# How to Avoid Data Lake Failures

FOUNDATIONAL **Refreshed:** 16 December 2019 | **Published:** 10 August 2018 **ID:** G00367848

**Analyst(s):** Nick Heudecker, Adam Ronthal

Business and IT leaders are overestimating the effectiveness and usefulness of data lakes in their data and analytics strategies. Data and analytics leaders can avoid data lake failures by comparing their skills, expectations and infrastructure capabilities with the scenarios in this report.

FOUNDATIONAL DOCUMENT

This research is reviewed periodically for accuracy. Last reviewed on **16 December 2019**.

## Key Findings

- Proponents of data lakes often exaggerate their benefits by promoting them as enterprisewide solutions to all data and analytics problems. But it's frequently impossible to meet their inflated expectations in a reasonable time frame.
- Data lakes are rarely started with a definite goal in mind, but rather with nebulous aspirations to "create a single version of the truth" or to "democratize our data." These are neither strategic nor tactically informative goals.
- Many data and analytics leaders have firsthand experience or anecdotal evidence of data warehouse implementation failures, but few have experience of data lake failures. This lack of experience has made many prospective implementers overconfident.

## Recommendations

Data and analytics leaders expanding their data management strategies with data lakes should:

- Improve their chances of success by building data lakes for the specific requirements of certain business groups, sets of users or analytics use cases, rather than taking a "big bang," enterprisewide approach.
- Avoid confusing a data lake implementation with a data and analytics strategy. A data lake is just infrastructure, not a substitute for a strategy encompassing objectives, stakeholders, outcomes, metrics and risks.

- Use a data lake implementation project as a way to introduce or reinvigorate a data management program by positioning data management capabilities as a prerequisite for a successful data lake.

## Table of Contents

Analysis....2
Failure Scenario 1: “Enterprise Data Lake”....4
Governance Challenges....4
Semantic Consistency Challenges....5
Performance and Flexibility Challenges....5
Political and Cultural Challenges....5
How to Avoid This Failure Scenario....5
Failure Scenario 2: “Data Lake Is My Data and Analytics Strategy”....6
Mistaken Attempts to Replace Strategy Development With Infrastructure....6
Lack of Organizational Clout or Social Capital....7
Underestimation of the Immaturity of Data Management Capabilities....7
Misunderstanding of the Diverse Requirements of a Data and Analytics Platform for Digital Business....8
How to Avoid This Failure Scenario....10
Failure Scenario 3: “Infinite Data Lake”....10
Outdated or Irrelevant Data....10
Continuation of Immature Data Life Cycle Management Capabilities....11
Eventual Performance and Cost Challenges....11
How to Avoid This Failure Scenario....11
Gartner Recommended Reading....12

## List of Figures

Figure 1. Factors Contributing to Three Data Lake Failure Scenarios....4
Figure 2. Digital Business Technology Platform — Conceptual Version....9

# Analysis

Enterprises are implementing, or planning to implement, data lakes for a variety of reasons. The most common is to provide data to information consumers more quickly, thus enabling them to

bypass data warehousing and data mart environments. Other reasons include the desire to create experimental environments for data scientists or to replace existing data warehousing environments.

Most clients who contact Gartner to discuss this topic believe that a data lake's characteristics of flexible data storage and diverse processing options will simplify the data management tasks of managing governance, data life cycles, data quality and metadata. The popular view is that a data lake will be the one destination for all the data in their enterprise and the optimal platform for all their analytics. This view rests on three assumptions that have not proved correct:

