Data Catalogs Are the New Black in Data Management and Analytics

Published 13 December 2017 - ID G00338777 - 19 min read

By Analysts Ehtisham Zaidi, Guido De Simoni, Roxane Edjlali, Alan D. Duncan

Demand for data catalogs is soaring as organizations struggle to inventory distributed data assets to facilitate data monetization and conform to regulations. Data catalog projects will fall short of their full potential if data and analytics leaders don't link them to broader data management needs.

Overview

Impacts

- Data catalogs enable data and analytics leaders to introduce agile information governance, and to manage data sprawl and information supply chains, in support of digital business initiatives.
- Data catalogs are being adopted in an effort to remedy a lack of discipline in metadata management, but the profusion of vendor offerings and approaches, and the hype, create uncertainty about the value and selection of an optimal data catalog solution.
- Embracing data catalogs without an information governance plan that links them to the broader metadata management needs will leave data and analytics leaders unable to effectively manage and monetize their data assets in the longer term.

Recommendations

For data and analytics leaders evaluating data catalogs as part of a data management strategy:

- Use data catalogs to curate the inventory of available distributed information assets and to map information supply chains, by making them an essential component of your data management strategy.
- Proceed in the knowledge that tool-specific embedded data catalogs (such as those delivered as part of a Hadoop distribution, a cloud-based data lake or a data preparation tool) will improve data usability, trust and shareability *only* in the context of that particular tool.

- Deploy data catalogs with the capability to scale beyond narrow (or tactical) use-case requirements, in order to support strategic data and analytics needs, by ensuring they are connected to the broader enterprise metadata management and information governance programs.

Strategic Planning Assumptions

Through 2019, 80% of data lakes will not include effective metadata management capabilities, making them inefficient.

By 2019, data and analytics organizations that provide agile, curated internal and external datasets for a range of content authors will realize twice the business benefits of those that do not.

Analysis

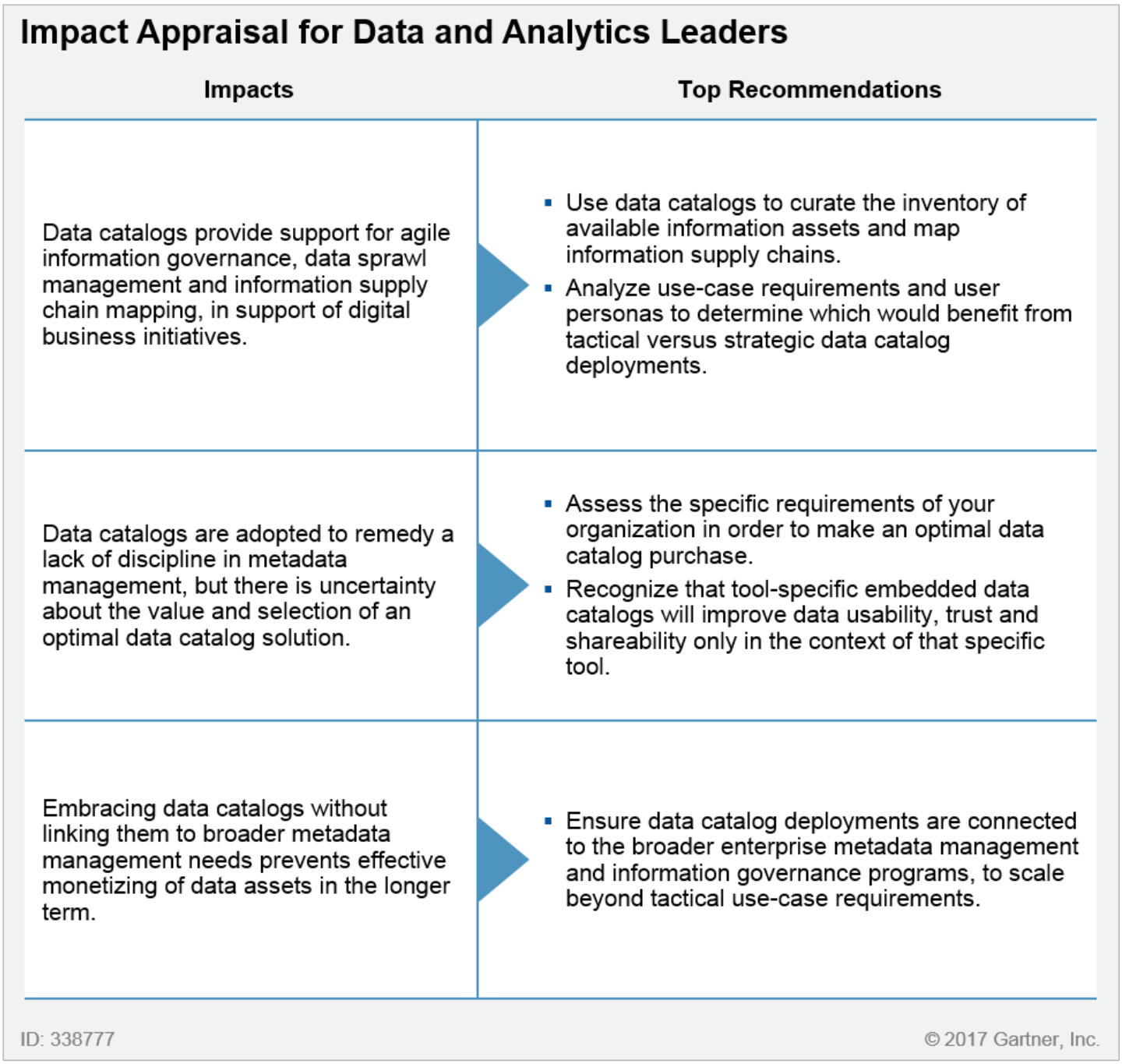
This document was revised on 14 March 2018. The document you are viewing is the corrected version. For more information, see the Corrections (https://www.gartner.com/technology/about/policies/current_corrections.jsp) page on gartner.com.

