



polygraphus/Getty Images

A smarter way to jump into data lakes

Mikael Hagstroem, Matthias Roggendorf, Tamim Saleh, and Jason Sharma

An agile approach to data-lake development can help companies launch analytics programs quickly and establish a data-friendly culture for the long term.

Increases in computer-processing power, cloud-storage capacity and usage, and network connectivity are turning the current flood of data in most companies into a tidal wave—an endless flow of detailed information about customers’ personal profiles, sales data, product specifications, process steps, and so on. The data arrive in all formats and from a range of sources, including Internet-of-Things devices, social-media sites, sales systems, and internal-collaboration systems.

Despite an increase in the number of tools and technologies designed to ease the collection, storage, and assessment of critical business information, many companies are still unsure how best to handle these data. Business and IT leaders have told us they remain overwhelmed by the sheer volume and variety of data at their disposal, the speed at which information is traversing internal and external networks, and the cost of managing all this business intelligence. Increasingly, they are also being

charged with an even more complicated task: harnessing meaningful insights from all this business information.

These executives must expand their data-management infrastructures massively and quickly. An emerging class of data-management technologies holds significant promise in this regard: data lakes. These storage platforms are designed to hold, process, and analyze structured and unstructured data.¹ They are typically used in conjunction with traditional enterprise data warehouses (EDWs), but in general, they cost less to operate than EDWs. Cost savings result because companies can use affordable, easy-to-obtain hardware and because data sets do not need to be indexed and prepped for storage at the time of induction. Data are held in their native formats and reconfigured only when needed, as needed. Relational databases may also need to be managed as part of the data-lake platform, but only to ease end users' ability to access some data sources.

There is a lot for companies to like about data lakes. Because data are loaded in "raw" formats rather than preconfigured as they enter company systems, they can be used in ways that go beyond just basic capture. For instance, data scientists who may not know exactly what they are looking for can find and access data quickly, regardless of format. Indeed, a well-maintained and governed "raw data zone" can be a gold mine for data scientists seeking to establish a robust advanced-analytics program. And as companies extend their use of data lakes beyond just small pilot projects, they may be able to establish "self-service" options for business users in which they could generate their own data analyses and reports.

However, it can be time consuming and complicated to integrate data lakes with other elements of the technology architecture, establish appropriate rules for company-wide use of data lakes, and identify the supporting products, talent, and capabilities needed to deploy data lakes and realize significant business benefits from them. For instance, companies typically lack expertise in certain data-management approaches and need to find staffers who are fluent in emerging data-flow technologies such as Flume and Spark.

In many cases, companies are slowing themselves down. They are falling back on tried-and-true methods for updating technology architectures—for instance, engaging in long, drawn-out internal discussions about optimal designs, products, and vendors and holding off on building a data-lake solution until they have one that is just right. In the meantime, opportunities to deploy advanced-analytics programs that will support digital sales and marketing and new-product development simply pass them by.

Companies should instead apply an agile approach to their design and rollout of data lakes—piloting a range of technologies and management approaches and testing and refining them before getting to optimal processes for data storage and access. The companies that do can keep up with rapidly changing regulatory and compliance standards for data—for instance, the European Union's General Data Protection Regulation, which is slated to take effect in May 2018. Perhaps more important, they can bring analytics-driven insights to market much faster than their competitors while significantly reducing the cost and complexity of managing their data architecture.

¹ "Structured" data (such as an Excel spreadsheet) are well organized and therefore easily identified by search algorithms; "unstructured" data (such as an audio file) are less organized and therefore less likely to be responsive to search algorithms.