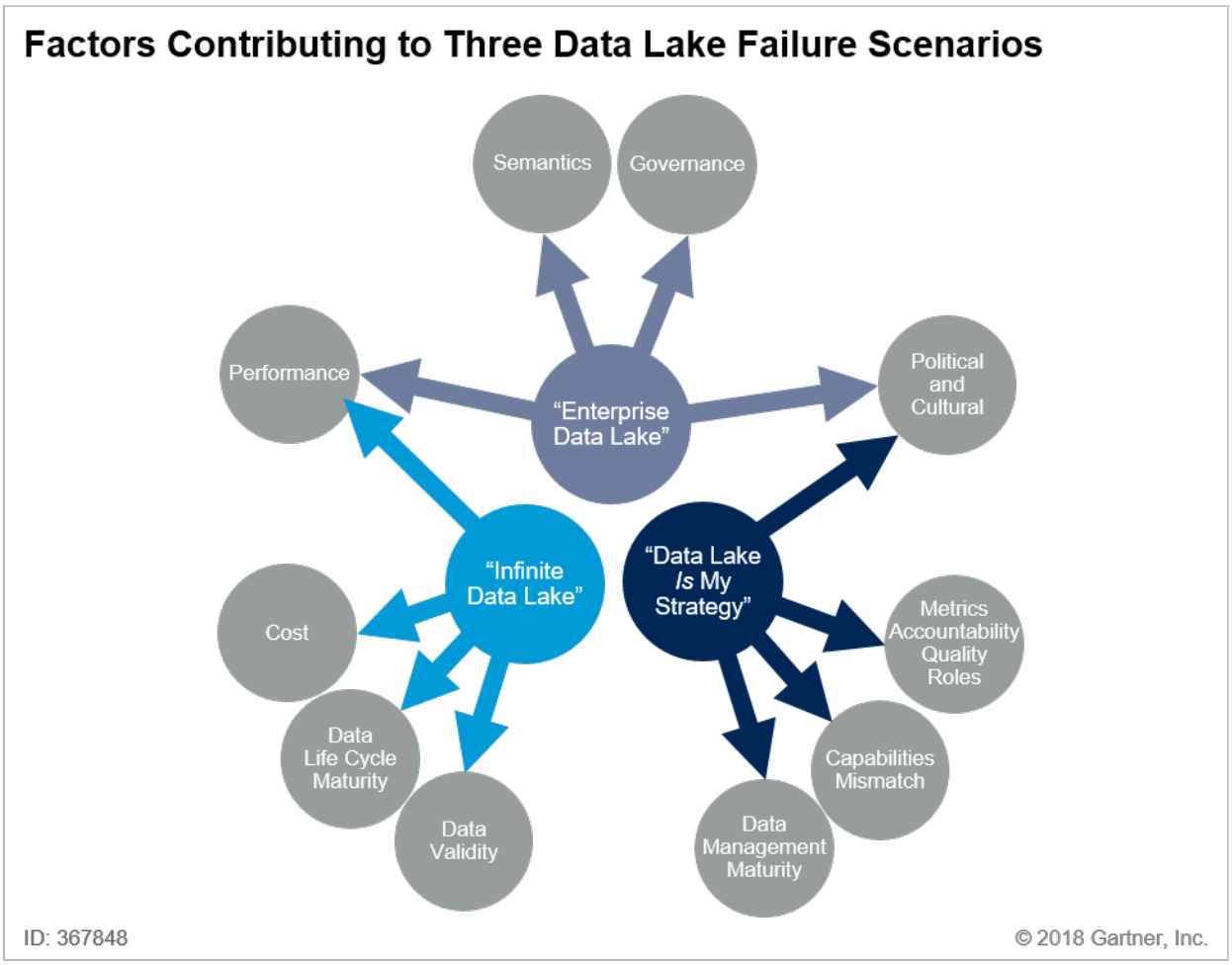
- The first is that everyone in the enterprise is data-literate enough to derive value from large amounts of raw or uncurated data. The reality is that only a handful of staff are skilled enough to cope with such data, and they are likely doing so already.
- The second is that the enterprise will be able to define cohesive governance and security policies across all datasets residing on a single cluster of physical infrastructure. The same attempt was made with data warehouse implementations, but proves far less successful with data lakes because the data they contain isn't modeled. Creating policies for data without context is impossible.
- The third is that data lake implementation technologies perform far better than they actually do, which leads to wild overestimations of their benefits.

Nevertheless, organizations are pressing ahead enthusiastically with data lake implementations. The lack of documented data lake failures, whether personally experienced or communicated in the popular press, has convinced many organizations that data lakes are a magical answer to their data and analytics requirements. Many of these organizations will likely fail. They just haven't failed *yet*.

To improve your chances of success, this report identifies the most common data lake failure scenarios that Gartner has identified from conversations with clients — and advises how to avoid them.

Figure 1 summarizes factors that contribute to three common scenarios for data lake failure.

Figure 1. Factors Contributing to Three Data Lake Failure Scenarios




Source: Gartner (August 2018)

## Failure Scenario 1: "Enterprise Data Lake"

Enterprises implementing an enterprise data lake aim to unify multiple data silos into a single piece of physical infrastructure. The intention is to readily provide all data to different groups throughout the organization and to centralize data access for analytics. Phrases like "single version of the truth" are frequently used by proponents of such lakes.