Interest in and adoption of a new generation of data catalogs is surging among Gartner clients. [1] The consequences of this trend affect data and analytics leaders in three important ways:

- First, data catalogs offer a fast and inexpensive way to inventory and classify the organization's increasingly distributed and disorganized data assets and map their information supply chains (see Note 1) to limit data sprawl.
- Second, they face a huge challenge in selecting an optimal data catalog solution amid the growing hype, and the availability of multiple vendor offerings.
- Third, without a comprehensive strategy linking data catalog initiatives to broader metadata management programs, they will struggle to manage and monetize data assets as well as complying with strict new regulatory and compliance rules.

This report examines the implications for data and analytics leaders and the steps they should take to exploit data catalogs effectively by linking them to broader enterprise information management needs.

Figure 1. Impacts and Top Recommendations for Data and Analytics Leaders



Source: Gartner (December 2017)

Impacts and Recommendations

Data Catalogs Support Digital Business via Inventory of Distributed Data Assets, Mapping Information Supply Chains and Managing Data Sprawl

Data and analytics leaders are scrambling to deploy some type of information governance model to address user-enabled data sprawl. Distributed data platforms — most notably data lakes — enable new use cases, but the data itself is often undocumented and ungoverned. Organizations are turning to a new generation of data catalogs, to develop inventories of their information assets and make them more accessible, usable and understandable. At the same time, they are trying to avoid the

mistakes of the past, when much more limited and tactical catalog solutions resulted in disorganized documentation efforts.

