ANGLING FOR INSIGHT IN TODAY'S DATA LAKE

October 2017

Michael Lock
Senior Vice President, Analytics and Business Intelligence

ABERDEEN

As data grows and diversifies, many organizations are finding that traditional methods of managing information are becoming outdated. This report aims to understand the performance implications of a data lake, and the common characteristics of those that leverage the technologies effectively.

A Growing Thirst for Better Data

The data lake analogy was conceived to help bring a common and visual understanding to the benefits of distributed computing systems able to handle multiple types of data, in their native formats, with a high degree of flexibility and scalability. While the analogy might not be perfect, the goal of a data lake is certainly well-aligned with the challenges so many companies struggle with today.

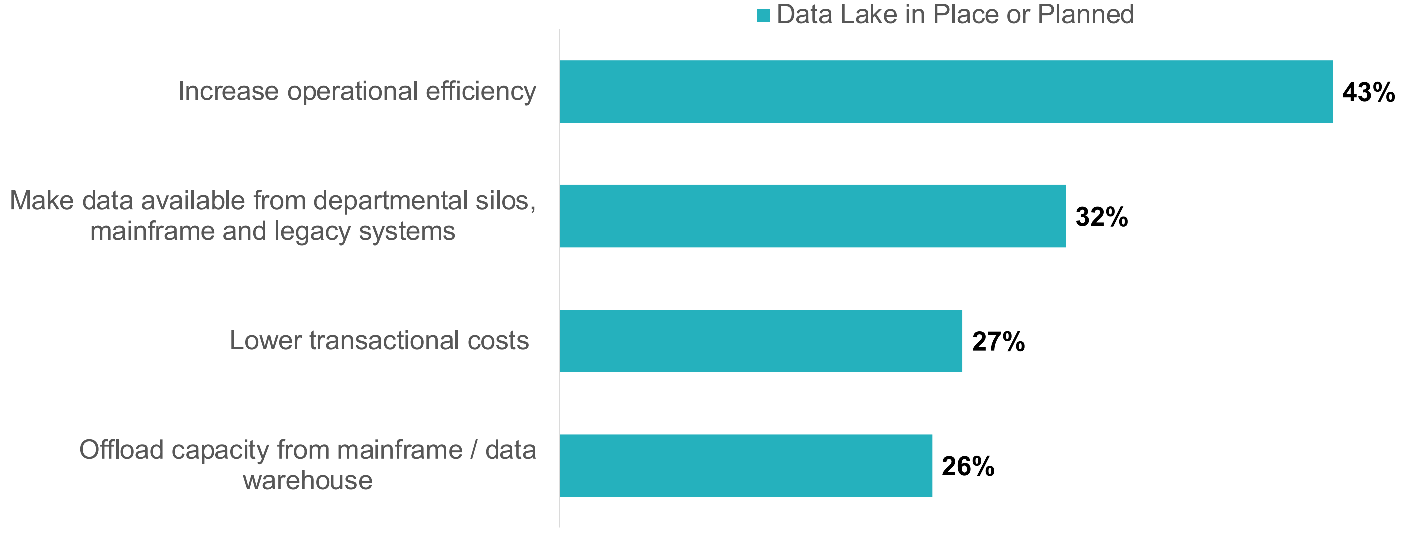
According to recent Aberdeen research, the average company is seeing the volume of their data grow at a rate that exceeds 50% per year. Additionally, these companies are managing an average of 33 unique data sources used for analysis. This type of rapid volume growth and complexity can wreak havoc on the internal efficiency of companies that depend heavily on data, hence many of these companies are responding by implementing data lake technologies (Figure 1).

The Aberdeen maturity class framework is comprised of three groups of survey respondents. Using self-reported performance across several key metrics, each respondent is classified into one of three categories:

- **Best-in-Class:** Top 20% of respondents based on aggregate performance
- **Industry Average:** Middle 50%
- **Laggard:** Bottom 30%

Sometimes the research cites a fourth category, **All Others**, which refers to the Industry Average and Laggard categories combined.

Figure 1: Why Invest in a Data Lake?