This scenario typically fails because of a variety of governance, performance and organizational challenges.

### Governance Challenges

The most acute challenge is to resolve diverse data governance requirements in a single platform. Governance capabilities must support local application data, as well as regional and global data

assets. But data lake implementation technologies, such as Hadoop and cloud-based object stores, aren't optimized for these different levels of granularity.

Technological limitations aren't the only governance challenge. Data stored in a data lake intentionally lacks the context enforced by various modeling and integration processes. After all, the point is to provide data faster, and those processes slow things down. Without the necessary context, it's impossible to reconcile and enforce a uniform governance model. This brings us to the second reason why this scenario fails: semantic consistency.

## Semantic Consistency Challenges

A major driving force behind enterprise data lakes is the belief of IT and business leaders that simply putting data in the same place removes any ambiguity about what that data means. This is despite data originating from potentially dozens of systems with no shared data or metadata model — customer data, for example, may derive from CRM, ERP and marketing databases. Data lakes are not optimized for semantic enforcement or consistency. Actually, the opposite is true: They are optimized for semantic flexibility, and allow anyone to introduce context to data if they have the skills to do so.

## Performance and Flexibility Challenges

Data lake infrastructure is optimized to store and process large amounts of data, frequently organized as massive files and processed in batches. It's not optimized for high numbers of concurrent users or highly diverse concurrent workloads. Performance degradation and failures are common in this scenario, such as when running extraction, transformation and loading tasks concurrently with training machine learning models and running interactive queries.

On-premises data lakes face another performance challenge in that they effectively have a static configuration. Introducing additional memory or storage requires long lead times to acquire hardware, provide power, networking and monitoring, and eventually to install software.

## Political and Cultural Challenges

The data silos that this scenario promises to eliminate are a reflection of the organization, not a limitation of technology. Organizational politics are built into architectures, which means it takes more than a new piece of infrastructure to resolve organizational silos. Some IT departments have built enterprise data lakes only to find that business units either refused to copy data into them or weren't provided with the appropriate tools and expertise to get data ingested.

## How to Avoid This Failure Scenario

Putting all enterprise data in a single location is a long-held vision in the world of IT, and the characteristics of data lakes appear finally to make it realizable. The reality, however, is often quite different, as described above.

Reconciling the various governance, performance, political and cultural issues may take months, more frequently years. In the meantime, more autonomous parts of the organization will create their own data lakes. These business unit lakes will be optimized for specific workloads, users and skills, and will likely be much more successful than their more aspirational enterprise data lake counterparts. These local business unit data lakes are more likely to succeed because the various challenges they face can be narrowly scoped and overcome, rather than posing intractable problems.

These smaller data lakes also point to a much larger trend for the decentralization of data and analytics within organizations. By contrast, to implement an enterprise data lake — a consolidation effort — is to run counter to the direction of many enterprise realities when it comes to data.

**Recommendations:**

- Build data lakes for specific business units or analytics applications, rather than try to implement some vague notion of a single enterprise data lake. The more you can refine the use case, data sources and users, the greater your chance of success. Work to resolve local issues of governance and performance, while also addressing data quality and data integration challenges. This will further improve your chances of success as additional data lakes are implemented across your enterprise.
- Make eventual consolidation of disparate data lakes a low priority. Consolidate based on workload characteristics, such as data volumes, data processing requirements, analytics and SLAs, rather than by business unit. For more information on data lake consolidation, see Note 1.

## Failure Scenario 2: “Data Lake *Is* My Data and Analytics Strategy”

A growing trend among data and analytics leaders, particularly new chief data officers, is to create an overarching data and analytics strategy supporting broader digital business initiatives. Some of these leaders look to data lakes as a quick substitute for more formal strategy development. Others have an ego-driven perspective on data lakes: They see them as means by which to be viewed as thought leaders, or to introduce major change to an enterprise they have recently joined.

This scenario typically fails for multiple reasons. These include a misunderstanding of what constitutes a data and analytics strategy, lack of organizational clout or social capital, underestimation of data management capabilities’ immaturity, and misunderstanding of the diverse requirements of a data and analytics platform for digital business.