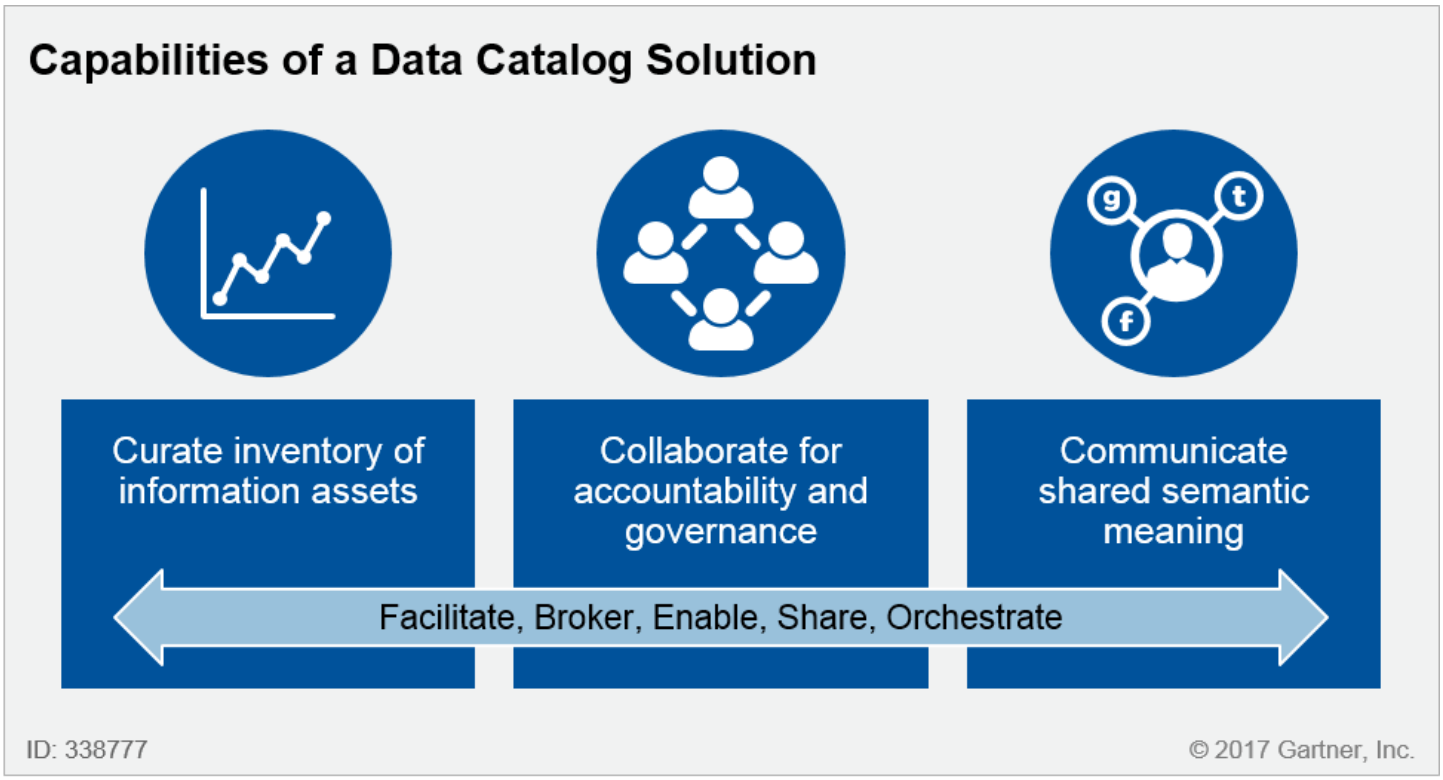
---

*A **data catalog** maintains an inventory of data assets through the discovery, description and organization of datasets. The catalog provides context to enable data analysts, data scientists, data stewards and other data consumers to find and understand a relevant dataset for the purpose of extracting business value.*

One particular use of data catalogs is to document and communicate the context, meaning and value of data that has been loaded to a data lake. Data lakes store data in an untransformed raw form, including data of immediate interest, potential interest and data for which the intended usage is not yet known. However, most data lakes quickly turn into "technology projects" that deliver limited value without an overarching metadata management strategy and the necessary disciplines to document and communicate the context and meaning of the data. Vast amounts of data that is expensively sourced, collected and stored, and rendered in its native formats, becomes untrustworthy and almost unusable (see "Metadata Is the Fish Finder in Data Lakes").

Data catalogs offer an ideal solution to this problem of ungoverned data propagation. In effect, they create a metadata repository that lets data consumers enrich, inventory, share and collaborate with the data across the enterprise. Data catalogs generally offer several core capabilities, as shown in Figure 2.

Figure 2. High-Level Summary of the Capabilities of a Data Catalog Solution



Source: Gartner (December 2017)

Data catalogs offer other short-term benefits in breaking down the barriers to operationalizing data assets and workloads in scenarios such as:

- Business-focused environments, where explicit emphasis on the importance of data is still in its infancy (see "Toolkit: Enabling Data Literacy and Information as a Second Language").
- Regulatory and compliance situations that are predicated on having well-documented understanding and traceability of the organization's information holdings (see "Why Privacy Is an Opportunity to Drive Data Value").
- Distributed parallel computing environments for analytics, such as data lake environments (see "Metadata Is the Fish Finder in Data Lakes").

Purely tactical catalog deployments can limit data and analytics leaders' ability to monetize data assets for digital business outcomes, and to satisfy the demands of new regulatory/compliance requirements such as General Data Protection Regulation (GDPR), Basel Committee for Banking Supervision 239 (BCBS239) and the Health Insurance Portability and Accountability Act (HIPAA). [2]

A recent Gartner Research Circle survey on data and analytics trends identified "data risk and information governance" and "deriving value from data" as the second and third most common challenges organizations face in their data and analytics programs (see Note 2).