Stages of data-lake development

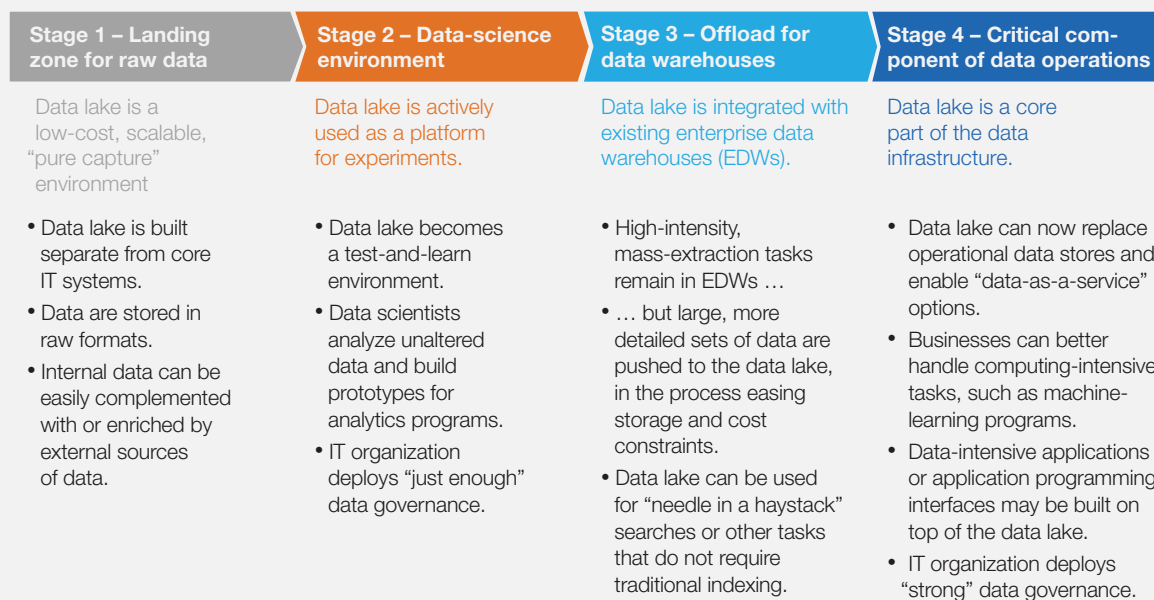
Companies generally go through the following four stages of development when building and integrating data lakes within their existing technology architectures (exhibit):

- Landing and raw-data zone.** At the first level, the data lake is built separate from core IT systems and serves as a low-cost, scalable, “pure capture” environment. The data lake serves as a thin data-management layer within the company’s technology stack that allows raw data to be stored indefinitely before being prepared for use in computing environments. Organizations can deploy the data lake with minimal effects on the existing architecture. Strong governance, including
- Data-science environment.** At this next level, organizations may start to more actively use the data lake as a platform for experimentation. Data scientists have easy, rapid access to data—and can focus more on running experiments with data and analyzing data, rather than focusing solely on data collection and acquisition. In this sandbox, they can work with unaltered data to build prototypes for analytics programs. They may deploy a range of open-source and commercial tools alongside the data lake to create the required test beds.

rigorous tagging and classification of data, is required during this early phase if companies wish to avoid creating a data swamp.

EXHIBIT

Companies may go through any or all of these four stages of building and integrating data lakes within technology architectures.



- **Offload for data warehouses.** At the next level, data lakes are starting to be integrated with existing EDWs. Taking advantage of the low storage costs associated with a data lake, companies can house “cold” (rarely used, dormant, or inactive) data. They can use these data to generate insights without pushing or exceeding storage limitations, or without having to dramatically increase the size of traditional data warehouses. Meanwhile, companies can keep high-intensity extraction of relational data in existing EDWs, which have the power to handle them. They can migrate lower-intensity extraction and transformation tasks to the data lake—for instance, a “needle in a haystack” type of search in which data scientists need to sweep databases for queries not supported by traditional index structures.
- **Critical component of data operations.** Once companies get to this stage of rollout and development, it is very likely that much of the information that flows through the company is going through the data lake. The data lake becomes a core part of the data infrastructure, replacing existing data marts or operational data stores and enabling the provision of data as a service. Businesses can take full advantage of the distributed nature of data-lake technology as well as its ability to handle computing-intensive tasks, such as those required to conduct advanced analytics or to deploy machine-learning programs. Some companies may decide to build data-intensive applications on top of the data lake—for instance, a performance-management dashboard. Or they may implement application programming interfaces so they can seamlessly combine insights gained from data-lake resources with insights gained from other applications.

