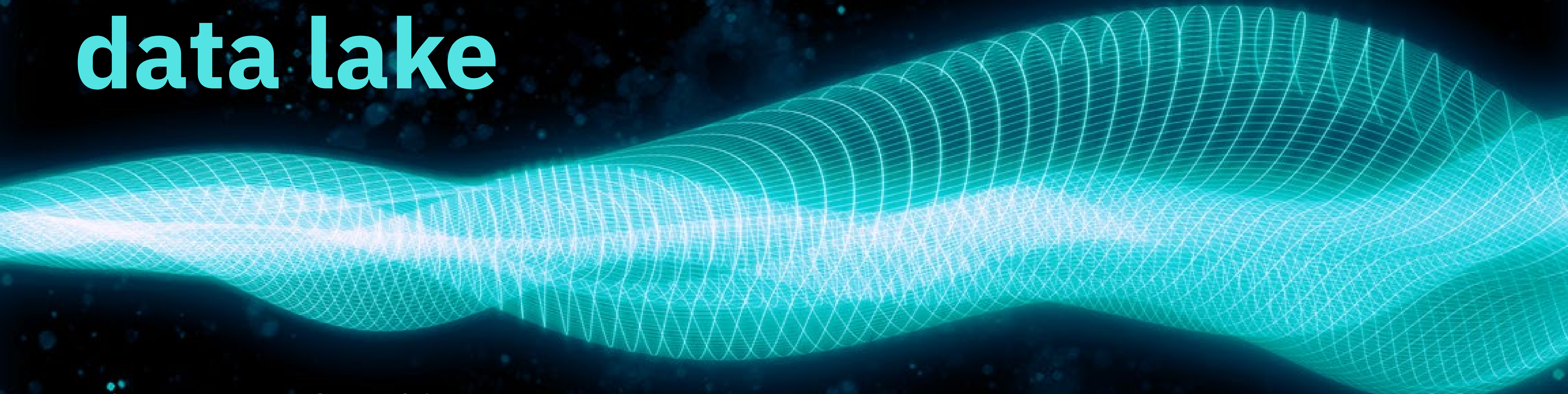
Build a better data lake



Drive smarter, data-driven decisions by capitalizing on a broader variety of data

Contents

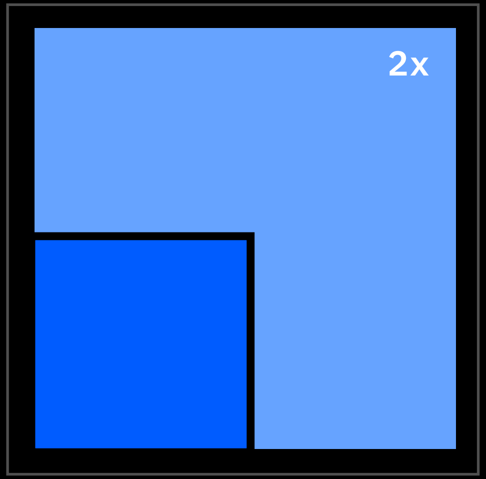
01 Introduction

Data volumes continue to grow at a rapid pace. But data growth is only one part of the big data story. New artificial intelligence (AI), Internet of Things (IoT), mobile and other technologies are generating not only large volumes of data but also a wide variety of data, from a broad array of sources. Organizations can capture customer sentiment expressed on social media, data streaming in from sensors, typed physician notes, weather data, audio from call-center interactions, email correspondence, clickstream data and much more.

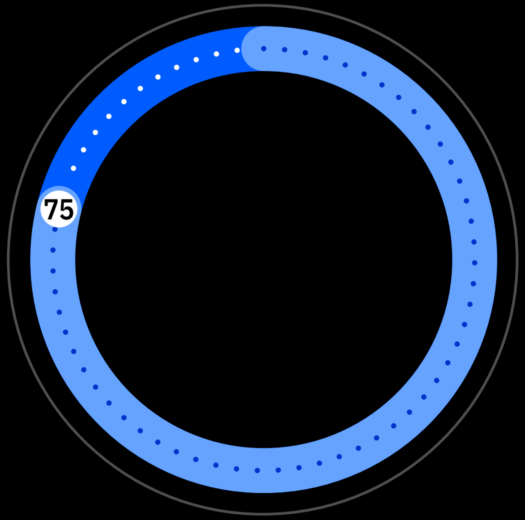
All of this data can be extremely useful for producing new insights that have a real impact on the business. For example, organizations can use this data to better understand their customers, pinpoint ways to enhance operational efficiency, improve fraud detection and identify new marketplace opportunities.

Unfortunately, traditional data warehouses and data marts cannot help organizations capitalize on the large volume and variety of data available today. Designed for structured data, data warehouses are unable to store, query or analyze semi-structured and unstructured data. Because they primarily output predetermined reports, data warehouses also lack the agility to accommodate ad hoc and real-time needs. Relying on data warehouses alone will leave the majority of data unused and unanalyzed.