### Mistaken Attempts to Replace Strategy Development With Infrastructure

A strategy defines how you win. At a high level, a data and analytics strategy must answer questions such as “What data? For what purposes? And by whom?” At a slightly lower level, a strategy must:

- Identify and guide the allocation of critical resources
- Define how to measure success

- Adapt to changing circumstances dynamically
- Act both proactively and reactively

Data and analytics leaders looking to avoid defining a strategy by, as it were, simply throwing everything into a data lake may have a vague notion of how they think their enterprise would win. It may be along the lines of "All the data is in the lake, so just go fishing for it!" But that leaves issues of measurement, reproducibility, quality, accountability and governance unresolved (see "Use the Gartner Data and Analytics Compass to Drive Strategy" and "Generally Accepted Information Principles for Improved Information Asset Management").

A data lake can be a technology component that *supports* a data and analytics strategy, but it cannot *replace* that strategy. You cannot simply buy some technology and think you have a strategy.

## Lack of Organizational Clout or Social Capital

Newly hired executives want to make an impact immediately. They often do so by undertaking major strategic initiatives, innovation projects or organizational changes. Implementing a data lake might look like a great way to make an impact, but many Gartner clients describe data lakes that were ignored or abandoned by business units because there was no motivation or urgency to use them.

Adding a data lake to your data management infrastructure means the affected business units have to change their processes. You have to overcome the inertia that business processes create just to get data into your data lake. Then you have to overcome more organizational inertia to get users to analyze that data. That's a big challenge. Unless you have significant organizational clout or social (and political) capital, it's unlikely you will succeed. And an early failure on a major initiative isn't a good look for a newly hired executive.

## Underestimation of the Immaturity of Data Management Capabilities

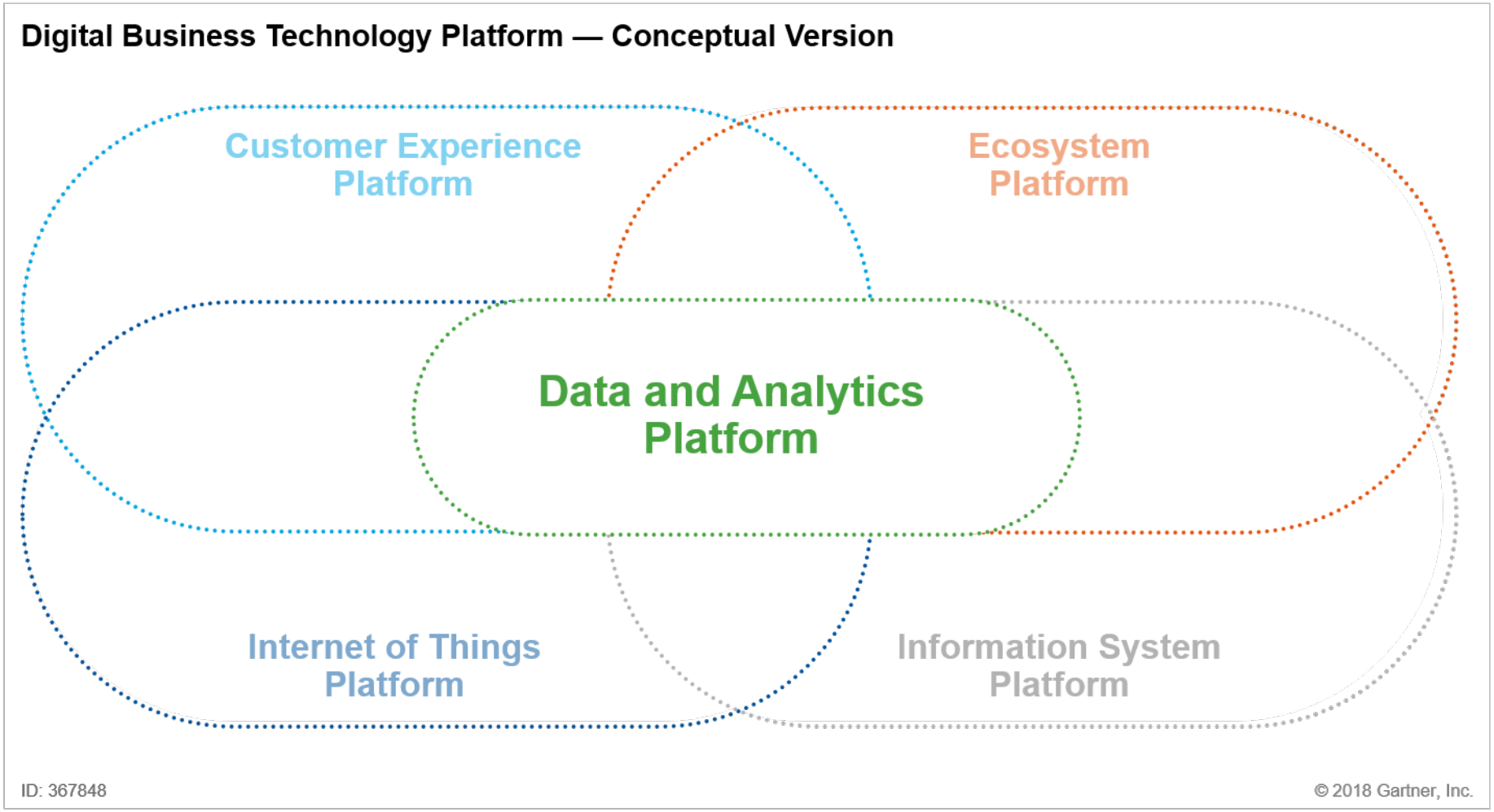
Data and analytics leaders commonly view implementing a data lake as a way to skip building essential data management capabilities, in the belief that the lake will be self-organizing, self-securing and self-governing. This belief is mistaken. Data lakes require *more* management capabilities because the data they contain lacks context. Data is pulled from source systems of record and placed in a lake, often with little connection to the originating source. (This problem is minimized in the purpose-built data lakes described in the section on the "Enterprise Data Lake" scenario above.)