A broader data catalog approach that leverages information governance and metadata management practices enables organizations to address these challenges and realize longer-term benefits, such

as:

- Collating and communicating the up-to-date information asset inventory that is available to the organization.
- Creating the common glossary of business terms that defines the semantic interpretation and meaning of the organization's data, thereby providing the means for mediating and resolving definitional inconsistencies.
- Enabling a dynamic and agile collaboration environment to enable business and IT colleagues to comment on, document and share data.
- Providing data usage transparency with lineage and impact analysis.
- Monitoring, auditing and tracing data in support of information governance processes.
- Capturing metadata to enhance internal analysis of data use and reuse, query optimization and data certification.
- Contextualizing information within its business usage by capturing, communicating and analyzing what data exists, where it comes from, what contexts it is used in, why it is needed, how it flows between processes and systems, who is accountable for it, what it means and what value it has. In effect, creating an information supply chain.

With these crucial capabilities, data catalogs (also known as information catalogs, see Note 3) can become a viable part of the solution to managing information as an organizational asset. This is the emerging business discipline of "infonomics" — managing and accounting for information with the same or similar rigor and formality as other traditional assets, such as financial or capital assets (see Note 4). To optimize information's ability to generate business value, data and analytics leaders must adopt an asset management mindset.

Most organizations fail to link their data catalog initiatives with their broader and more strategic enterprise information management (EIM) programs (such as data governance, data quality and metadata management). This link to the broader strategy is critical; it ensures that data catalogs break free from tactical or isolated "point" solutions, into more strategic organizationwide initiatives covering multiple analytics and operational use-case requirements (see "Magic Quadrant for Metadata Management Solutions" and "Reset Your Information Governance Approach by Moving From Truth to Trust").

Without a strategy that connects data catalogs with EIM initiatives, data and analytics leaders will end up with multiple data cataloging solutions trying to solve the same challenges for each project. This will lead to redundancy, governance chaos and an unnecessarily high total cost of ownership (TCO). In the long term, these problems will worsen. Due to multiple "point" data catalogs that fail to

scale, organizations could end up with redundant catalog solutions providing multiple versions of their metadata and possibly leading to governance chaos. The result is that, at exactly the moment when the organization as a whole needs to leverage data assets across the business, these tactical point data catalogs are unable to support that need.

Recommendations for data and analytics leaders:

- Use data catalogs to curate the inventory of available information assets and map your information supply chains.
- Partner with business users to establish the requirements for inventorying data assets in order to promote their use and value. Adopt infonomics practices to facilitate, mediate and champion the underlying business value of data.
- Analyze use-case requirements and user personas to identify which ones would benefit from tactical data catalog deployments in order to increase productivity and reduce time-to-market. For example, data lake use cases rather than those that need a more strategic implementation of the data catalog — such as the cataloging of data across on-premises and cloud data stores.

Growing Hype and Numerous Vendor Offerings/Use Cases Create Confusion in Data Catalog Selection

Gartner has seen a huge push from vendors in respect of data catalog solutions as a part of the metadata management market. Some are even rebranding versions of their metadata management tools as "data catalogs" to exploit customer interest. Self-service data preparation tools are also adding limited data catalog features (or partnering with stand-alone data catalogs) to ensure that their users can perform basic classification and inventorying of integrated data within their data preparation flows (see "Market Guide for Self-Service Data Preparation"). We are even witnessing cloud service providers delivering data catalogs as part of their cloud services.

Modern data catalogs allow users to assign different levels of trust (certification) to both raw and harmonized datasets, and can also allow certified users to rate integrated datasets for accuracy. Some modern data catalogs have the ability to track the activities of data consumers in order to understand actual data usage, what data is most important, which datasets are related and the nature of that relationship. They also track data lineage and incorporate data and query usage analytics and certification metrics. These capabilities help business users identify which integrated datasets have been certified, or not, for use by information stewards.

This increased intelligence makes the data catalog more effective in helping data consumers create queries and in associating specific consumers, or groups of them, with particular datasets. As a result, data catalogs can start meeting the needs of citizen data scientists, citizen integrators and business/data analysts — end users who are not IT coding experts or experienced data engineers. This opens the organization's data assets, and their effective use, to a much larger constituency;

effectively allowing organizations to take the first steps toward monetizing their information assets (see "Seven Steps to Monetizing Your Information Assets").

Data catalog implementations can vary depending on the use-case requirements, tool capabilities and maturity of each organization. There are also many vendor solutions and technology options through which companies can add data cataloging capabilities to their existing data management architecture. Among the most significant are:

- Broader metadata management solutions that embed data cataloging as a capability (see "Magic Quadrant for Metadata Management Solutions").
- Data preparation tools that either provide basic cataloging capabilities for their data preparation flows, or partner with other data catalogs (see "Market Guide for Self-Service Data Preparation").
- Data integration tools and data lake management solutions that embed data cataloging as a feature within a broader solution (see "Magic Quadrant for Data Integration Tools").
- Smart data discovery tools that embed basic data cataloging capabilities to aid information stewards and business analysts by automating the extraction, interpretation and contextualization of meaning from the application metadata of the originating data source.

