SPOTLIGHT

Sponsored by: Hitachi Vantara

This IDC Spotlight looks at the demand for better intelligence about data and the benefits of a data catalog.

Data Catalogs: Intelligence in the Modern Data Environment

March 2022

Written by: Stewart Bond, Research Director

Introduction

AT A GLANCE

WHAT'S IMPORTANT

Modern data and work environments demand better intelligence about data, which enables organizations to become intelligent enterprises in a digital-first world. The cornerstone capability that enables gathering, analyzing, and leveraging intelligence about data is cataloging.

We are at a point in history when data — its creation, management, analysis, and use — is universally accepted as a foundational asset for every organization. Executives now openly articulate the need for their organizations to be more data driven, to be "data companies," and to increase their enterprise intelligence. A global IDC survey fielded in the summer of 2021 indicated that 77% of organizations have a senior-level executive responsible for leading enterprise intelligence, which refers to an organization's capacity to learn and ability to synthesize information and deliver insights at scale. According to worldwide responses received by IDC from IT decision makers earlier in 2021, leveraging data and improving decision making to remain competitive and exploit changing market conditions is the second highest area of strategic interest at the board level after customer engagement.

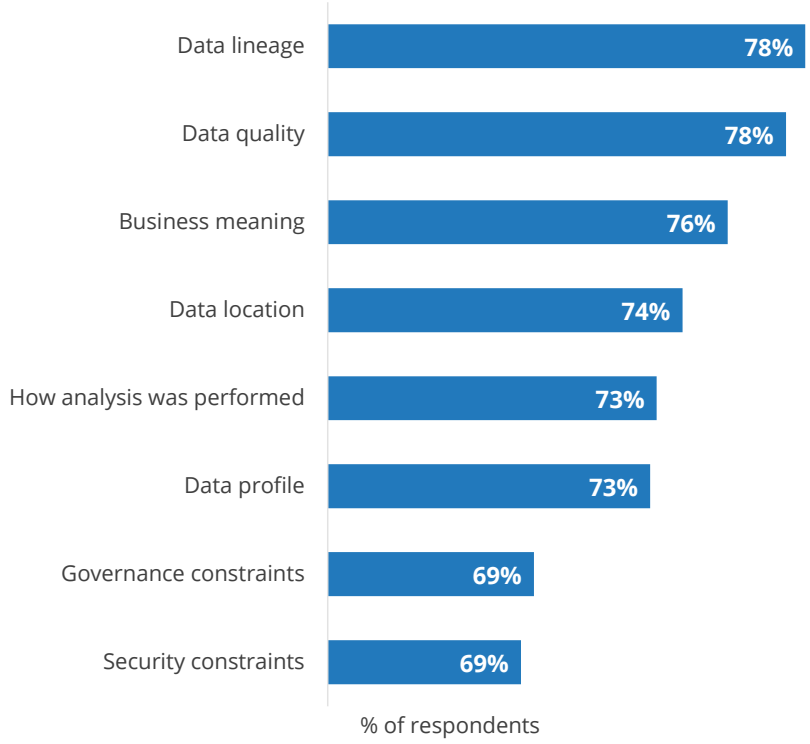
With the desire to be more data driven, there is an expectation that nearly two-thirds (62%) of employees are using data to make decisions (IDC's *Data Culture Survey,* December 2020). But only a third of people strongly believe their actions are actually driven by data. Half (50%) of the respondents to the same survey said they are overwhelmed by the amount of information available and cannot find the signal in the noise, while slightly less than half (46%) said there isn't enough information available to make decisions! In IDC's December 2021 *Data Trust Survey,* only a quarter of the people (27%) noted that they completely trust data. Clearly, gaps exist between expectations placed on the use of data and the ability to trust and use data in delivering data-driven business outcomes.

When the term "big data" emerged, it was defined by the three Vs: volume, variety, and velocity. Organizations tried to take control of big data by centralizing it in a data lake. Today, we have a larger and broader notion of data, regardless of where it is located; data is now characterized as being highly distributed, diverse, and dynamic. These three Ds are characteristics of the modern data environment that are contributing to the gaps that exist between expectations and reality of data use that are getting in the way of achieving enterprise intelligence in the advent of a digital-first world.

The reason half the people think there is too much data and the other half think there is too little is because there isn't enough intelligence about the highly distributed, diverse, and dynamic data. Figure 1 shows what respondents to IDC's *Data Culture Survey* (December 2020) expect or demand to know prior to making a data-driven decision.

