata enrichment

Data Product Hub

Typical use cases
Accelerate AI initiatives: Surface data points from lakehouses as data products, which can be shared directly in watsonx.ai projects or through common APIs in Python notebooks.

Data sharing: Package commonly requested data assets from IBM Knowledge Catalog into data products to make available to data consumers for self-service access through a data marketplace (accessible via Python notebooks or CSV files in IBM Cloud Object Storage).

Third-party lakehouses: Convert valuable assets from third-party lakehouses such as Snowflake (e.g., metrics, dimensions, etc.) into personalized SQL queries available to business analysts for self-service access in data marketplaces.

Key capabilities

Internal marketplace
Discover data products with ease. Create a single pane of glass for self-service data consumption.

Data as a product
Sharing data in a large-scale manner requires a special level of compliance. Data products are optimized for large-scale discovery and delivery to consumers across an organization.

Open ecosystem
Adapt to an organization’s existing technology stack. Onboard assets from virtually any tool, IBM or third party into a single hub and transform them into data products for easy discovery and access by data consumers.

How generative AI is revolutionizing Data Product Hub

LLMs enable data marketplaces to enhance self-service capabilities through text-to-SQL.

Users are empowered to perform analyses and access the correct data through low-code/no-code UIs.

Data Product Hub will be supporting this capability in future.

$1.2B

reduction of financial risk of data governance burden across organizations

IBM Data Intelligence streamlines access to trusted, quality data while addressing privacy and compliance requirements