The ultimate objective for data and analytics leaders should be to make the data catalog a cornerstone of curating information as an asset, by documenting and communicating the information supply chain — particularly, but not exclusively, for decision management.

Below, we provide a sample list of vendors who offer data cataloging capabilities either as features of a broader solution or as stand-alone tools themselves:

- Adaptive (http://www.adaptive.com/)
- Alation (https://alation.com/)
- Alex Solutions (http://alexsolutions.com.au/)
- Attivio (https://www.attivio.com/)
- Cambridge Semantics (https://www.cambridgesemantics.com/)
- Collibra (https://www.collibra.com/)
- Data Advantage Group (http://www.dag.com/)
- Datum (http://www.datumllc.com/)
- Global IDs (http://www.globalids.com/)

- IBM (https://www.ibm.com/)
- Infogix (http://www.infogix.com/)
- Informatica (https://www.informatica.com/)
- Teradata (Kylo) (https://kylo.io/)
- Linq (https://www.linq.it/)
- Microsoft Azure (https://azure.microsoft.com/)
- MooD International (http://www.moodinternational.com/)
- Oracle (http://www.oracle.com/)
- Podium (https://www.podiumdata.com/)
- SAP (https://go.sap.com/index.html)
- Smartlogic (https://www.smartlogic.com/)
- Unifi (https://unifisoftware.com/)
- Waterline Data (https://www.waterlinedata.com/)
- Zaloni (https://www.zaloni.com/)

The vendors of data catalog solutions listed above could belong to one of the three distinct subclasses into which each data catalog offering may be subdivided:

- Generalist, business-oriented data catalogs for broader use in information governance and infonomics — targeted at the chief data officer (CDO).
- Data catalogs that are specifically intended for inventory, curation and classification of the data provisioned in data lakes — targeted at data scientists and data engineers.
- Vendor-specific solutions that only provide this capability for the vendor's own environments — and are therefore targeted at the vendor's incumbent customers.

Recommendations for data and analytics leaders:

- Assess the specific data requirements of your organization to make the optimal decision when engaging a vendor about data catalogs.

- Evaluate the basic data catalog features of existing data management tools as a starting point for tactical requirements, such as inventorying multistructured data in data lakes.
- Ensure that your data catalog functionality requirements are balanced with other critical aspects such as vendor execution and vision, service and support, and your requirements for information security, information governance and TCO.
- Understand that tool specific-embedded data catalogs (for example, data catalogs delivered as part of a Hadoop distribution, a cloud-based data lake or a data preparation tool) will improve data usability, trust and shareability *only* in the context of that tool.
- Compare the embedded data cataloging capabilities of tools that support specific use cases (such as data lake management), with broader metadata management solutions (for more use cases) to avoid redundancy, a high TCO and information governance challenges (see "Magic Quadrant for Metadata Management Solutions" and "Toolkit: Sample RFI and Vendor Rating Spreadsheet for Evaluating Metadata Management Solutions").

Siloed Data Catalog Projects Will Limit the Inventory of Information Assets and Data Monetization

The tactical deployment of data catalog solutions aimed at specific use cases or projects makes data more usable for that specific purpose. However, it does little to enable the data to be treated and leveraged as an organizational asset which can be reused in digital initiatives that span multiple use cases, or business functions, across distributed data in the cloud or on-premises.

Instead of thinking in narrow terms — of "data catalog solutions" alone — data and analytics leaders should think about broader metadata management as a discipline of which data catalogs are one part of the enabling technology. Data catalogs offer capabilities that should fit into an organization's overall metadata management strategy, in order to create a business-oriented, shareable and reusable metadata repository in which common definitions of information assets are stored.

Documenting and communicating both business and technical metadata is the key to cataloging, identifying and evaluating the organization's information assets and how they are managed. However, data and analytics leaders evaluating data catalogs must ensure that their proposed solution has the capability to scale beyond tactical deployments and link them to the organization's broader metadata management needs and capabilities (see "Magic Quadrant for Metadata Management Solutions").

Modern metadata management solutions go beyond just data cataloging capabilities to also leverage and support metadata repositories, business glossaries, data lineage, impact analysis, rule management and metadata ingestion and translation. These more complex solutions offer the following benefits:

- Connecting operational and analytical uses of data.

