bcg.perspectives
by THE BOSTON CONSULTING GROUP

LOOK BEFORE YOU LEAP INTO THE DATA LAKE

By Rash Gandhi, Sanjay Verma, Elias Baltassis, and Nic Gordon

To fully capture the tremendous value of using big data, organizations need nimble and flexible data architectures able to liberate data that could otherwise remain locked within legacy technologies and organizational processes.

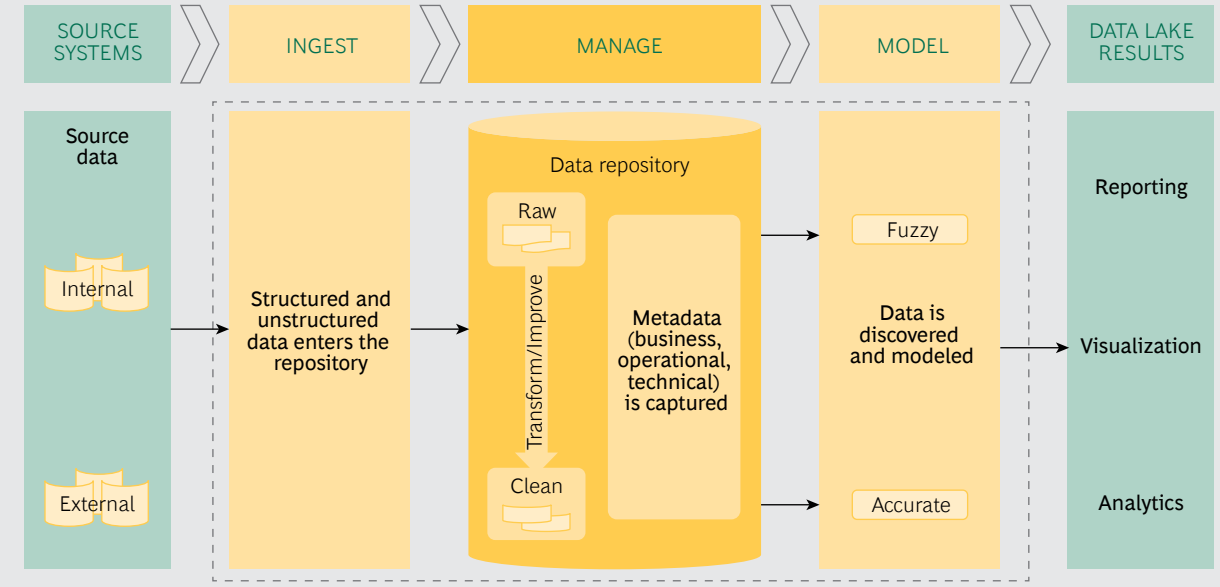
Rapid advances in technology and analytical processing have enabled companies to harness and mine an explosion of data generated by smartphone apps, website click trails, customer support audio feeds, social media messages, customer transactions, and more. Traditional enterprise data warehouse and business intelligence tools excel at organizing the structured data that businesses capture—but they stumble badly when it comes to storing and analyzing data of the variety and quantity captured today and doing so at the speed now required. Companies need data architectures that can handle the diversity of data available now (semistructured data, unstructured data, log files, documents, videos, and audio, for example) and yield even more accurate predictive modeling and customer insight at a highly detailed level.

Enter the "data lake," a term that refers to a large repository of data in a "natural," unprocessed state. Data lakes' flexibility and size allow for substantially easier storage of raw data streams that today include a multitude of data types. Data can be collected and later sampled for ideas, tapped for real-time analytics, and even potentially treated for analysis in traditional structured systems. But before organizations dive into the data lake, it's important to understand what makes this new architecture unique, the challenges organizations can face during implementation, and ways to address those challenges.

What Exactly Is a Data Lake?

Historically, organizations have invested heavily in building data warehouses. Significant up-front time, effort, and cost go into identifying all the source data required for analysis and reporting, defining the data model and the database structure, and developing the programs. The process often follows a sequence of steps known as ETL: extract source data, transform it, and load

EXHIBIT 1 | How a Data Lake Works



Source: BCG analysis.

it into the data warehouse. Making changes to an existing data warehouse requires sizable additional investment to redesign the programs that extract, transform, and load data—we estimate that 60% to 75% of development costs come in the ETL layer.