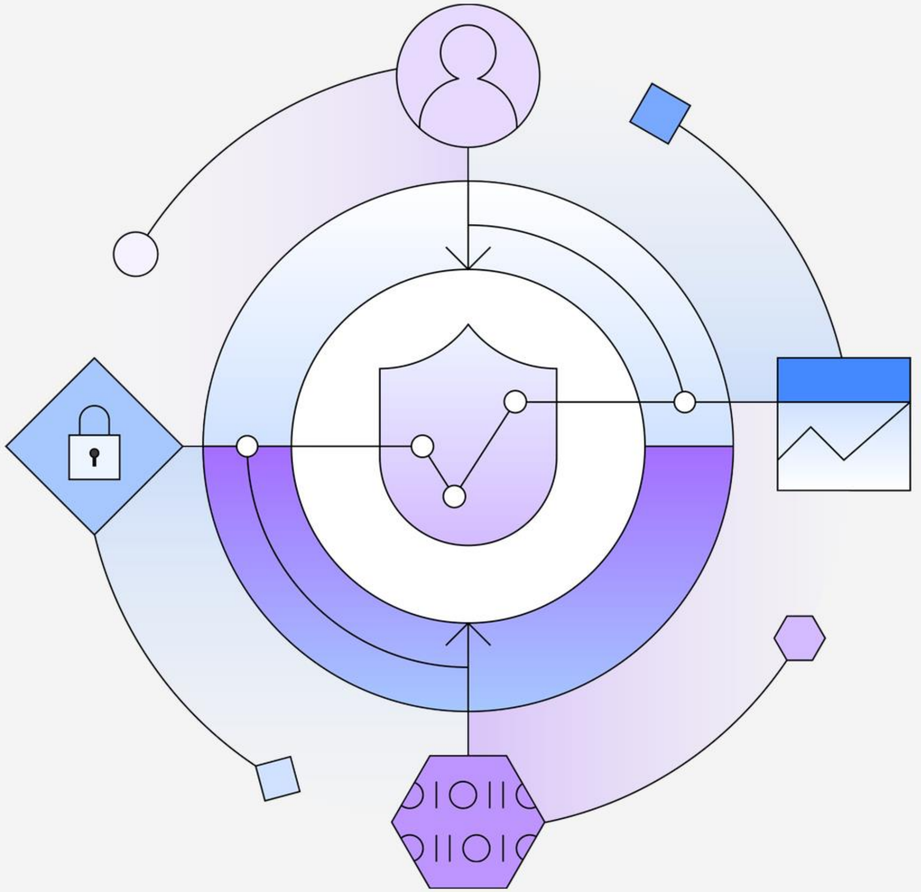
01 Establish a governed data foundation

02 Protect sensitive data with privacy-preserving techniques

03 Support compliance with industry regulations

04 Deliver quality, trustworthy data, maximizing business value

05 Simplify management of data products for analytics and AI



01
Establish a governed data foundation

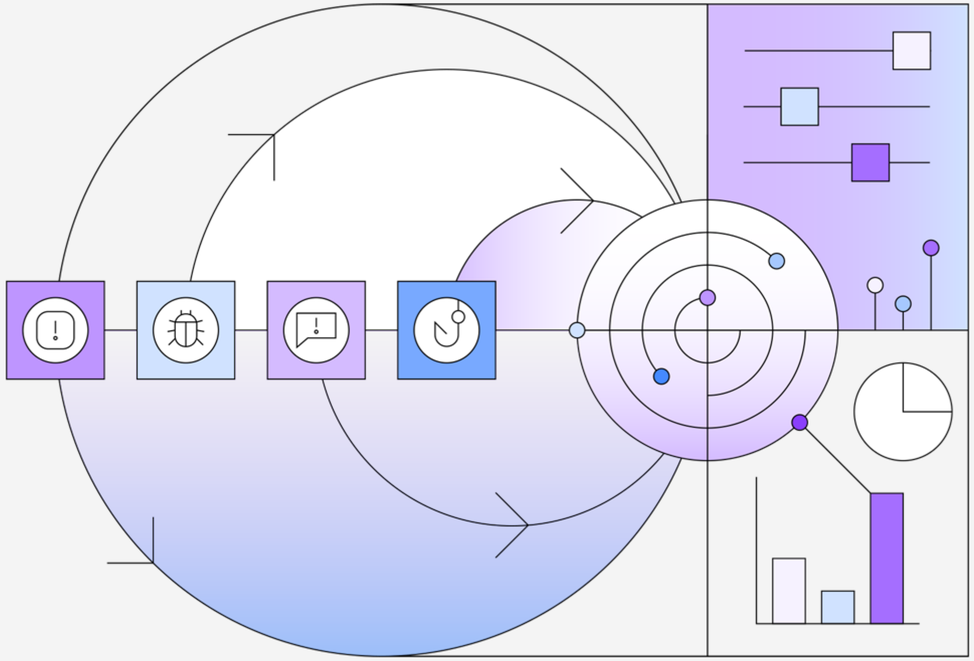
– Consistent understanding of data

– Advanced data discovery and curation

– Simplified policy management

Key capabilities of IBM Knowledge Catalog

– **Creation of a business glossary** of governance artifacts, terms, classifications, data classes, reference data, policies and rules
– **Customizable workflow management** ensures proper review, approval and publication processes for governance artifacts
– Discover, curate and provide governed access to data with advanced data discovery and **AI-powered metadata enrichment** capabilities
– Evaluate data quality using automated data quality analysis based on business rules to measure, identify and prevent data quality issues
– Achieve high-quality, comprehensive views of data using **knowledge graph-based matching** driven by machine learning