Metadata management, data quality, data lineage and data integration, among other things, are crucial prerequisites for a successful data lake. They cannot be afterthoughts. If you don't have a rich data management discipline within your organization, it's unlikely your data lake will be successful (see "Making Big Data Normal Begins With Self-Classifying and Self-Disciplined Users").

## Misunderstanding of the Diverse Requirements of a Data and Analytics Platform for Digital Business

Every digital business platform has a data and analytics platform at its core. That data and analytics platform provides intelligence to the various platforms around it, such as Internet of Things and customer experience platforms. Figure 2 shows a conceptual version of a digital business platform.

Figure 2. Digital Business Technology Platform — Conceptual Version




Source: Gartner (August 2018)

The data and analytics platform appears conceptually similar to a data lake. It isn't. A data lake might provide the data and analytics platform with refined data or trained machine learning models, but the performance SLAs required by surrounding platforms are often much greater than data lake technology can support at scale. Digital business platforms also require consistent semantics and governance. Increasingly, these platforms must also provide justification for why or how a given decision was made or analytical result reached. These requirements are typically well beyond data lake capabilities (see "Building a Digital Business Technology Platform").

### How to Avoid This Failure Scenario

This scenario is the most difficult to avoid because the decision to pursue it is made largely for political, rather than business, reasons. Accepting that success is unlikely is the best way to avoid this scenario. You must define a data and analytics strategy. There is no technology infrastructure that enables you to skip this step.

**Recommendations:**

- Start data and analytics initiatives with a baseline measurement of the "as is" state, and show a clear path to financial and business objective contributions. Link data and analytics initiatives to three types of value:
  - Information value (which improves the information management process)
  - Business value (which improves business processes with data and analytics)
  - Stakeholder value (what the data and analytics mean for stakeholders, such as customers, partners, shareholders and society at large) (see "Data and Analytics Strategies Need More-Concrete Metrics of Success").
- Use this opportunity to introduce or reinvigorate a data management strategy. Position data management capabilities as a prerequisite for a successful data lake implementation. This will help to mature things like data quality, governance and metadata management.

## Failure Scenario 3: "Infinite Data Lake"

Organizations implementing an "infinite data lake" believe that all data maintains its original value indefinitely, and that data doesn't depreciate like other enterprise assets. Accordingly, these organizations expect their data lake infrastructure to scale indefinitely. The value proposition is essentially that organizations no longer have to be concerned with data life cycle or storage optimization because the data lake will accommodate any amount of data, now and in the future.

This scenario has all the risks of the "Enterprise Data Lake" failure scenario, as well as some others, as follows.

### Outdated or Irrelevant Data

Every business changes. From the products and services it offers, to the markets it offers them in, to the customers it sells them to. By storing massive amounts of historical data, the infinite data

lake may skew analysis by supplying data that isn't relevant to current market conditions or the changing mission-critical priorities of the business. Also, making sense of all that historical data means keeping, and understanding, the metadata describing it. This increases the overhead of curating the lake.