- Enabling technical data analysts to become more productive by leveraging data catalogs for faster data lineage and traceability analysis, improved consistency and quality in root-cause analysis, and more accurate connection and correlation of datasets at the technical/physical level.
- Using data dictionaries to identify synergies between data used for different business initiatives (both data content and meaning), enabling greater data sharing and reuse, and improving data consistency.
- More effective understanding and communication of the semantic meaning of data.
- Intelligent decisions about the information life cycle — from data interoperability, to standards, to archiving, to disposal and deletion.
- Provision of information audit trails and other information-risk-related assurances, as required by regulators, legislation, business partners or customers.

Data and analytics leaders should focus on augmenting and accelerating the metadata-harvesting process and the semantic understanding derived by data analysts. To do so, they can exploit the rapidly emerging catalog features that are particularly suitable for analytics requirements. These features aid information stewards and business analysts in two ways:

- By automating the extraction, interpretation and contextualization of meaning from the application metadata of the originating data source.
- By using the underlying datasets themselves to infer the meaning of the data's content.

Recommendations for data and analytics leaders:

- Establish a metadata management practice within the organization to address and exploit metadata value and ensure that this practice links and connects the data catalog initiative with the overall metadata management program.
- Identify the top capabilities that your metadata management strategy will need, and evaluate the data catalog features needed to support those capabilities.
- Prioritize mission-critical scenarios that drive effective proofs of concept for metadata management solutions and data catalogs.
- Avoid data catalogs that do not have the ability to scale out beyond tactical-use-case requirements and connect to the broader enterprise metadata management and EIM initiatives, should your needs demand this.

Acronym Key and Glossary Terms

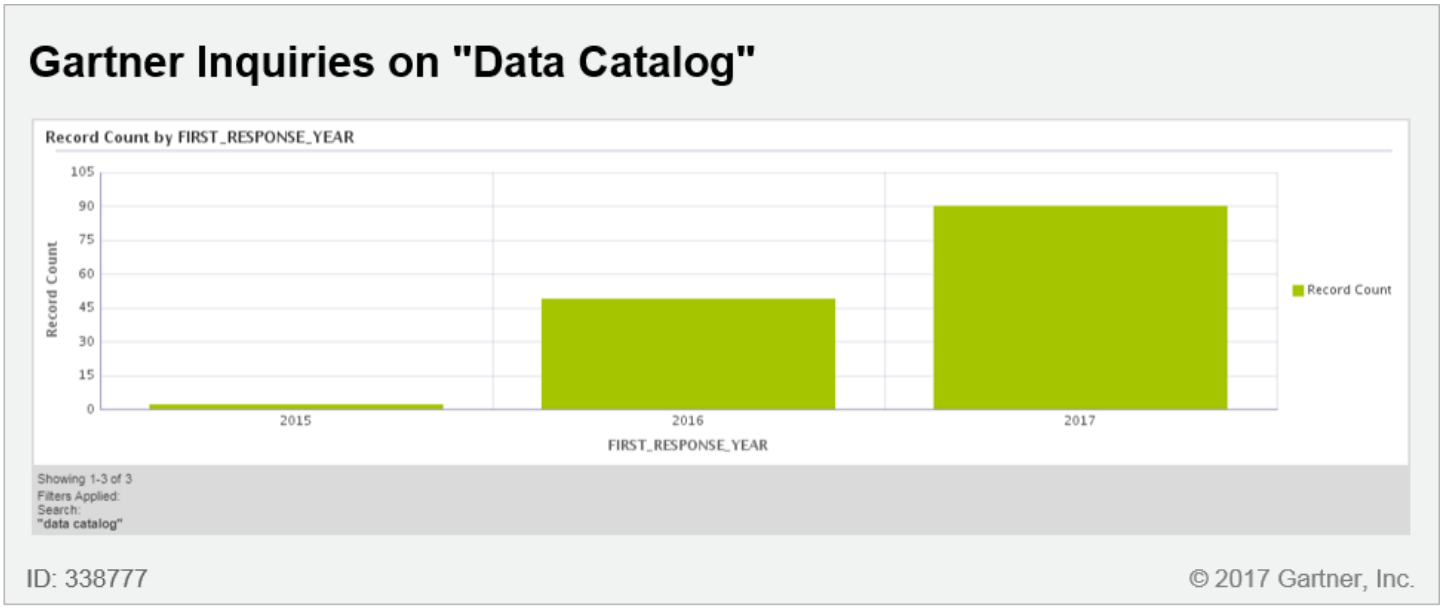
BCB239	Basel Committee on Banking Supervision regulation 239 (global)
EIM	enterprise information management
GDPR	General Data Protection Regulation (Europe)
HIPAA	Health Insurance Portability and Accountability Act (U.S.)
TCO	total cost of ownership