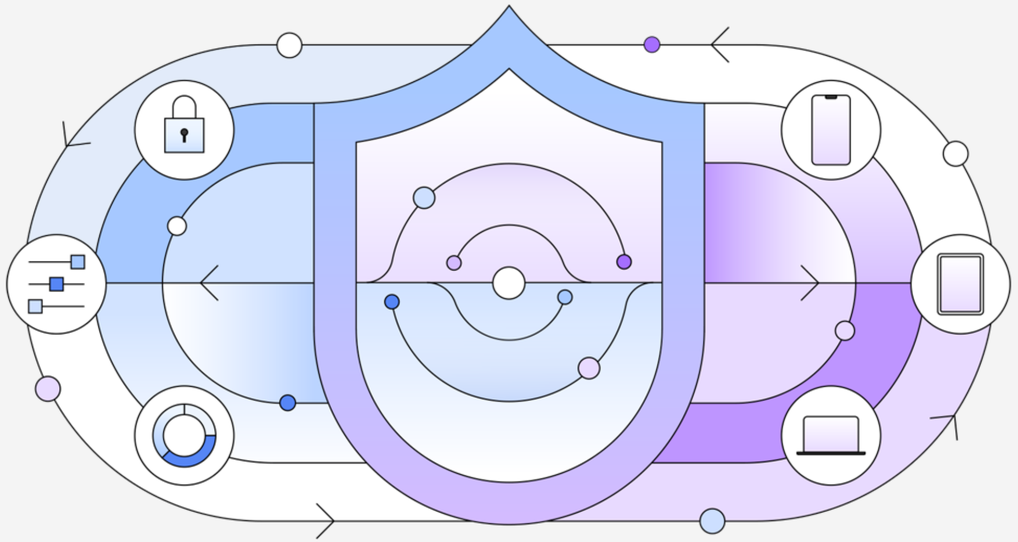


02
Protect sensitive data with privacy-preserving techniques

- Mitigate data privacy risks
- Protect distributed data assets
- Improve data utilization with trusted data

Key capabilities of IBM Knowledge Catalog

- **Create data** protection rules to control data access within catalogs, virtualized data assets, data refinery and masking flows
- **Dynamic enforcement** of ABAC rules as sensitive data is accessed in governed enforcement points
- Identify sensitive and personally identifiable information in order to help **manage risk**
- Protect sensitive data with **advanced de-identification** techniques such as format-preserving masking and hashing
- Protect data outside of Cloud Pak for Data with masking flows and **integrations with IBM watsonx.data**, IBM Data Virtualization and IBM Security Guardium

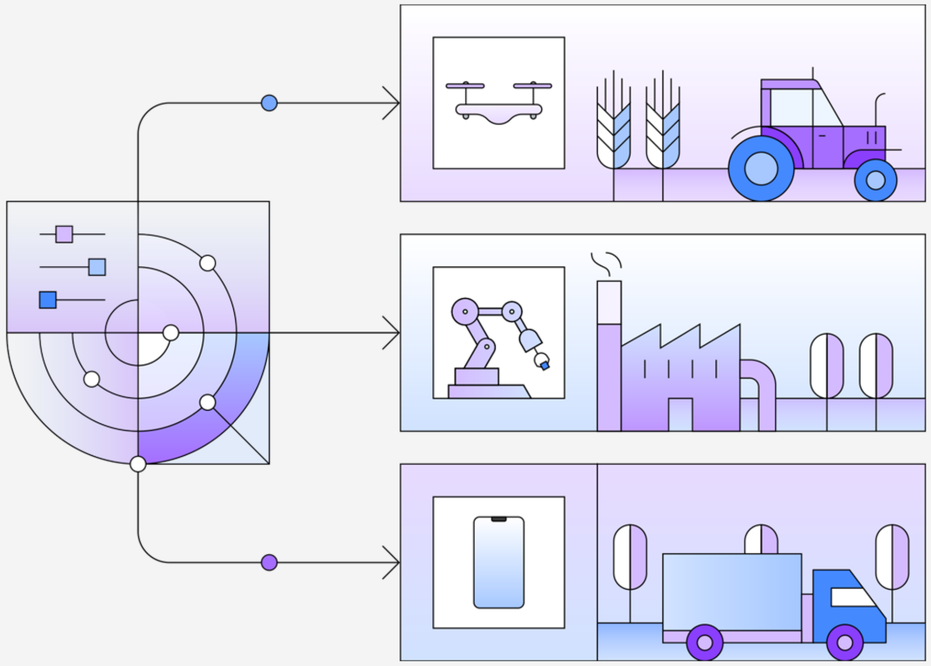


03
Support compliance with industry regulations

- Simplify compliance with new regulations
- Promote auditability with data lineage
- Proactive data quality management

Key capabilities of IBM Knowledge Catalog

- **Design and manage** policies and rules to describe organizational guidelines, industry regulations and standards
- Accelerate regulatory compliance and reporting with **built-in Industry Alignment Vocabularies**
- **Simplify compliance** by visualizing relationships between data and governance artifacts including policies, rules, and classifications
- Gain visibility into where data came from, how it has changed and who is using it with **automated data lineage** in IBM Manta
- Assign data quality SLAs to your critical data elements and simplify curation through **AI-driven data quality rules**