## Continuation of Immature Data Life Cycle Management Capabilities

Organizations with a low level of data life cycle maturity focus on local optimizations for the problems at hand. Reproducing work is often impossible, due to a lack of shared metadata, as well as of a shared understanding of the quality of different datasets. An infinite data lake perpetuates this situation. There is also the risk that such a lake contains data that it shouldn't, such as stale or otherwise deprecated (and no longer useful or valid) datasets. Essentially, this scenario ignores the data hygiene practices that are essential for effective use of data lakes, regardless of their intended use cases.

## Eventual Performance and Cost Challenges

Regardless of the data lake infrastructure chosen, storing increasingly massive amounts of data for an unlimited time will lead to scalability and cost challenges. Admittedly, scalability challenges are less of a risk in public cloud environments, but cost remains a factor. On-premises data lakes implemented with Hadoop, a common implementation choice, are more susceptible to cost challenges. This is because their cluster nodes require all three dimensions of computing, namely storage, memory and processing, but as a predominantly batch-processing option, the purchased memory and processing power largely sit idle.

## How to Avoid This Failure Scenario

This failure scenario is challenging to avoid because many people are obsessed with the idea that just having more data will solve their problems and help them seize new opportunities. There is no doubt that possession of data can confer competitive advantage, but it must be timely data and relevant to current challenges and market opportunities. Having more data for the sake of it doesn't deliver beneficial business outcomes. It creates liability.

**Recommendations:**

- Create a rigorous data life cycle policy for your data lake. You might, for example, require deletion of any data that hasn't been accessed in the last 30 or 60 days, or in the past four quarters.
- Implement automated data quality standards as part of the process of ingesting data into your lake. Reject any data below a predefined quality standard or metric. Ensure that your data scientists, data engineers or business analysts are responsible for data quality or are partnered with data quality specialists.

## Gartner Recommended Reading

*Some documents may not be available as part of your current Gartner subscription.*

"Best Practices for Designing Your Data Lake"

"Use Design Patterns to Increase the Value of Your Data Lake"

"Is the Data Lake the Future of the EDW?"

"Metadata Is the Fish Finder in Data Lakes"

"Efficiently Evolving Data From the Data Lake to the Data Warehouse"

### Evidence

- A Data Warehouse Evolution survey of 175 IT leaders conducted in March 2017.
- An Information Infrastructure Modernization survey of 111 IT and data and analytics leaders conducted in September 2016.
- Over 1,000 inquiries about data lakes from users of Gartner's client inquiry service, and one-to-one meetings at events, in the past three years. These interactions included discussion of architecture, organizational impact, strategy and implementation.
- H. Fang, "Managing Data Lakes in Big Data Era: What's a Data Lake and Why Has It Became Popular in Data Management Ecosystem" [sic]
- C. Madera and A. Laurent, "The Next Information Architecture Evolution: The Data Lake Wave"
- A. Halevy, F. Korn and others, "Managing Google's Data Lake: An Overview of the GOODS System"