Moreover, data warehouse solutions typically provide historical, or backward-looking, views. And here is where the challenge arises: organizations today are demanding that data tell them not just what happened in the past but also what is likely to happen in the future. They seek predictive and actionable insights, gleaned from a variety of data accessed through both batch and real-time processing to inform their strategies.

Traditional data warehouses are not ideal solutions to this challenge. They are slow to change and costly to operate, and they can't be scaled cost-efficiently to process the growing volume of data. Data lakes can fill the void.

Given companies' storage requirements (to house vast amounts of data at low cost) and computing requirements (to process and run analytics on this volume of data), data lakes typically use low-cost, commodity servers, in a scale-out architecture. Servers can be added as needed to increase processing power and data capacity. These systems are typically configured with data redundancy to ensure high resilience and availability. Much of the big data software is open source, which drives down costs. The total cost of establishing and running a data lake can be five to ten times lower than the cost of using traditional SQL-based systems.

A company's data lake can be built on any of multiple technology ecosystems (for example, Hadoop, Drill, and Cassandra), the most notable of which is the well-established Hadoop. Both upstarts (including Cloudera, MapR, and Hortonworks) and traditional IT players (such as IBM, HP, Microsoft, and Intel) have used Hadoop in constructing their data lakes.

Data lakes are highly flexible, and they enable a responsive "fail fast" approach to analytics that can drive significant value. In their simplest form, data lakes have three core functions (see Exhibit 1):

- To ingest structured and unstructured data from multiple sources into a data repository
- To manage data by cleaning, describing, and improving it

- To model data to produce insights that can be visualized or integrated into operating systems

What Is Different About Data Lakes?

A data lake brings new approaches to data management on several fronts.

Content Variety. A data lake is designed to store and process content in a wide variety of states (including multistructured, unstructured, and structured content), unlike traditional data warehouses, which can meaningfully store and process only structured content. As an example of the power of data lakes, an analysis of unstructured data such as e-mails or voice calls linked to customer transactions can help identify fraudulent trading activity. A data warehouse is limited to storing only structured data, making such analyses difficult and time-consuming.

Data Structure. Traditional data architectures mandate a database structure that is defined up front. Data architects prescriptively model and define the physical database prior to transforming and loading data into it, a process referred to as "schema on write."

But companies increasingly need an architecture in which users are free to access and structure data dynamically on the fly; that process is sometimes referred to as "schema on read" or "late-binding execution." Instead of using the sequence common to data warehouses—extract, transform, load (ETL)—it employs the ELT approach, swapping the load and transform steps so that the raw loaded data is cleaned and transformed in the data lake. That offers an advantage when the accuracy of the data is imperative, such as for regulatory reporting. Because data quality validation happens as needed in the data lake, you don't need to create a big IT project to clean all the data, thus saving time and cost.

The flexibility of a schema-on-read model enables users to experiment with a variety of data and create innovative business insights dynamically. The schema-on-read attribute of a data lake can be one of its major strengths, especially when the properties of the data are correctly described and when data quality is fully understood. Failure to adequately govern those properties, however, can quickly undermine the data lake's business value. When data is not properly described, for example, it can't be understood and searched; when its quality isn't measured, it can't be trusted.

Timeliness of Data. Information can be streamed directly into a data lake, providing an almost real-time view of the data. This can be important when business decisions rely on real-time analytics, such as making credit decisions. With traditional SQL data architectures, a delay typically occurs as source data is cleaned and then loaded into a data warehouse, in hourly, daily, or weekly batches. Furthermore, traditional SQL systems use heavy controls to insert (that is, to store) data, which slows down data throughput. Users of data warehouses may, therefore, not have the most up-to-date information, which is a shortcoming that can undermine the quality of the insights derived through data analytics.

Data Quality and Validation. Because raw data gets loaded into a data lake, its quality and trustworthiness is tested at the time it is accessed for analysis. Traditional SQL-based data warehouse programs, on the other hand, undergo extensive testing of the programs used to extract, transform, and load data, which can help ensure the high quality of the data moved into them.

