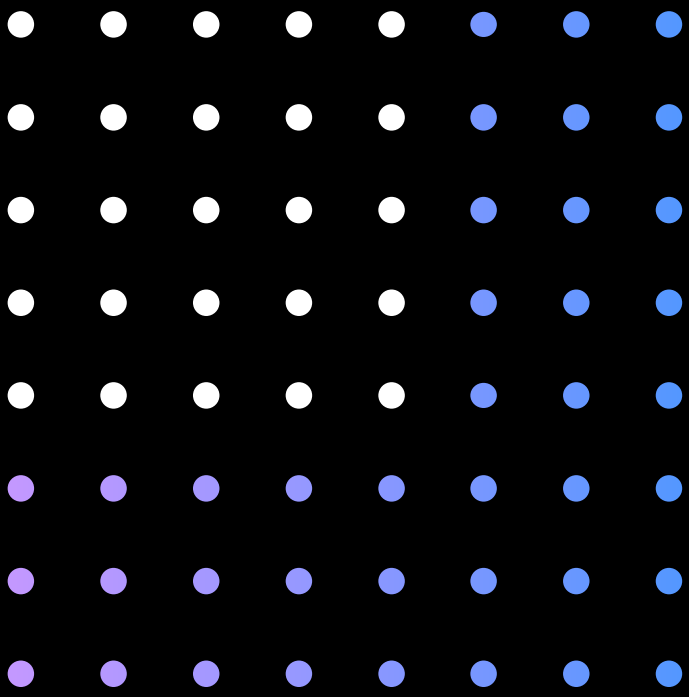
Deliver business-ready data with intelligent data cataloging and data lake governance

IBM Watson Knowledge Catalog provides a machine learning-powered data governance platform to help with data lake challenges



IBM

Contents

Key takeaways

- Few organizations are seeing the value they expected from the data lakes they have built to store and analyze their data for trusted insights.
- DataOps solve the challenges organizations face with inefficiencies in accessing, preparing, integrating and making data available to consumers while adhering to corporate and regulatory policies.
- Common data lake challenges include the difficulty and cost of importing new data sources into the data lake; an inability to integrate internal and external data sets; a lack of confidence around data governance; no access to self-service data preparation tools; and an inability to find and understand the data that's in the data lake.
- An enterprise data governance platform with cataloging, data quality and data discovery can transform a failing data lake project into a true source of business value.
- IBM® Watson® Knowledge Catalog, powered by IBM Cloud Pak™ for Data, provides a machine learning (ML) catalog for data discovery, data cataloging, data quality and governance. It helps data users quickly discover, curate, categorize and share data assets, data sets and analytical models.
- When organizations lack a deep understanding of their data, it becomes more difficult to trust and use this information with all forms of artificial intelligence (AI), including ML and deep learning.

Solving data lake challenges with a DataOps approach

Ten years ago, the journey began to find a flexible, versatile approach to build a central data store where all enterprise data could reside. The solution was the data lake—a general-purpose data storage environment that would store practically any type of data. It would also allow business analysts and data scientists to apply the most appropriate analytics engines and tools to each data set, in its original location.

Typically, these data lakes were built using Apache Hadoop and Hadoop Distributed File System (HDFS), combined with engines such as Apache Hive and Apache Spark. As these data lakes began to grow, a set of problems became apparent. While the technology was physically capable of scaling to capture, store and analyze vast and varied collections of structured and unstructured data, too little attention was paid to the practicalities of how to embed these capabilities into business workflows.

Through 2022, over 80% of data lake projects will fail to deliver value as finding, inventorying and curating data will prove to be the biggest inhibitor to analytics and data science success.[1] As a result, questions such as: "What data should we put in the data lake?", "Who is going to use it?", "How do we make it easy for them to find?", "Where did this data come from?" And "How do we prevent data from being misused?", often went unanswered. These critical limitations in addressing people, process and technology issues effectively led to unsuccessful data lake implementations.

Today, many organizations have recognized their failures, have changed leadership teams for the data lake implementation and are launching a second, third or even fourth attempt to implement a data lake successfully–this time leading with data operations DataOps.

DataOps is a collaborative data management practice focused on improving the communication, integration, and automation of data flows between data managers and data consumers across an organization.