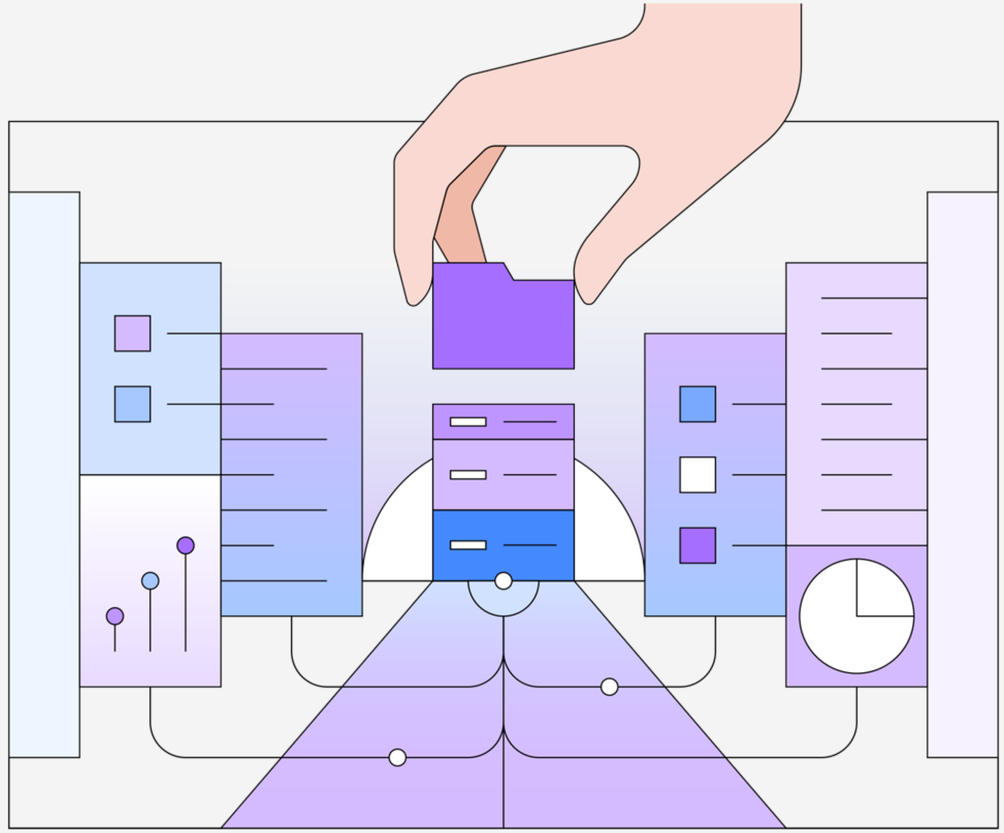


04
Deliver quality, trustworthy data, maximizing business value

- Automated data quality scoring out of the box
- Proactively monitor for changes in data quality
- Remediate data quality issues at scale

Key capabilities of IBM Knowledge Catalog

- View **high-level quality scores** based on 7 key dimensions simply by running metadata enrichment
- **190+ data quality definitions** available out of the box and support for custom-built rules
- Apply data quality rules and **automatically monitor quality goals** using DQ SLAs
- **Prebuilt workflow engine** accelerates tracking of DQ issues for remediation
- **Interactively cleanse and transform data** with 100+ built-in operations; no coding skills required



05
Simplify management of data products

- Self-service access to governed data
- Share trustworthy data products across the organization
- Business-friendly publishing experience for data assets

Key capabilities of IBM Knowledge Catalog and IBM Data Product Hub

- Reliable marketplace for data consumers to **discover, understand and access governed data products** underpinned by contracts
- Enforce data contracts to embed governance into the data-sharing process
- Advanced data quality capabilities including automated quality profiling and SLA rules to **deliver right-quality data for analytics and AI**
- **Natural language-powered** semantic search to quickly find, understand and use governed, high-quality, relevant assets
- **Automated profiling and classification** of data assets for data consumers to gain insight into data asset content and context