### Note 1 Reconciling Disparate Data Lakes

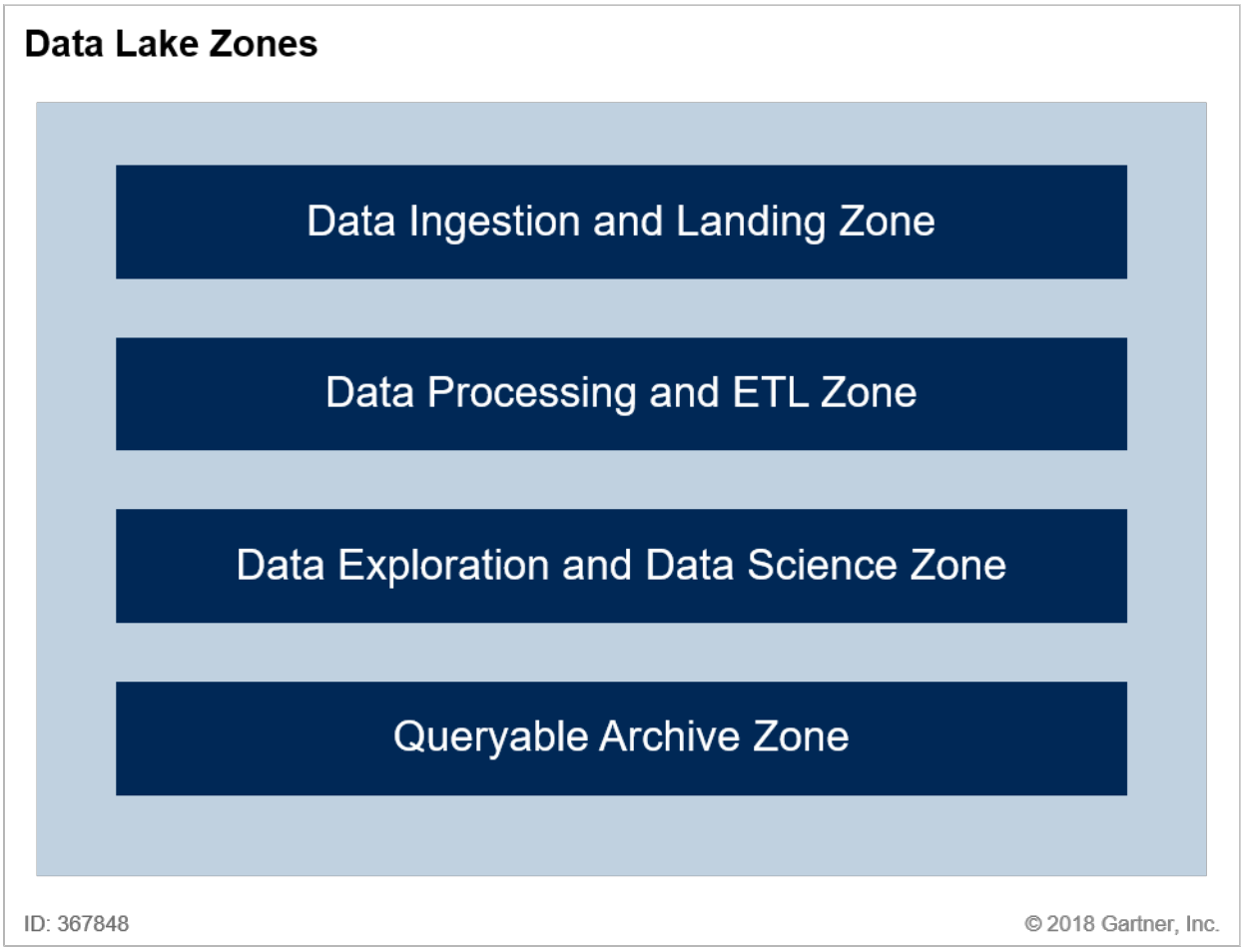
The goal of gathering all data in one location was never truly achieved in the data warehousing world. It's unlikely to be achieved in the data lake world, either — despite data lake proponents' promises. That said, the characteristics of data lakes enable us to get closer to this goal, provided we set expectations appropriately.

Governance, data life cycle and data hygiene practices should be applied to data lakes, just as to any other data management environment. If the technologies used lend themselves to consolidation based on data types, processing requirements and use cases, you may, by all means, pursue a consolidation and unification strategy.

Implementation, within a data lake, of data management zones for specific use cases and end-user populations can be an efficient means of consolidation. But this is only the case if the underlying technology can support appropriate workload management, use case separation and/or isolation, and segmentation of data for security and governance purposes. Definition of specific use-case-based zones for different purposes (as in Figure 3) can serve as the basis for any data lake

consolidation effort. Each zone will have different governance requirements, support different types and structures for data (relational, nonrelational, schema on write, schema on read), and serve different roles and skills within the organization (data scientist, data engineer, business analyst).

Figure 3. Data Lake Zones




ETL = extraction, transformation and loading

Source: Gartner (August 2018)

**Recommendations for data lake consolidation:**

- Define specific data zones within the consolidated data lake, based on use case or level of refinement for a given dataset.
- Define governance policies, based on how data is used within these zones and the end-user populations they serve. More adept users may require less governance.
- Maintain data life cycle and hygiene policies to ensure the validity and currency of data within individual zones. These policies will vary by use case.

- Monitor performance requirements and associated workload management configurations to ensure SLAs can be met for each zone.
- Enforce data quality and security standards when data passes from one zone to another.

**GARTNER HEADQUARTERS**

**Corporate Headquarters**
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

**Regional Headquarters**
AUSTRALIA
BRAZIL
JAPAN
UNITED KINGDOM

For a complete list of worldwide locations,
visit http://www.gartner.com/technology/about.jsp