Introducing DataOps

DataOps brings best practices from DevOps, data management and data governance into a common framework, with a collaborative way of developing and maintaining data flows across multiple stakeholders. DataOps is designed to solve challenges organizations face that are associated with inefficiencies in accessing, preparing, integrating and making data available to consumers while adhering to corporate and regulatory policies. These efficiencies can be found in a business unit, an analytics team or even an operational process.

Following this methodology requires addressing the people, process and technology issues that mean the difference between successful and unsuccessful data lake implementations. From a technology side, DataOps stresses the importance of using a fully integrated, end-to-end platform for data ingestion and integration, data quality, data governance and data consumption to create a governed data lake. Data quality validation rules should run automatically as part of the ingestion process to sustain a continuous data pipeline across the enterprise. The ingestion process should be fully integrated with the data catalog, which becomes the heart of your pipeline. Data consumers should be able to access data quality scores and data profiling results from the data catalog and trust that the organization is working with the same data in context.

The growth of data is outpacing organizations' ability to get value from it. When organizations were asked what are the biggest challenges for using systems of insight they responded: 1) 40% is merging existing business processes to source data to analyze it and 2) 39% is sourcing, gathering, managing, and governing the data as it grows.[2] Today, it's not just a case of protecting the huge time and resource investments that have already been made in data lake technologies—it's the fact that there is no alternative. From implementing AI or even to perform comprehensive analysis, it's vital to have a full view of as much data as possible, which means you need an architecture that's capable of holding and analyzing and governing all that data in one place. In many cases, a governed data lake is the only realistic option for meeting these requirements.

Today's businesses can—and must—find a way to extract value from their data lake by ensuring it supports a business-ready data pipeline for DataOps.

Challenges to using enterprise data lakes

Sharing data

When a team within an enterprise acquires or creates a new data set, it's likely to have a strong sense of the data's value, and the sensitivities surrounding it. If it contains commercially confidential information, personally identifiable information (PII) or customer data, for example, the team will know how that information should and shouldn't be used, and will take precautions to make sure nobody in the team misuses it.

They will also be conscious that outside of their team, other potential users of the data may not share the same understanding of the value of the data, or the risks associated with misusing it. These risks will naturally make the team extremely cautious about sharing the data or storing it anywhere that's not under their control.

This is bad news for data lakes. If the business sees the data lake as simply an uncontrolled dumping ground for data, they will be very reluctant to entrust their valuable data to it. As a result, other parts of the business won't be able to benefit from that data, and the whole concept of using the data lake as a self-service repository for sharing enterprise data falls apart.

Integrating data
Even when a team does agree to its data being integrated into the data lake, it can be a torturous process. The original concept of the data lake is to import data in its raw format, without requiring the complex extract, transform and load (ETL) processes of a traditional data warehouse. However, the reality is that almost all data sources require some degree of preprocessing before they can be useful for any kind of meaningful analysis.

As a result, integrating a new source of data into a data lake can often take months. And because much of this data has previously been held in small operational silos rather than enterprise systems, there may be dozens or even hundreds of sources to integrate in total.

This means that in many cases, the information that business analysts or data scientists need has not yet been added to the data lake, and may not be added for months or even years. Again, this can be a significant barrier to adoption.

Storing data
While the cost of commodity storage and compute resources has decreased dramatically over the past few years, Hadoop clusters aren't free. Storing massive quantities of data in a data lake is much more cost-effective than storing it in a high-performance data warehouse appliance, but the cost can still be significant.

Moreover, unlike data that's traditionally stored in data warehouses the value-to-volume ratio of the big data held in a data lake is comparatively low. You may need to store a very large haystack to locate the handful of high-value needles.

Finding data
Assuming you have identified the most valuable data sets to store, persuaded your stakeholders to share them and succeeded in integrating them into your data lake, you still need to make it possible for other users to find, understand and use them properly. The quality of data in the data lake is yet another challenge. You're not sure if the data is of high or low quality, but it's being fed into the lake.