Evidence

[1] This report is based on information from a number of sources, including the following:

- Interactive briefings in which vendors provided Gartner with updates on their strategy, market positioning, recent key developments and product roadmaps.
- Feedback about tools and vendors captured during conversations with users of Gartner's client inquiry service.
- Client inquiries to Gartner's data and analytics team: We have taken 150 inquiries during the past two years specifically on the term "data catalog," with more than 10 inquiries per month on the subject. In 2017, we have (at the time of writing) taken 90 inquiries on this subject, which is a year-over-year increase of 50% compared with the same period in 2016 (see Figure 3).

Figure 3. Gartner Inquires on "Data Catalog" (2015-2017)



Source: Gartner (December 2017)

- According to Google Trends (https://trends.google.co.uk/trends/) , the term "data catalog" was trending nearly 50% higher, on average, than either "information catalog" or "metadata management solution" during a one-year period from late July 2016 to late July 2017.
- We are also seeing a huge push from the vendors; they are creating rebranded versions of their metadata management tools and calling them "data catalogs" in order to win the business.

[2] New data regulations and compliance requirements:

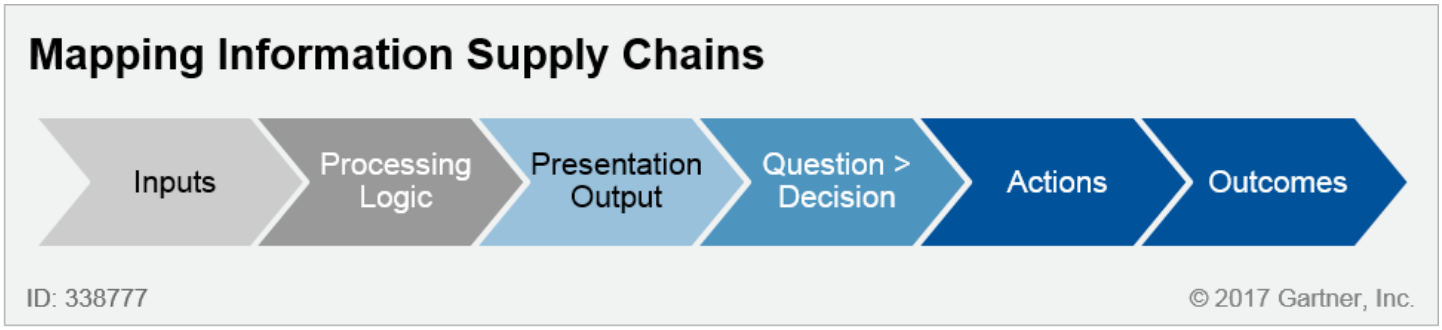
- The European GDPR, global BCBS239, and the U.S.'s HIPAA are three prominent examples of regulatory and compliance mandates for more rigorous governance of information assets. However, many jurisdictions and domains of business have some form of regulatory expectation for the responsible and transparent processing and storing of data. Where such legislative regimes are in place, data and analytics leaders should seek to work in partnership with the legal counsel of their organization in order to fully understand and respond to any implications that may arise from a data catalog initiative.

Note 1
Information Supply Chains

Physical supply chains are concerned with defining and optimizing the processes that source raw materials and turn them into finished goods, together with the resulting distribution and consumption of those goods to where they are needed.

Analogous disciplines, such as the of mapping information supply chains, can be used within data management and infonomics to more explicitly identify, map and manage the processes by which information is sourced, distributed and consumed by a business's operational and decision-making processes (see Figure 4).