Access and Security. Tools to capture basic technical metadata for the data ingested in a data lake are available, but that information still must be enriched with business and operational metadata that enables users to access and fully exploit data. The rich metadata and associated security policies enforced in a traditional data warehouse enable construction of complex user-access models. Fine-grained security policies and role-based access privileges grant effective control of user access to content.

Effort and Cost. Data lakes are significantly easier and less expensive to implement than traditional SQL-based data warehouses, owing to the commoditized nature of the platform, the lower cost of open-source technologies, and the deferment of data modeling until the user needs to analyze the data. In contrast, the cost of a data warehouse solution can reach millions of dollars for large companies as a result of the need for up-front data modeling, the long period of time to design and build, the requirement for data integration, and the need to customize database, server, data integration, and analytics technologies.

Overcoming the Challenges of Data Lakes

Today's data lakes present two overarching challenges.

The first is the lack of tools or, at least, mature tools available for the Hadoop environment. Data warehouses have built these tools over more than two decades. Data lakes have much catching up to do. For example, data lakes don't yet have the level of security that users of data warehouses are accustomed to, although significant improvements have been made and the environment is vastly different from that of just a year ago. The same situation exists in terms of data quality and validation. Metadata tools within the Hadoop environment used for checking data quality have been maturing over the past several years. But without robust controls, users could lose trust in the accuracy of the data needed to derive value, such as for regulatory reporting.

The second is the skills gap. Not enough people have sufficient experience working in the data architecture of the Hadoop environment. Eighty-three percent of respondents to a 2016 survey by CrowdFlower said there is a shortage of data scientists, up from 79% in 2015. In 2020, Gartner predicts, the US will face a shortage of 200,000 data scientists.

The good news is that, once these challenges are addressed, the process of developing a data lake is useful in and of itself. Companies simply won't be able to model data in the ways they need to in the future without the flexible architecture that building a data lake creates. But to get the greatest value from these efforts, we recommend the following steps.

Identify the highest-value opportunities. The shift toward big data architectures has forced many senior executives to take a fresh look at the major components of a data architecture strategy and the game-changing capabilities they enable. The first step in this work is to think through the highest-value use cases for big data. Across industries, we often find uses in real-time customer ad targeting, real-time risk and fraud alert monitoring, regulatory reporting, and IT performance optimization. (See "Big Data and Beyond," a collection of BCG articles about big data.) Once companies determine these use cases, they must identify the target organization and the processes and technologies required for the transformation.

Keep the right goals in mind. A data lake typically has three uses.

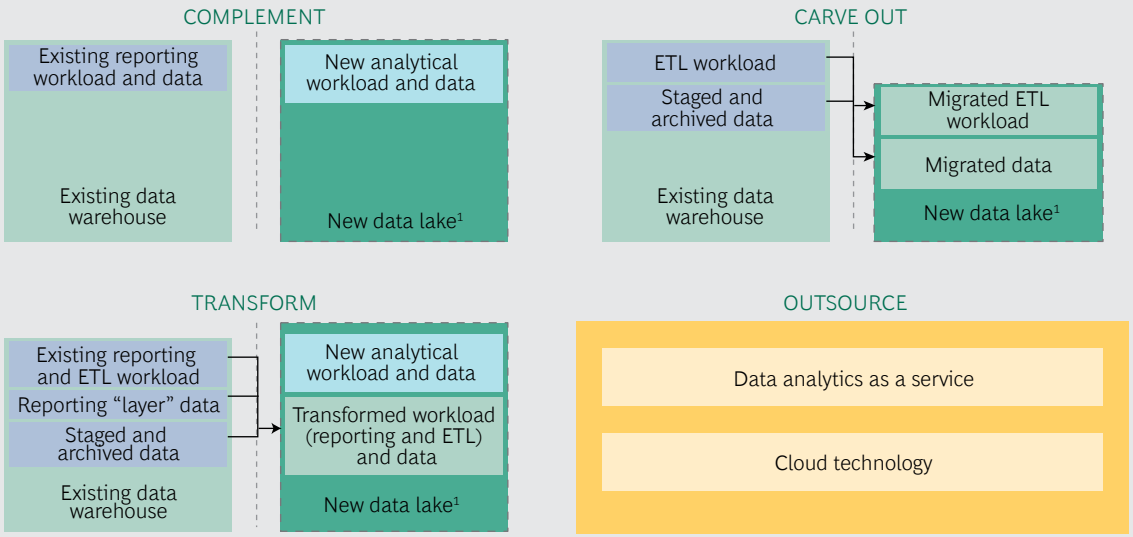
The first, and currently primary, use is *insight generation*—pulling global data for reporting or visualization purposes, for example, or running machine-learning jobs to determine new connections and relationships.