Challenges to using enterprise data lakes

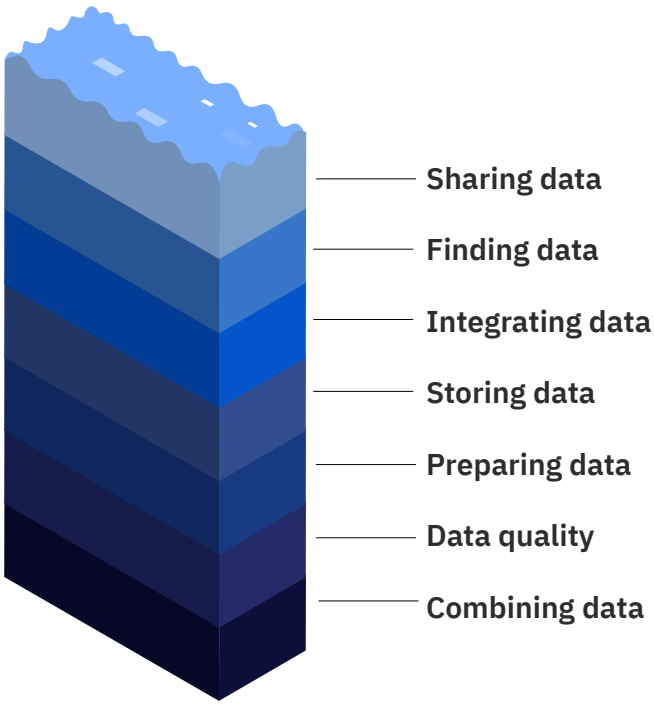


Figure 1. Enterprises that have adopted data lake technologies may encounter one or more of these common issues.

Unfortunately, in most data lakes, this isn't easy to achieve. Data is often stored without any context, making it difficult or impossible for a new user to decode it without consulting the original owner. Terminology is often so domain-specific that a metric used in one area of the business might be known by a completely different name—or defined in a subtly different way—by another. The potential for confusion and misinterpretation may be so large that many data sets are effectively worthless, or even dangerous, for an analyst who's not already familiar with them.

Combining internal and external data
Finally, even the largest data lake should not attempt to hold every possible data set that a company's data scientists will want to use. For example, it wouldn't make sense to import a complete replica of Google Maps, Weather.com or Bloomberg in your data lake, just because one of your data scientists wants to perform geospatial analysis, or integrate weather data or stock prices into an algorithm.

Because your data lake won't hold all the data your business analysts need for analytics, they'll have to spend time searching for it in multiple applications. Since a very large proportion of useful analyses are likely to involve the combination of internal and external data sets, this once again raises the barrier to entry and from the user's perspective, reduces the perceived value of the data lake.

Preparing data
There are many factors that make data preparation challenging—from understanding where to find the data to then formatting it. Preparing data for use in analytics is the most inefficient and time-consuming task for data users. Data users spend most of their time finding, cleaning and formatting information, instead of focusing on data analysis,modeling and deriving insights for business impact.

Limited accessibility to governed data sets has also caused an over reliance on IT during the preparation phase. This limited access signals the need to improve self-service capabilities and data literacy skills across the enterprise to alleviate this roadblock.

Data quality
Dumping data into a data lake can make it unusable. As there are no data quality or validation rules being applied to data before it's fed into a data lake, it does not provide data that can be trusted and used. High-quality data is an essential characteristic that determines the reliability of data for making decisions. Data is a valuable asset that must be managed as it moves through an organization. As information sources are growing more numerous and diverse, and regulatory compliance initiatives more focused, the need to integrate and access information from these disparate sources in consistent, trusted and reusable ways is critical.

A holistic approach to building governed data lakes

Most data lakes leverage Apache Hadoop and its broad ecosystem of open source projects for their data storage layers and analytics engines. Unsurprisingly, the open source community around Hadoop has recognized the problems that face current data lake implementations, and many projects have recently sprung up that aim to address the various problems individually. Similarly, there are a number of proprietary tools on the market that purport to solve the same issues.

It may be tempting, then, to remediate the problems of your data lake piecemeal, as they arise. When the number of data sets rises too high to be manageable, add a cataloging tool. When users complain that they can't find the data they need, bolt-on a front-end with a search function. And when your data stewards can no longer keep track of where your data came from or who's using it, deploy data lineage tools and a data governance framework.

It sounds simple, but in practice, this piecemeal approach tends to come at the cost of massively increasing complexity and reducing maintainability, especially as the scale and scope of the data lake increase. In the same way that adding new data sources to a data lake increases the complexity of your ETL requirements, the addition of new tools tends to increase the complexity of the data lake's non-functional requirements.