The time and capabilities required for companies to grow their data lakes from simple landing zones to critical components of the data infrastructure will vary depending on companies’ objectives and starting points. At each stage of development, companies need to examine complicated questions relating to the size and variety of their data sets, their existing capabilities in data management, the level of big data expertise in their business units, and product knowledge in the IT organization. For instance, how sophisticated are analytics tools in the current environment? Is the company using traditional development tools and methodologies, or newer ones? How many concurrent data users does the company typically require? Are workloads managed dynamically? How quickly do end users need access to data? At various points in the data-lake development process, companies can get mired in these details and lose momentum; leaders in the IT organization or the business units inevitably fan out to tackle other “urgent” projects.

The data lake’s journey from “science project” to fully integrated component of the data infrastructure can be accelerated, however, when IT and business leaders come together to answer these and other questions under an agile development model. In our experience, an agile approach can help companies realize advantages from their data lakes within months rather than years. Quick wins and evidence of near-term impact can go a long way toward keeping IT and business leaders engaged and focused on data-management issues—thereby limiting the need for future rework and endless tweaking of protocols associated with populating, managing, and accessing the data lake. An agile approach can put IT and business leaders on the same page. Such collaboration

is critical not just for determining a technical path forward for the data lake but also for establishing a data-friendly work environment and seizing new business opportunities based on insights from data.

Building a data lake: An agile approach

Most organizations understand the need for agile methodologies in the context of software development. Fewer have applied agile in the context of data management. Typically, the IT organization takes the lead on vetting potential technology options and approaches to building data lakes, with little input from the business units. Under an agile approach, IT and business leaders jointly outline and address relevant technology and design questions. For instance, will the data lake be built using a turnkey solution, or will it be hosted in the cloud (using private, public, or hybrid off-site servers)? How will the data lake be populated—that is, which data sets will flow into the lake and when? Ideally, the population of the data lake should be based on the highest-priority business uses and done in waves, as opposed to a massive one-time effort to connect all relevant data streams within the data lake.

Indeed, the most successful early adopters are designing their data lakes using a “business back” approach, rather than considering technology factors first. They are identifying the scenarios in which business units could gain the most value from the data lake and then factoring those scenarios into the design (or redesign) of the storage solution and rollout decisions. Companies are then incrementally populating the data lake with data for specific groups or use cases, as needed. And rather than going all in on one designated solution, companies are piloting two or three final candidates from different providers to assess

the real-world performance, ease of integration, and scalability of their offerings.

This agile approach to rollout can ensure that performance or implementation challenges will be caught early. It incorporates feedback from the business units. It also leaves room for agile development teams to tinker with processes and data-governance protocols as the data lake fills up, analytics and storage technologies change, and business requirements evolve.

As data lakes move from being pilot projects to core elements of the data architecture, business and technology leaders will need to reconsider their governance strategies. Specifically, they must learn to balance the rigidity of traditional data oversight against the need for flexibility as data are rapidly collected and used in a digital world. Under an agile approach to governance, businesses can apply sufficient oversight as new sources enter the data lake, avoiding some of the more rigid engineering practices required in traditional data warehouses and then refining rules and processes as business requirements dictate to get to an optimal solution. For instance, data scientists might be given free rein to explore data, even as business cases for certain categories of data are still being identified. Meanwhile, frontline users might face stricter controls until use cases are more firmly established.