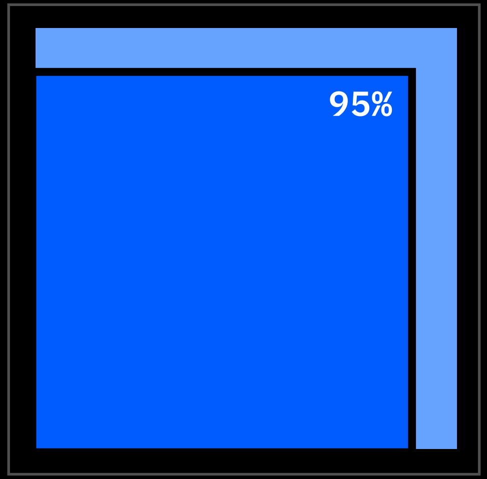
Organizations must find new ways to take advantage of big data. To overcome the limitations of traditional data warehouses, many have begun incorporating data lakes into their data management strategy.



2x increase: Growth in number of enterprise respondents with over 100 TB of unstructured data between 2016 and 2017.[1]



By 2019, 75 percent of analytic solutions will incorporate 10 or more exogenous data sources from second-party partners or third-party providers.[2]



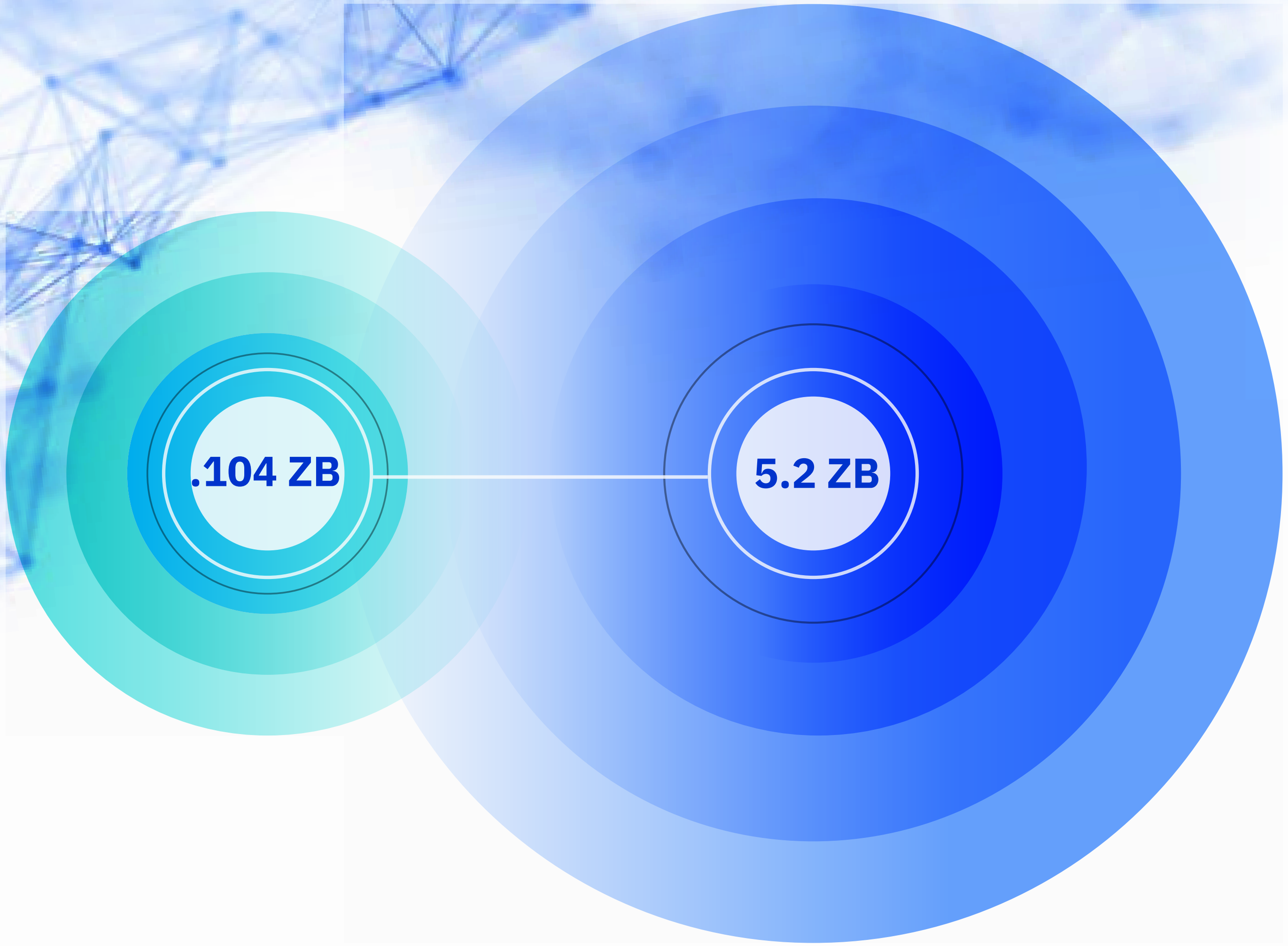
By 2025, real-time **IoT data will make up more than 95%** of real-time data.[3]

[1] Forrester, "Predictions 2018: The Honeymoon for AI Is Over," November 9, 2017.
[2] Crowdflower, "2016 Data Science Report."
[3] IDC for Seagate, "Data Age 2025: The Evolution of Data to Life-Critical," April 2017.