Figure 4. Mapping Information Supply Chains



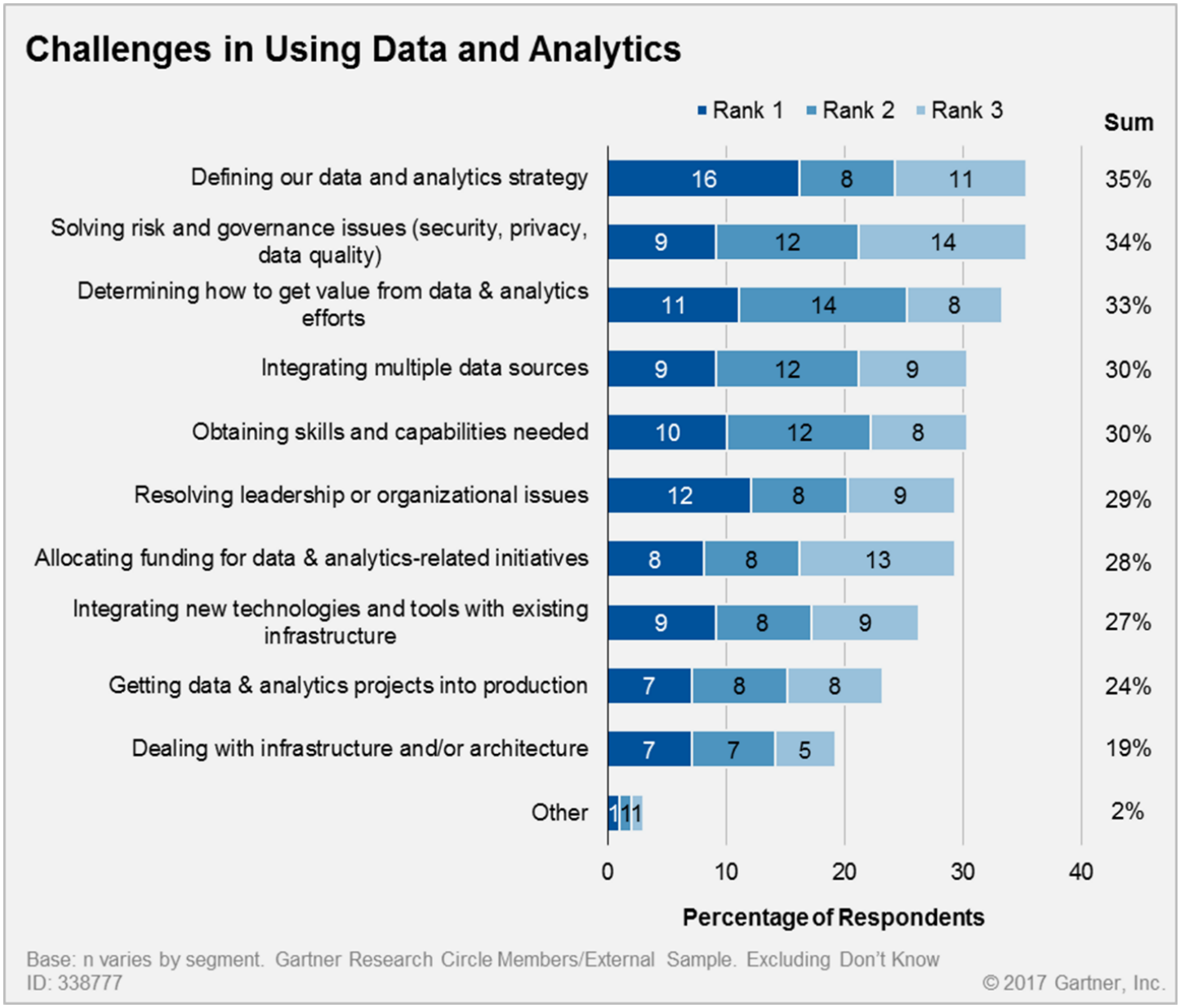
Source: Gartner (December 2017)

By mapping the information supply chain, the utility, value of and accountability for data is more explicitly communicated. Also, information governance requirements are identified and opportunities highlighted for improving business outcomes and delivering new organizational value from better investment, treatment and curation of the data.

Note 2
Top Data and Analytics Challenges

The 2017 Gartner Research Circle Data and Analytics Trends Survey was conducted online from 23 May 2017 to 26 June 2017, among members of the Gartner Research Circle — a Gartner-managed panel of IT and business leaders — as well as an external sample source. In total, 196 respondents completed the survey. It was developed collaboratively by a team of Gartner analysts and was reviewed, tested and administered by Gartner's Research Data Analytics team. Figure 5 shows the top data and analytics challenges revealed by the survey.

Figure 1. Impacts and Top Recommendations for Data and Analytics Leaders



Source: Gartner (December 2017)

Note 3
"Information Catalog" or "Data Catalog"?

We observe these terms being used interchangeably, and they can be treated as being synonymous. Business-oriented roles more typically refer to an "information catalog," whereas the more technically oriented roles may be more likely to refer to a "data catalog."

Note 4
Infonomics

Infonomics is the emerging discipline of managing and accounting for information with the same or similar rigor and formality as is applied to other traditional assets, such as financial, physical, intangible or human assets.

Infonomics proposes that information meets all the criteria which formal company assets meet. Although infonomics is not yet recognized by generally accepted accounting principles (GAAP), it is increasingly incumbent on organizations to behave as if it were, in order to optimize information's ability to generate business value.