The second use is *operational analytics,* which is more time sensitive than insight generation. An example is assessing the credit risk associated with a transaction while the transaction is being processed; the risk associated with the transaction amount, location, or type is scored in real time based on the customer's profile and history.

The big data landscape is constantly changing, and new components—such as Mahout, Tez, and Pig—are maturing, making it possible to mostly replace traditional data warehouses for these first two uses.

Indeed, today's big data technologies provide the scope and scale to cover many of

EXHIBIT 2 | Four Big Data Operating Models



Source: BCG analysis.
Note: ETL = extract, transform, and load.
[1]Infrastructure and data analytics services can be sourced internally and/or externally (for example, in the cloud).

the analytics needs of today's enterprises, opening the door to the third, and currently nascent, use of the data lake: *transaction processing,* which is time sensitive and must guarantee the integrity and consistency of data access. Take, for example, a payment made from your bank account—the money needs to leave your account immediately and, likewise, the account balance and the record of payment must be instantly updated. With the emergence and continued evolution of SQL technologies, it may eventually be possible to cover some portion of transaction-processing needs with big data platforms. Hadoop-based big data platforms are still maturing, but the risk of any associated growing pains should not deter a potentially radical architectural vision. A forward-looking implementation allows a company to embrace the true potential of data lakes as they mature.

Select the right operating model. We typically see four key models for implementing a data lake: complement, carve out, transform, and outsource. (See Exhibit 2 for an illustration of the models.)

- With the **complement** model, a company builds a data lake alongside a data warehouse to support new use cases that cannot be accomplished effectively with a traditional data warehouse, such as multistructured data analysis and predictive analytics.

- With the **carve-out** model, companies build a data lake to replace parts of the existing data warehouse solution that are better suited for storage and processing in a data lake. This is typically done to reduce IT costs such as the cost of off-loading expensive ETL development to the data lake.

- In the more radical **transform** model, the data lake progressively replaces the broader suite of relational-database platforms that process data and deliver insights across customer, product, and business management processes. The motivations for such a radical move include transforming a company into a data-driven digital business, increased agility, and improved efficiency.

- With the **outsource** model, a company frees itself from building and maintaining a data lake. It can achieve this by adopting cloud technology and thereby

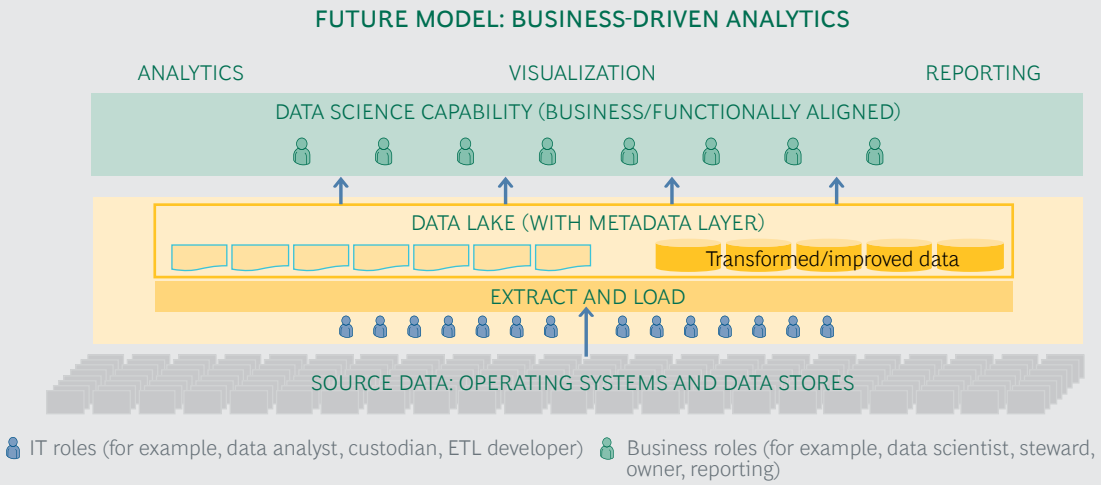
reducing capital investments for infrastructure and specialist skills. Furthermore, companies can leverage analytics as a service by sending data to a vendor, which processes the data and returns results or insights.