% of Respondents n = 193, Source: Aberdeen, September 2017

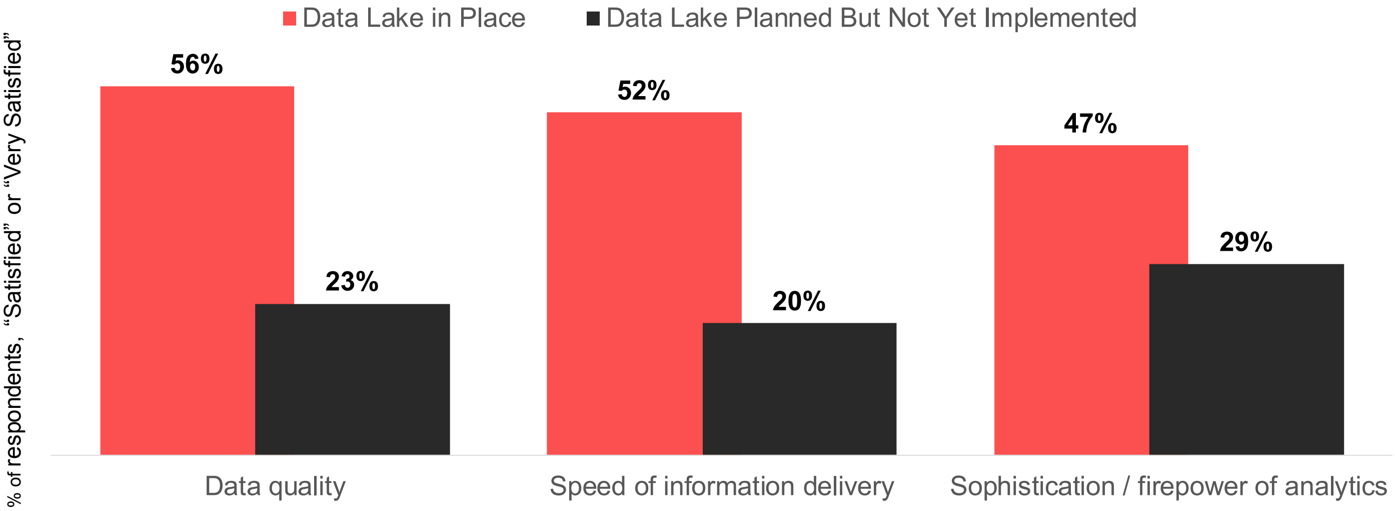
Companies look to more sophisticated data lake infrastructure to help exploit the influx of new data types, simultaneously making use of existing legacy data systems. While investment in data lake technology can be enticing, most organizations have made significant prior investments in legacy data warehouse and mainframe technologies. Not only can a data lake help companies exploit the potential of newer and more diverse data types, but it can also help make legacy systems more efficient by offloading capacity to the newer, more flexible infrastructure.

Not only can a data lake help companies exploit the potential of newer and more diverse data types, but it can also help make legacy systems more efficient by offloading capacity to the newer, more flexible infrastructure.

Empowering Today's Evolved User Base

Not only is the complexion of data changing in today's business environment, but the face of the traditional data user is evolving as well. Decision makers today, in a variety of different job roles, have heightened expectations when it comes to the data that fuels their most important decisions. At the same time though, quality and latency issues plague many of today's organizations. When comparing similar companies in the market for a data lake, those who have deployed the technology are more likely to report user satisfaction when it comes to critical metrics like quality and timeliness.

Figure 2: Elevated User Satisfaction with a Data Lake



n = 193, Source: Aberdeen, September 2017

These companies are also more likely to be satisfied with their ability to apply more advanced and sophisticated techniques to this foundation of data, ultimately an important characteristic in the context of a data lake. These architectures boast massive scalability in handling data, so there are certain organizations implementing a data lake for the sheer brute force of handling rapid data growth. However, many data-driven organizations today look to the flexibility of a data lake to help support new and different analyses performed against a variety of data. An effective data lake, and many of the associated open-source technologies, offer the potential to accomplish these more sophisticated analyses within an effective timeframe.

Data Lake "Leaders" Defined

Along with the potential analytical firepower of a data lake comes the inherent complexity and challenge of implementing and managing these environments.

In other words, an effective data lake involves far more than just writing a check for the requisition of software and hardware. Using findings from Aberdeen's 2017 Big Data survey, data lake "Leaders" were defined by their performance against the following three metrics:

- **Efficiency of data capture.** Today's most impactful analyses depend on data from a variety of different sources and applications. Top companies are able to reduce the amount of time finding and gathering data, enabling more time for analysis.

  (Leaders are more than twice as likely to report that the speed / efficiency of collecting data is "highly effective.")

- **Data accessibility.** With the right data captured from a variety of sources, leading companies are then able to expose that information to data professionals and business decision makers without an oppressive amount of red tape, or bureaucracy from IT.