FIGURE 1: ***Data Intelligence Expectations***

Q When making data-driven decisions, which of the following do you expect and demand to know?



n = 455

Source: IDC's Data Culture Survey, December 2020

At the top of the list is intelligence about the data: Where did it come from? What quality level is it? What does it mean to the business? Where is it? What does it look like? Despite the fact that this type of intelligence is expected and demanded, organizations are not doing a great job providing it. In IDC's December 2021 *Data Trust Survey,* respondents were asked to rank 10 data management activities by how well they believed their organization performed them. Profiling data, reporting data quality scores, locating the best data for solving problems, documenting the lineage of data, and enabling self-service access to data by nontechnical users ranked lowest.

Not only is the data distributed, diverse, and dynamic, but so too are the people who work with the data, in terms of responsibilities and roles within the organization and where they are physically located. IDC has defined a new generation of data-native workers: Generation Data, or Gen D for short. Gen D is not just made up of those workers who have "Data" in their job title; it also includes many other roles throughout the organization, all of which are part of the data culture. Gen D is not a chronological generation, but a vocational generation. If you are asked to make data-driven decisions, expect to know the provenance of data, have a sense for fake data or information, question recommendations made by artificial intelligence (AI)-based systems, communicate using data, or contribute to the knowledge of data in the organization, you are part of Gen D.

The global pandemic also had a significant impact on where people are physically located, and as we emerge from its disruption, we are seeing a shift toward more remote and hybrid work environments. In hybrid work models, organizations face the challenge of achieving employee experience parity. Hybrid workers do not require technical parity to have experience parity. What they need is parity of access to people, systems, knowledge, data, and data intelligence, which is critical to achieving experience parity.

Modern data and work environments demand better intelligence about data, which enables organizations with data to become intelligent enterprises in a digital-first world. The cornerstone capability that enables gathering, analyzing, and leveraging intelligence about data is *cataloging.*

Benefits of a Data Catalog

The demand for better intelligence about data isn't just a future requirement — it has increased over the past few years with the introduction of data privacy regulations, and it is a critical requirement in digital transformation initiatives. The market has responded with a new generation of data catalog solutions. Data catalogs have existed for years in technical metadata management software, spreadsheets, and centralized document repositories — and inside the heads of data-native workers. However, these catalogs are not enough to keep up with the demands of a highly distributed, diverse, and dynamic population of Gen D workers for highly distributed, diverse, and dynamic data. The new generation of data catalog is automated, intelligent, business aware, and instrumented, transforming intelligence about data into knowledge to meet the needs of a new generation of data-native workers.

This new generation of data catalog:

- Captures technical, business, lineage, operational, governance, and relationship metadata (data about data) from multiple types of data repositories and formats across the enterprise to inform data-native workers about where data is and what it means, if it needs to be protected, and the context in which it should be used.
- Leverages modern AI and machine learning (ML) and rules-based automation to improve the efficiency of gathering and applying data intelligence at scale.
- Captures metadata about analytic models, visualizations, dashboards, and reports that deliver information and insight to operational and analytic decision makers, human and machine.
- Supports multiple-persona user experiences to boost the diversity of the data-native worker, enabling crowdsourcing and collaboration among users to improve corporate data knowledge.

Data intelligence is the future of the unified data platform in the future enterprise. File systems, databases, data warehouses, object stores, and data lakes all have their purpose and merit, but not one of these has been sufficient in unifying data types, data formats, and data management technologies. Centralizing data into a data warehouse or data lake platform isn't the answer to solving the problems of highly distributed, diverse, and dynamic data. Collecting and centralizing intelligence about the data is the data platform, because data intelligence answers the who, what, where, when, why, and how questions of data, from production through to consumption. Data-driven decisions are informed by data intelligence, where data-native workers and machines can synthesize information, improve their capacity to learn, and automate insights at scale.