Ensure data quality, security, and governance. Regardless of where data is stored, businesses cannot make effective data-driven decisions if their information is not robust, secure, and trustworthy. Leading organizations enforce data governance and data quality policies, processes, tools, and stewardship to ensure that data is fully described and understood and of high quality. (See "How to Avoid the Big Bad Data Trap," BCG article, June 2015.) They also ensure full traceability and lineage of data. Effective implementations make certain that relevant data is sufficiently anonymized and that access is strictly controlled and restricted to relevant authorized users.

Build the organization. With secure, high-quality data in place, companies must next build an organization that is ready and able to make the best use of that data. They need to embed a data consciousness on the business side of the house, adding data scientist roles and skills there, not just in IT, so that their business team can tap the data lake whenever it needs to. This reduces reliance on IT for analytics, visualization, and the production of reports. (See Exhibit 3.) This kind of organization structure enables flexible consumption of data based on business needs, adding agility and subtracting cost. Leading organizations are already positioning themselves for self-service analytics on the business side. In fact, to ensure that they are capturing the utmost value from the data lake, they are building these capabilities company-wide, adding roles and skills that focus on data ownership and stewardship.

DATA LAKES OFFER huge potential to transform businesses. Using them well requires understanding their strengths and limitations and taking a pragmatic approach to implementation. (See "Changing the Game with a Data Lake: An Interview with Centrica's David Cooper and Daljit Rehal," BCG article, September 2016.) Companies that implement data lakes will have the opportunity to build a flexible architecture that enables data delivery at the right time and in the right format. The ultimate goal: accurate and actionable insights that drive business value.

EXHIBIT 3 | Build Data Capabilities on the Business Side of the Organization, Not Just in IT



Source: BCG analysis.
Note: ETL = extract, transform, and load.

About the Authors

Rash Gandhi is a principal in the London office of The Boston Consulting Group and a member of the Technology Advantage practice. His areas of expertise include business-aligned IT strategies and architectures, big data enablement, transformation of application development and maintenance functions, and performance of technical health checks and reviews to derisk programs. You may contact him by e-mail at gandhi.rash@bcg.com.

Sanjay Verma is a partner and managing director in the firm's San Francisco office. He is a core member of the Technology, Media & Telecommunications practice. Prior to joining the firm, Verma launched and led Cloud Labs, a business unit of Flex. He also was the architect of the Oracle TimesTen In-Memory Database. You may contact him by e-mail at verma.sanjay@bcg.com.

Elias Baltassis is a director in BCG's Paris office and a core member of the Technology Advantage and Financial Institutions practices. Prior to joining BCG, he was a founding member and managing director of a world-leading big data company. He has led a broad range of big data and analytics projects in financial services, private equity, retail, and telecommunications. You may contact him by e-mail at baltassis.elias@bcg.com.

Nic Gordon is an associate director in the firm's London office and a core member of the Financial Institutions practice. Prior to joining the firm, he held the senior positions of chief data officer, global head of business intelligence and analytics, global head of data services, and global head of data strategy and architecture for leading banks around the world. You may contact him by e-mail at gordon.nic@bcg.com.

The Boston Consulting Group (BCG) is a global management consulting firm and the world's leading advisor on business strategy. We partner with clients from the private, public, and not-for-profit sectors in all regions to identify their highest-value opportunities, address their most critical challenges, and transform their enterprises. Our customized approach combines deep insight into the dynamics of companies and markets with close collaboration at all levels of the client organization. This ensures that our clients achieve sustainable competitive advantage, build more capable organizations, and secure lasting results. Founded in 1963, BCG is a private company with 85 offices in 48 countries. For more information, please visit bcg.com.

9/16