  (81% of Leaders are "satisfied" or "very satisfied" with data accessibility, compared with only 34% of Followers.)

- **Timeliness of information.** All users have an effective window of time, during which the right piece of information can impact their decisions. The ability to get information on time, regardless of the length of their "decision window," is critical.

  (Leaders report that 84% of information is delivered on time, compared with 61% for Followers.)

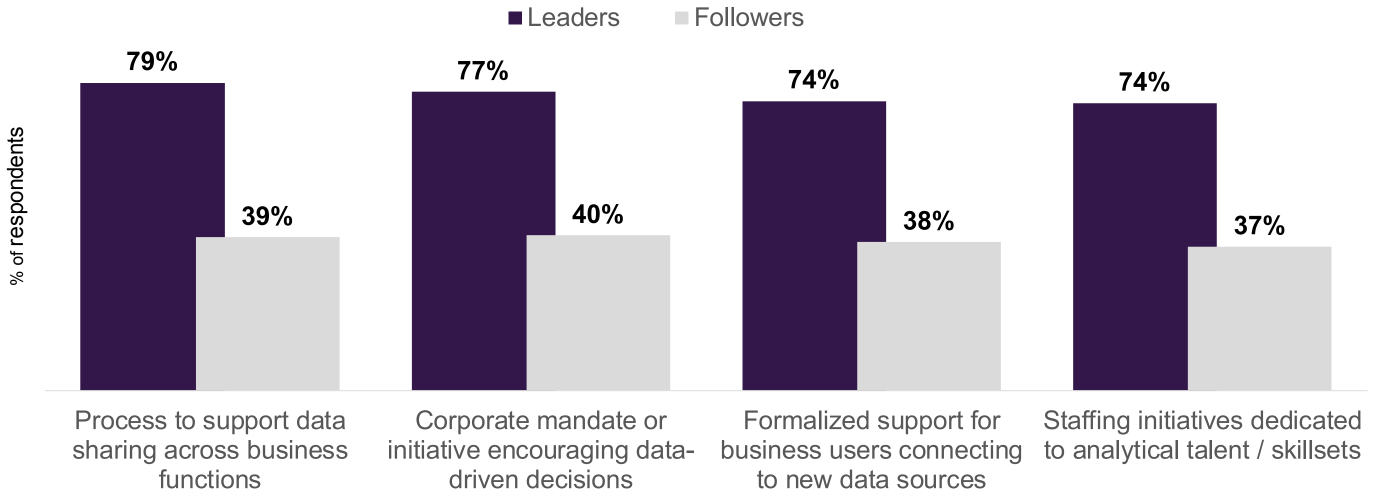
As an addition to the maturity class framework outlined in the sidebar on page 2, Aberdeen sometimes segments respondents into two separate categories as follows:

- **Leaders:** Top 40% of respondents based on aggregate performance
- **Followers:** Remaining 60% of respondents based on aggregated performance

Separating the Best from the Rest

Capitalizing on the potential of a data lake requires more than just deploying technology. Leaders put in place a variety of capabilities that help support their performance. One critical capability is enabling the access and distribution of data to help fuel these newer and potential game-changing analyses. Leaders are more likely to have processes in place for sharing data across business functions, and support business users in their efforts to connect to new data sources on their own (Figure 3).