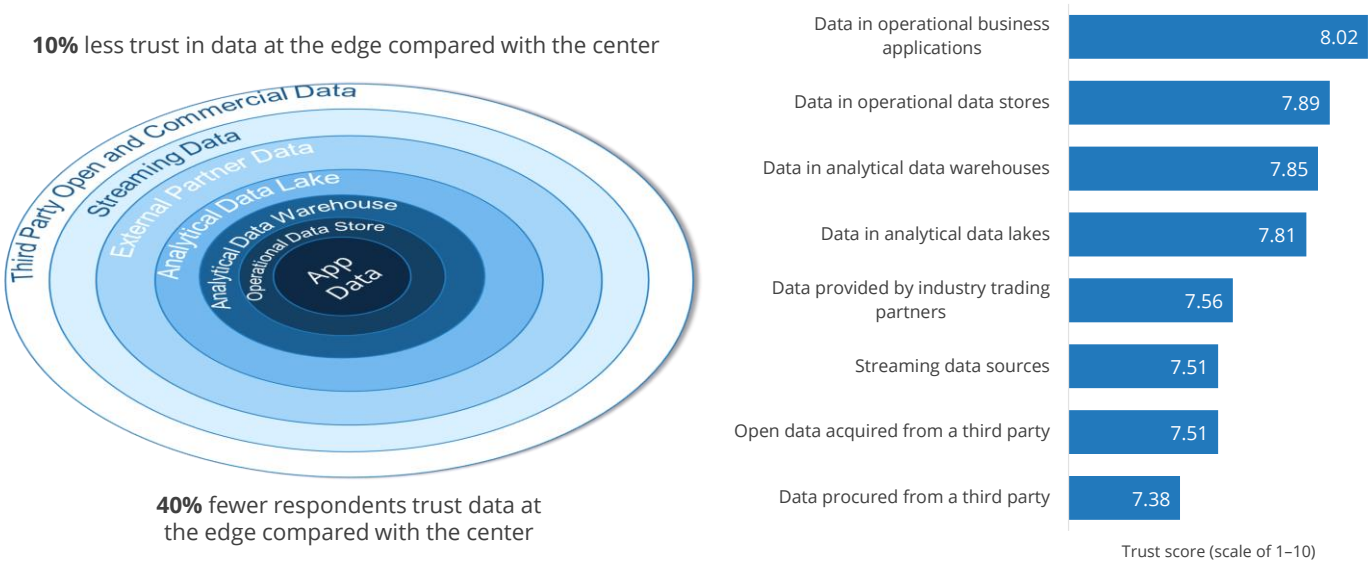
Trends in Data Cataloging

Data is both the lifeblood of a digital-first world and the common thread that runs through the future enterprise, but without trust in the data, there is no future in digital. Trust in data requires transparency of data location, schema, origins, transformations, quality scores, meaning, and context. Data cataloging helps provide this transparency, allowing organizations to gain control of highly distributed, diverse, and dynamic data, improving the organizational ability to trust data and deliver trusted data-driven outcomes.

Figure 2 shows what type of data is trusted the most in the organization and what type is trusted the least.

FIGURE 2: ***Trust Degradation***

Q ***How much trust do you have in each of the following?***



n = 500

Source: IDC's Data Trust Survey, December 2021

The highest level of trust in data exists at the origin of enterprise data: the applications through which the business creates transaction, interaction, and domain-specific data. As data is extracted and ingested into operational and analytical data stores, it changes shape and can lose its business context and meaning. Data catalogs are being used to harvest intelligence about data, maintaining context and meaning throughout its lineage from data producer to data consumer.

According to IDC, nearly two thirds (63%) of organizations reported that higher levels of data trust have positive impacts on business metrics. Customer satisfaction benefits the most, with an improvement of 42%, and revenue benefits the least at 29% — but even a 29% increase in revenue may be considered substantial in many markets: Conversely, 11% of organizations reported that lower levels of data quality and trust have a negative impact on business metrics, particularly with respect to operational costs. Data intelligence provides transparency of data and builds data trust in Gen D to deliver trusted business outcomes with measured benefits.

Technology is only part of the solution to gaining more control over highly distributed, diverse, and dynamic data. Process and method are also critical, and this is where DataOps has emerged. DataOps is a combination of technologies and methods with a focus on quality for consistent and continuous delivery of data value, connecting producers and consumers of data to extract value for any data-driven outcome. DataOps applies the continuous development, testing, and deployment principles of DevOps to data and analytics pipelines, adding statistical process controls to manage the quality of data and data products. Intelligence about data provided by data catalogs is critical in the continuous testing of data definitions, values, and context of data flowing within pipelines against acceptable tolerances, policies, and thresholds to stop bad data from being used in decision making and protect against data governance and compliance exceptions.

The three Ds of modern data and the need for continuous collection and analysis of metadata to optimize and control the path between data producer and consumer while keeping within governance and compliance guardrails demands intelligent automation. Increasingly, the use of machines and probabilistic methods in gathering and learning from intelligence about data enables intelligent automation in data management and governance functions. Respondents to the IDC *Data Trust Survey* indicated that most organizations are using a combination of deterministic and probabilistic methods across data management functions, but in those that primarily use probabilistic methods, trust is 40% higher than in those that primarily use deterministic methods.

There has been a significant shift to remote work, and many believe that shift will remain and become permanent. As we transition to hybrid work environments — onsite and remote — improving the hybrid employee experience, including access to data, is paramount. Ensuring parity of access requires intelligence about data to be centralized and available to workers everywhere and anywhere work is happening.