Instead of having an integrated end-to-end platform that can integrate data, perform quality operations on your data and catalog your data for effective use by your business analysts, you will typically find that each tool has its own ways of managing failure, and its own approach to logging. As a result, troubleshooting and problem resolution can become highly time-consuming.

Another, more important shortcoming of the piecemeal approach becomes apparent when you take a less technical, more conceptual view of the problems that data lakes commonly experience. The key insight is that scalability, findability, integration, data quality and governance are not separate problems: they're inextricably interrelated. Solving them will require a much more holistic approach.

Scalability, findability, integration, data quality and governance are not separate problems: they're inextricably interrelated. Solving them will require a holistic approach to information management.

IBM Watson Knowledge Catalog Data discovery, data cataloging and data quality

The IBM Watson Knowledge Catalog powered by IBM Cloud Pak for Data helps data users quickly discover, curate, categorize and share data assets, data sets, analytical models and their relationships with other members of the organization. It helps data governance teams to define business glossary, policies and rules and provides advanced workflows for governance. The catalog serves as a single source of truth for data engineers, data stewards, data scientists and business analysts to gain self-service access to data they can trust and use with confidence.

Solutions such as IBM Watson Knowledge Catalog powered by IBM Cloud Pak for Data can provide all the capabilities required to solve the major problems of today's data lakes in a single, comprehensive platform. The catalog helps address the root cause of these interrelated problems: the widespread failure of data lakes to provide effective tools to capture, store and manage metadata, and track data lineage.

In many ways, the value of a data lake depends on the metadata it contains, just as much as it depends on the data itself. Without metadata to explain where a data set came from, who created it, what it contains, who's permitted to use it, and how it's being used, the data itself is practically worthless. Users won't be able to find it, and even if they do, they'll not understand what it means or trust it with confidence or know how they can use it.

Watson Knowledge Catalog
Delivering trusted and meaningful data

Organize your data

Know

Data must be complete, applicable and accessible everywhere. Discover, classify and understand all types of data.

Govern your data

Trust

Data must be secure, clean and easy to find to encourage trusted self-service access. Understand where the data came from and its quality.

Democratize your data

Consume

Ability to drive self-service discovery and automate decision making to evolve the business. Provide a view of all information to those that need it and allow them to access it.

Figure 2. IBM Watson Knowledge Catalog provides a broad range of capabilities for data discovery, data cataloging and data governance.

A single source of truth and single point of access

IBM Watson Knowledge Catalog powered by IBM Cloud Pak for Data addresses these issues by making metadata a key priority. At its heart is a powerful cataloging engine that indexes all the data sets and analytic assets your business has access to, regardless of where the data resides such as in your data lake, data warehouse or transactional system, or even in a set of spreadsheets. Regardless of whether they're structured or unstructured or stored on premises or hosted in the cloud. Furthermore, the catalog can also include external data sets and sources, such as proprietary data services that your company subscribes to, or open data APIs.

As well as providing a single source of truth about all your data sets, the data catalog also provides a single point of access. AI-powered search and suggest capabilities help business analysts, data scientists, data quality engineers and data governance teams find assets more easily, and present the available metadata to help users understand what they've found and assessed whether it is useful to them.

Embedded, self-service data preparation capabilities accelerate the time it takes to transform data for productive use in analytics and AI applications. Business analysts and data scientists don't have to waste time in preparing and analyzing the data. Integration with an enterprise-wide data preparation solution, like IBM InfoSphere® Advanced Data Preparation helps ensure that the governed data sets created through the catalog surface to those with the most context to drive business insights and actions for business users. This integration furthers collaboration across the data pipeline.

> Scalability, findability, integration, data quality and governance are not separate problems: they're inextricably interrelated. Solving them will require a holistic approach to information management.

The catalog also helps data stewards in the chief data officer's (CDO's) office by tagging and classifying data sets and automatically tracing their lineage and usage, and by leveraging the built-in business glossary to standardize business terminology across the data. As a result, it's easier for stewards to understand what each data set contains, where the sensitive or PII is, and who should be allowed to access it.