At the very least, however, companies should designate certain individuals as owners of data sets and processes, so that responsibilities are clear and decisions about data sources and access rights can be made quickly. Because data are not being structured up front, companies will also want to capture and store metadata on all the data sources flowing into the lake (either within the lake itself or in

a separate registry) and maintain a central data catalog for all stakeholders. Additionally, companies may need to reconfigure access rights as they iterate on data-management protocols—keeping in mind regulatory requirements and privacy issues related to holding personally identifiable information. Data owners must communicate these access rights to all relevant stakeholders.

Transformation at a global bank

Let's consider how a global bank applied agile principles to its development of a data lake. The bank had been struggling with several critical data challenges: low-quality business information, lack of specialists to manage different data sets arriving in different formats, aging data-warehouse technologies, and more than 1,000 data sources. The systems were kludgy. Incoming data sets had to be structured before they could be entered into four data-warehouse layers (output delivery, normal form, subject layer, and app layer) and before any usable reports could be created.

Outside of these technical challenges, business and IT leaders at the bank were not working collaboratively, which exacerbated the company's data problems. Data were being stored in isolated systems, so critical business information often remained trapped. But requests for access to certain data sets were slow to get a response because of poor coordination and communication across business units and IT operations. Data management was seen as "IT's job"; business leaders held the topic at arm's length and thus struggled to articulate their data needs.

Senior leaders at the bank were concerned about losing customers, in part due to the company's inability to manage data adroitly. They decided to experiment with data-lake

technologies to try to ease the extraction, structuring, and delivery of data sets. Seeking to work as quickly as its software developers, the company used an agile development model and rolled out the data-lake project in phases.

Senior leaders convened an agile data team involving subject-matter experts from the business units and from the IT organization to consider the business impact of and use cases for improved data quality and access before determining which areas of the company would have initial access to the data lake.

The agile data team conducted in-depth interviews with business users to identify pain points and opportunities in existing data-management practices. The team's plan was to release waves of new data services and applications in four-month windows—implementing new data-management tools, developing data-delivery services with the business units, and refining processes based on customers' feedback. Within months of the initial launch of the agile data project, the bank was able to load data relevant to particular business use cases into a common environment and identify the critical data elements required for providing services to the business units.

Success in high-profile areas of the business enabled the bank to extend the usage of the data lake to other areas in subsequent months. The shift from structuring all the data up front to documenting a back-end process only for utilized data was significant. The bank was able to break down data silos; information from systems could now be found in one place, and employees were able to access multiple forms of data (demographic, geographic, social media, and so on) to gain a 360-degree view of

customers. Collaboration between the business units and the IT group increased, as did employees' and customers' satisfaction scores.



More and more companies are experimenting with data lakes, hoping to capture inherent advantages in information streams that are readily accessible regardless of platform and business case and that cost less to store than do data in traditional warehouses. As with

any deployment of new technology, however, companies will need to reimagine systems, processes, and governance models. There will be inevitable questions about security protocols, talent pools, and the construction of enterprise architecture that ensures flexibility not just within technology stacks but also within business capabilities. Our experience suggests that an agile approach to the implementation of data lakes can help companies climb the learning curve quickly and effectively. ◆

Mikael Hagstroem (Mikael_Hagstroem@McKinsey.com) is a partner in McKinsey's Charlotte office, **Matthias Roggendorf** (Matthias_Roggendorf@McKinsey.com) is a senior expert in the Berlin office, **Tamim Saleh** (Tamim_Saleh@McKinsey.com) is a senior partner in the London office, and **Jason Sharma** (Jason_Sharma@McKinsey.com) is an associate partner in the Silicon Valley office.

The authors wish to thank Prasoon Sharma for his contributions to this article.

Copyright © 2017 McKinsey & Company. All rights reserved.

Digital/McKinsey


August 2017

Designed by Global Editorial Services

Copyright © McKinsey & Company

McKinsey.com

 [@DigitalMcKinsey](https://twitter.com/DigitalMcKinsey)

 facebook.com/DigitalMcKinsey