Considering Hitachi Vantara's Data Catalog, Part of Lumada DataOps Portfolio

Hitachi Vantara acquired Waterline Data in 2019 to address the needs of business and data professionals in finding and extracting value from big data. After creating the Data Catalog product, the organization acquired IO-Tahoe to add data quality and governance capabilities. The Data Catalog enables data-native workers to gain intelligent data insights through auto discovery, enrichment, privacy, lineage, and data rationalization capabilities, making data actionable and trusted. Hitachi Vantara's Data Catalog provides several benefits to data-driven organizations:

- Establishes a unified view of data to rapidly discover data across multiple repositories
- Builds trust in data through provenance validation, data quality improvements, and lineage
- Permits data curators and stewards to manage sensitive data and data quality
- Eliminates redundant duplicate data to reduce storage cost and lower risk of using incorrect data

Data exploration is possible using business glossary terms and internet-like search for data to help Gen D become more efficient: spending less time searching and more time using data in value-added activities. Data governance can be automated through application of data policies and integration of tagged data with security infrastructure, reducing data protection costs while lowering risk. Automated cataloging capabilities can also find data hidden from manual compliance reporting processes. Hitachi Vantara's Data Catalog allows data curation to become a full team effort through crowdsourcing knowledge about the quality and usefulness of data assets and identifying the most valuable assets. The catalog can also intelligently infer missing data lineages, providing users with additional insights to help them select the best available data for their projects.

A key differentiator of Hitachi Vantara's Data Catalog is the level of automation it provides. Its data fingerprinting technology works by analyzing the data values in each data set and accurately profiling the data. It uses that information to create a "fingerprint" for each column of data — using machine learning to intelligently and automatically tag and match data fingerprints to glossary terms and populate the data catalog. Users can refine matched terms and remaining unmatched terms through crowdsourcing and collaboration. The Data Catalog then layers on additional capabilities such as automatic data quality assessment, identification of standard and custom sensitive information, applying business rules for streamlining processes, and identification of duplicate data anywhere in an organization's modern data environment. This level of automation, combined with augmentation and awareness, helps make data asset management possible in the highly distributed, diverse, and dynamic data environments of organizations trying to compete in a digital-first world.

Hitachi Vantara offers the Data Catalog as part of its broader Lumada DataOps portfolio, which provides intelligent data management for digital innovation through advanced insights based on trusted data. The product portfolio is built on an open and composable architecture to deploy intelligent data pipelines that require ingestion and transformation, cataloging, data optimization, analytics, and edge intelligence.

Challenges

The two primary challenges ahead for Hitachi Vantara in taking its Data Catalog to market are related to organizational priorities in data management and user adoption of data intelligence technologies. Despite data intelligence being at the top of the list of things that data-native workers demand to know when making data-driven decisions, metadata management capabilities were the least important and least implemented function among respondents to IDC's *DataOps Survey.* This was also apparent in results from IDC's *Data Trust Survey,* where identifying lineage, locating data, profiling, and calculating data quality scores were at the bottom of the list of processes performed well by organizations. Organizations need to recognize the importance of data intelligence in building a data-literate culture among Gen D workers and prioritize metadata management solutions such as Hitachi Vantara's Data Catalog to turn intelligence about data into knowledge.

User adoption is a highly rated concern among organizations that have implemented data catalogs. Although catalogs can provide a lot of automation capabilities that deliver value with little effort, like anything else, you only get out of them what you put into them. Tacit knowledge added through crowdsourcing and working with the machine learning capabilities to better train the models will deliver even more value. Hitachi Vantara's Data Catalog provides capabilities that enable efficient contribution of knowledge and lower the barriers to adoption.

Conclusion

Despite the fact that executives have widely embraced the need for their organizations to be data driven, data-native employees remain hampered by the challenges of highly distributed, diverse, and dynamic data in being able to trust data and make data-driven decisions. Data catalogs are enabling a new generation of data-native workers in locating, understanding, and leveraging data to improve business outcomes and close the gap between expectations and abilities. Automation of data intelligence collection and curation is a key success factor in modern data environments. It is also a differentiator for Hitachi Vantara, and to the extent that the company can address the challenges described in this paper, it has a significant opportunity for success.

Automation of data intelligence collection and curation is a key success factor in modern data environments.

About the Analyst

Stewart Bond, *Research Director, Data Integration and Data Intelligence Software*

Stewart Bond is Research Director of IDC's Data Integration and Intelligence Software service. Mr. Bond's core research coverage includes watching emerging trends that are shaping and changing data movement, ingestion, transformation, mastering, cleansing, and consumption in the era of digital transformation. Having worked in the IT industry for over 25 years, from early experience in database and application development, through solution design and deployment, to strategic architectural consulting, Stewart has worked through some significant changes in the IT industry.

MESSAGE FROM THE SPONSOR

Hitachi Vantara's Data Catalog, part of its Lumada DataOps portfolio, enables powerful business insights based on trusted, actionable data through auto discovery, enrichment, quality improvement, privacy by design, and data rationalization.

- Better understand your data
- Increase your business agility
- Build high-quality data products
- Lower privacy and compliance risks

Read more about the product here: https://www.hitachivantara.com/en-us/products/data-management-analytics/lumada-data-catalog.html

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com