A single catalog for multiple data sources inside and outside the organization

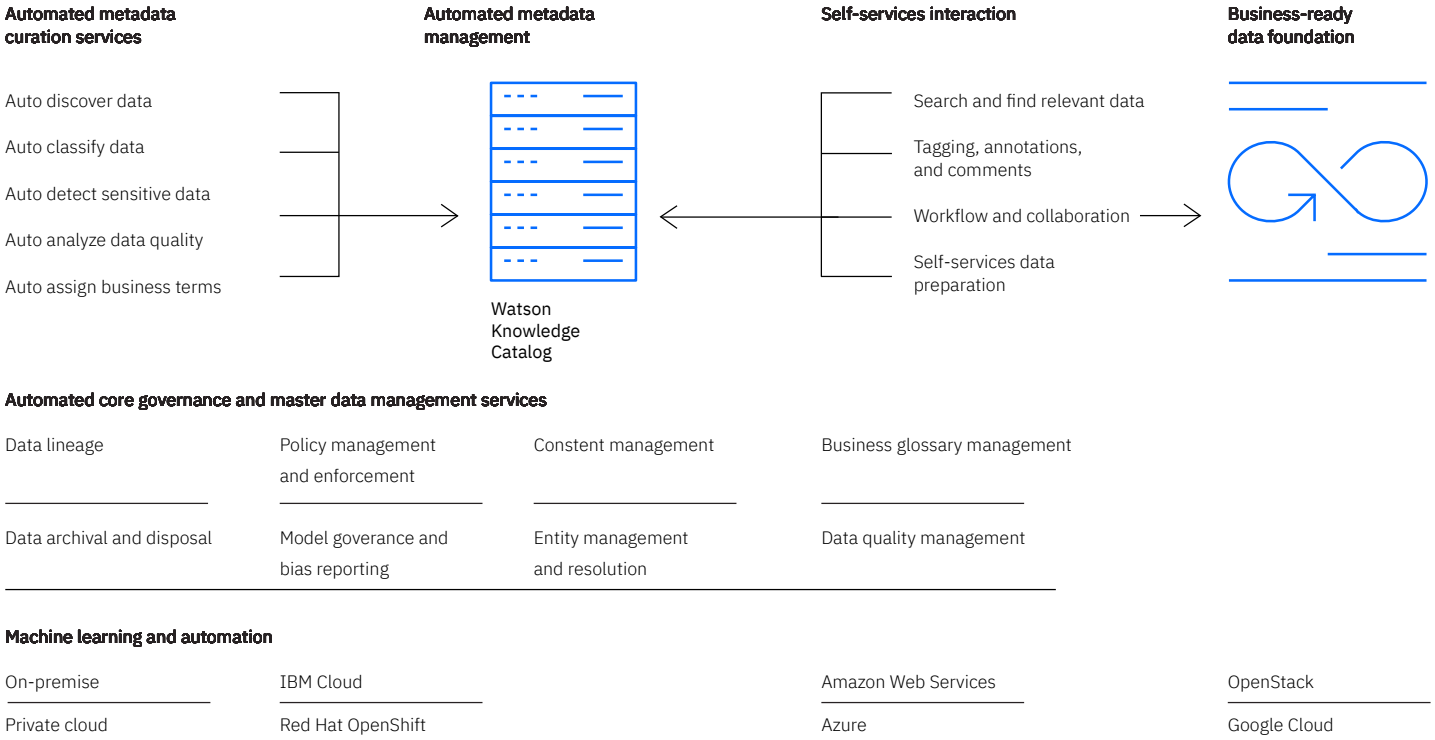


Figure 3. With IBM Watson Knowledge Catalog intelligent metadata index, data—both structured and unstructured—can reside in original systems, but users can discover it quickly for smarter analytics.

IBM Watson Knowledge Catalog makes metadata a key priority, providing a single source of truth and a single point of access to all the data sets your business has access

Built-in intelligent data discovery

To improve findability even further, the catalog allows users to tag and comment on data sets and analytic assets, enriching the metadata and adding extra context to help coworkers find what they need. The solution also includes built-in data discovery algorithms that use ML to auto classify the contents of each data set. By identifying common field types such as names, addresses, zip codes and social security numbers, the solution reduces the need for authors to annotate the data manually. It infuses automation and ML to automate data curation and metadata management. With built-in data quality functions, the solution enables deep data profiling, data quality and validation rules.