Figure 3: Strong Capabilities Supporting Data Maturity



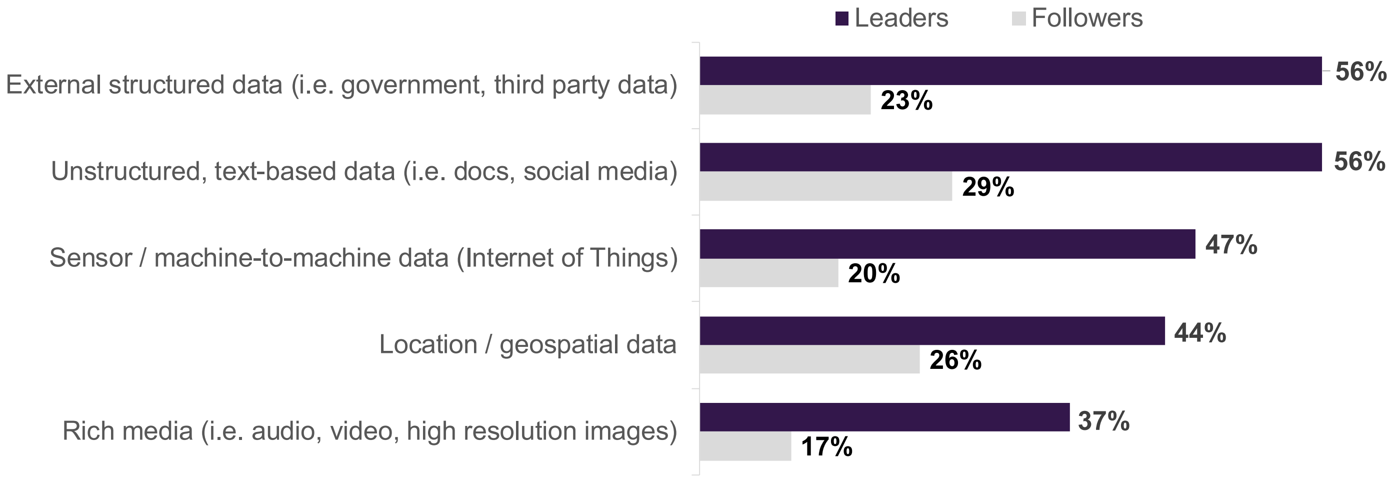
n = 127, Source: Aberdeen, September 2017

Additionally, leaders put programs in place to raise the analytical bar within their organizations. Internally, leaders are more likely to have encouragement (or even mandates) from senior leadership to help drive more use of data and the creation of data-driven insights. Leaders also make analytical skill sets a priority when looking externally for talent. These companies are twice as likely to have hiring and staffing initiatives weighted heavily on analytical talent and expertise.

An additional characteristic shared by Leaders is their ability to be flexible in the data they use for analysis. Just about every technology or service, dedicated to supporting a data lake environment will tout the ability to handle multiple forms of data in their native formats, staged and ready to be exploited for more cutting-edge analyses.

The research demonstrates that the average organization deals with 31 unique data sources that can feed into their analytical systems, but not all of that data is traditional, application-based data. Arguably, the most meaningful insights today are generated using of a variety of different data types, from external data, to unstructured information or machine-generated Internet of Things (IoT) data.

Figure 4: Leaders Crave a Diversity of Data



% of Respondents rating data types as "critical", n = 127

Source: Aberdeen, September 2017

The chart above demonstrates that data lake Leaders are adept at handling an assortment of complex data, partly because of the maturity of their technology, and partly out of necessity, as many of them rate these data types as "critical" for the purposes of analysis.

The Payoff: Leaders Deliver Results

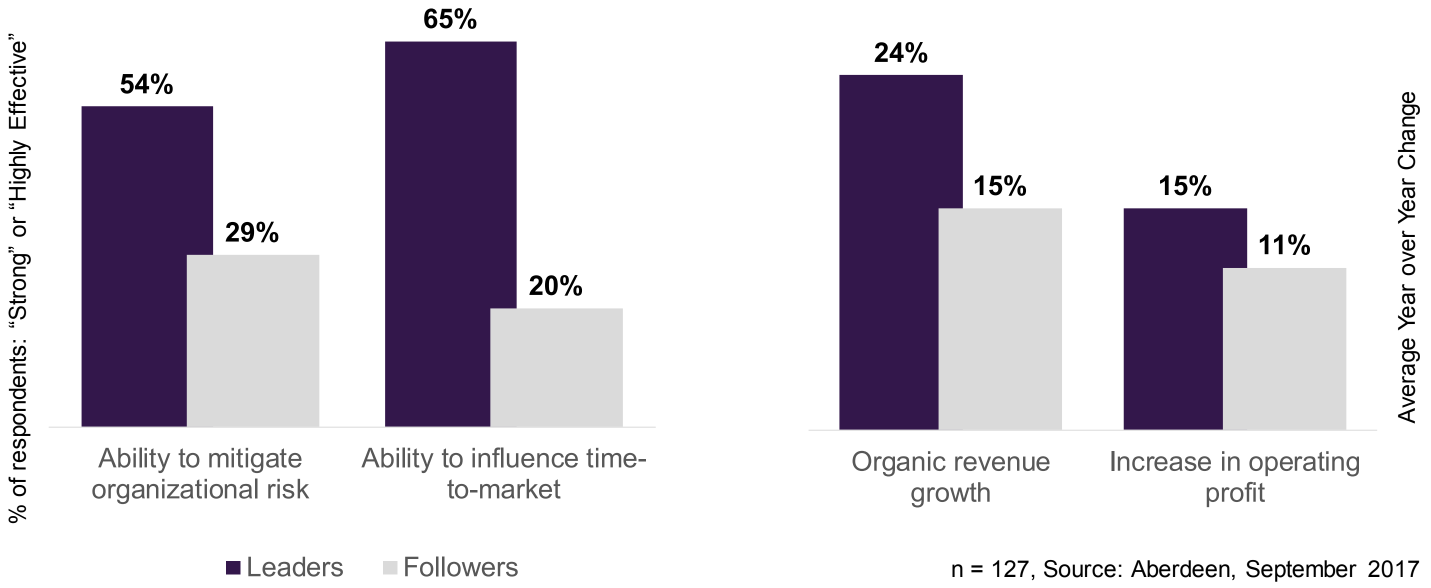
Aside from the metrics used to determine leading performance with a data lake, top companies are able to use this architecture to deliver results in multiple areas of their business.