Automated data operations provides a curated data pipeline, with data quality and governance and helps ensure that there's a continuous flow of high-quality governed data into the data lake.

In a similar way, the addition of an intelligent metadata model of your assets provides a unique way to automatically enforce like the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA).

IBM Watson Knowledge Catalog powered by IBM Cloud Pak for Data helps to deliver trusted, high-quality, business-ready data to essentially all data users.

All the components of the solution have been engineered as microservices, with a single set of design principles and a common approach to non-functional requirements, such as scalability, error management, security and logging.

IBM Watson Knowledge Catalog provides a ML enterprise governance platform—so it's ready for AI at scale.

Instead of the confusing errors and performance bottlenecks that are likely to result from a piecemeal, do-it-yourself approach, IBM Watson Knowledge Catalog provides a ML enterprise governance platform, so it's ready for AI at scale.

IBM Watson Knowledge Catalog is available in three variants:

- As a software as a service (SaaS) solution in the IBM Cloud™
- In IBM Cloud Pak for Data
- Integrated with IBM Watson Studio

Solutions like the IBM Watson Knowledge Catalog can unlock the value that data lake initiatives originally promised. Watson Knowledge Catalog with intelligent cataloging and governance capabilities helps build a trusted and governed data lake for AI.

Four benefits of building a governed data lake for AI

1. Build trust and confidence in data through quality and governance

- Data quality capabilities help you to improve the quality of your data and make high-quality data available in your data lake.
- Governance policies are automatically set and enforced—so when you find a data set, you know whether and how you are allowed to use it.
- You can curate your data as users add ratings, comments and other information that will help others determine whether or not a data set will be useful to them.

2. Empowers your data users

- Your line-of-business (LOB) teams share their data willingly because they are confident that it will be properly governed and protected from misuse.
- You can drive collaboration and transform data into trusted enterprise assets through dynamic data policies and enforcement.
- Your data gets more findable and reusable over time, as users add relevant tags and metadata to help others extract value from it.
- A single interface gives you access to every data set your organization owns, regardless of where it's stored.

3. Get your time back

- Automatic data discovery reduces the time and effort you need to spend adding metadata for new data sets.
- Automatic data curation and metadata management reduce the time it takes to discover metadata and assign terms and also reduces the business glossary creation time.
- With simple and intuitive self-service data preparation tools, your data users spend less time preparing data and more time discovering insights.
- You unleash your data scientists and your business analysts to provide better analytics in a shorter space of time.
- Intelligent, AI-powered search helps you find the data you need within seconds, instead of waiting weeks for another team to provide it.

4. Manage growing data and cost

- You can optimize storage costs by avoiding the expense of ingesting low-value data sets into the data lake.
- You can also see all the external data sets that your organization subscribes to, reducing the risk of paying for more subscriptions than you need.
- You can prioritize the ingestion of new data sources into the data lake based on users' demand for the data, helping you integrate the most valuable sources first.

Unlock the value of your data

Whether you work in the CDO's office, in the IT department or as a LOB data scientist or analyst, you and your colleagues share a common goal. If you can build a data lake that really delivers on its promises, you could not only make your own jobs much easier and more productive. Additionally, you could play a key role in giving your business a competitive edge that few organizations can currently rival.

If you can clean the waters of your data lake while your competitors are still floundering in the swamp, you will open up possibilities that they can only dream of. Genuine first-mover advantage awaits those who are the first to unlock the value of previously untapped data.

Conclusion

Know where all your data resides, who's using it, and its value to your business for analytics.

Critical to DataOps initiatives are data catalogs because they can help deliver automated open metadata management by integrating data governance, quality and active policy management.

IBM Watson Knowledge Catalog with intelligent cataloging and governance capabilities helps build a trusted and governed data lake for AI. The catalog embeds data integration, data quality and governance into your data lake environment to help deliver business-ready data for DataOps—and a single source of truth.

For more information

To learn more, visit:
ibm.com/cloud/watson-knowledge-catalog

IBM Corporation
New Orchard Road
Armonk,NY 10504

Produced in the United States of America, October 2019 IBM, the IBM logo, ibm.com, IBM Cloud, IBM Cloud Pak, IBM Watson and InfoSphere are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide.

1.Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders–Gartner, Sept 2019
2. The Forrester Wave: Machine Learning Data Catalogs, Q2 2018