With a clean and consumable foundation of data and an empowered user base, Leaders are able to expedite the flow of information across their organization and deliver critical information on time. This has several demonstrable effects on the internal operations for these top performers, but two in particular. First, these companies report a substantially higher degree of trust in their data, which ultimately helps them mitigate the risk associated with duplicated, corrupted, or simply absent data and other things that can cause major problems for any data-driven company. Secondly, data lake Leaders are able to utilize this internal data efficiency to help perform core activities more effectively (in

Data Lake Leaders are adept at handling an assortment of complex data, partly because of the maturity of their technology, and partly out of necessity, as many of them rate these data types as "critical" for the purposes of analysis.

this case, the process of bringing products and services to market). Leaders are more than three-times as likely to report a “strong” or “highly effective” go-to-market process (Figure 5).

Figure 5: Data Efficiency and Business Execution



n = 127, Source: Aberdeen, September 2017

In addition to this internal process efficiency, a well-conceived data lake helps set the stage for an elevated level of analytical activity. For leading companies, those analytical activities also translate to substantial ROI in the form of business growth and profit boost. Leaders saw markedly higher year-over-year improvements in operating profit and organic revenue growth as well.

Key Takeaways

Ultimately, companies implement a data lake for two main reasons. First, they want to take advantage of more advanced and sophisticated analytical techniques, applied towards a more complex and diverse foundation of information. Secondly, they want to make traditional activities — like data access and speed of retrieval — more efficient and easier to perform. While not every company succeeds at achieving both objectives simultaneously, the most effective ones are able to see results from implementation. The following are highlights of the most important takeaways from Aberdeen’s research:

- **Data lakes offer a comprehensive and holistic solution to big data.** The typical language used to describe the benefits of a data lake — flexibility, scalability, etc. — is practical, if a little vague.

The fact is, data-rich companies today face many challenging and unusual situations with respect to their data. They have a variety of data types (external, unstructured, IoT), coming from numerous sources, housed in cloud and on-premise deployments, growing at a rapid pace, and used for countless types of workloads and use cases. The ideals of a modern data lake are aimed at helping companies manage these challenges with a more complete big data solution.

- **The biggest change companies see is in the mindset of their users.** One could argue that the rapid advancement in technology is primarily responsible for today's increased user expectations. However, the fact remains that people demand cleaner, faster, and more relevant information. Leading companies today leverage their data lake infrastructure to help their users exploit a wider variety of data. This supports their ability to produce game-changing insights from its analysis.
- **An efficient data lake can be an engine for business performance.** In the analytical process of transforming raw data into deliverable insight, a top-notch data lake strategy produces higher quality, more relevant data, delivered in a timelier way. With a flow of data that is faster, more frictionless, and more trustworthy, companies are able to identify — and act upon — immediate opportunities for business growth and efficiency.

Related Research

Predictive Analytics: The Science of Soothsaying; August 2017

Modern MDM: The Hub of Enterprise Data Excellence; June 2017

Analytics in the Age of IoT: Today's Data-Driven Competitive Edge; May 2017

The Data Warehouse Evolved: A Foundation for Analytical Excellence; May 2017

About Aberdeen Group

Since 1988, Aberdeen Group has published research that helps businesses worldwide to improve their performance. Our analysts derive fact-based, vendor-neutral insights from a proprietary analytical framework, which identifies Best-in-Class organizations from primary research conducted with industry practitioners. The resulting research content is used by hundreds of thousands of business professionals to drive smarter decision-making and improve business strategies. Aberdeen Group is headquartered in Waltham, Massachusetts, USA.