02

What is a data lake and how can it help you?

Data lakes are next-generation hybrid data management solutions that can meet big data challenges and drive new levels of real-time analytics. Their highly scalable environment can support extremely large data volumes and accept data in its native format from a wide variety of data sources. Data lakes can help break down silos, enabling organizations to gain 360-degree views of information and conduct cross-department analytics.



By 2025, data subject to analysis will grow by a factor of **50** to **5.2 ZB.**[4]

[4] IDC for Seagate, "Data Age 2025: The Evolution of Data to Life-Critical," April 2017.

When designed and implemented correctly, data lakes offer five important benefits

1. **Streamlined data preparation**
   By storing data in its original format, a data lake can help reduce the amount of time spent on data preparation.

2. **Simplified data access**
   With a well-built data lake, you can extend access to more users, including not only data scientists but also line-of-business users and application developers. User-defined access allows them to work with data from multiple sources across the organization, on premises or in the cloud.

3. **Enhanced agility for data users**
   A data lake equipped with the proper tools can enable ad hoc queries and real-time analysis—while eliminating the time and costs involved with IT assistance.

4. **Reduced costs**
   Data lakes use commodity hardware, enabling you to scale them cost-effectively without excessive capital expenditures. You can even use a data lake as a repository for older data that would otherwise take up capacity in more expensive warehouses. By providing users direct access to data, data lakes can also help you avoid the cost of IT assistance. In addition, implementing proper data governance capabilities for your data lake allows you to avoid costs associated with correcting data quality issues.

5. **Improved decision-making**
   Analyzing data drawn from more sources lets you increase the depth of insights and enhance the accuracy of results. Governance features that help ensure data is relevant and trustworthy. Real-time analytics and AI capabilities allow you to seize new opportunities as they unfold.

Data lakes can be used in numerous real-world applications across industries. Here are a few examples:

Retail

- Determine what a customer is likely to purchase online and provide recommendations like an intelligent personal assistant.
- Identify a customer's "path to purchase" to understand customer buying patterns and conduct more micro-targeted marketing.
- Predict or proactively identify fraudulent activity from both inside and outside the organization.

Banking

- Predict the success or failure of discounts.
- Pinpoint the "next product to buy" and promote that product to customers.
- Identify which customers are likely to decrease their bank business and employ proactive marketing activities.

Hospitality and travel

- Track and predict customer preferences to guide proactive selling.
- Improve the customer experience and boost brand loyalty through customization and personalization.
- Conduct real-time pricing and analysis.

03 Potential data lake pitfalls

Not all data lakes are created equal. Haphazardly constructed data lakes that lack enterprise-grade capabilities will do little more than provide a repository for a large, disorganized agglomeration of data. These “data swamps” collect data without providing an easy or secure way to locate, access and analyze the information you need.

What pitfalls can lead to a data swamp?

No business case

Without clearly articulating and understanding how a data lake will benefit the business, you might fail to acquire the approvals and buy-in needed to move forward.

Poor integration

A data lake can supplement or in some cases replace a data warehouse. But unless you have a plan for integrated data management, you might not achieve the full value a data lake can deliver.

Wrong technology choices

Selecting the wrong platform or tools can add significant complexity and cost to implementation and ongoing management.

Inadequate governance and security

Enterprise-grade governance and security strategies are critical for protecting sensitive information, maintaining compliance and enabling users to take full advantage of data.

No long-term vision

A data lake requires a long-term commitment plus planning to accommodate continued data growth.

Remember: while a data lake can reduce the amount of the initial data preparation work, it does not eliminate it.

When entering data into a data lake, add robust metadata that describes the data source. Profile and validate the data to confirm its structure, content and quality. Assign responsibility for management, governance and security.

Your data lake strategies must provide for scalability. The volume, velocity and variety of data flowing into the data lake will continue to increase over time. As the benefits of the data lake are demonstrated across the enterprise, you will likely attract a growing number and variety of users. You must be ready to handle these increasing demands on your data lake—and your data lake team—so you can continue to maximize its value for your organization.

04

What is your business case for a data lake?

Assess current resources

Do you have the resources to support your use cases and achieve your business goals? What will it cost to augment these resources? As you identify resource gaps and plan for ways to close them, be sure to factor in the time and costs for:

- New technology, hardware, software and services
- Skilled professionals not currently on staff
- Internal IT support
- User training

Define your business goals

What do you hope to accomplish by analyzing all this newly accessible data? Defining your goals will help you identify high-priority use cases. Identifying use cases will in turn help you narrow down the types of data and data sources to focus on and enable you to start pinpointing the tools and strategy required. For example, if your top priority is monitoring wear and tear on critical manufacturing equipment, you might decide to analyze born-in-the-cloud IoT data in a cloud-based data lake.

Gain approval and buy-in

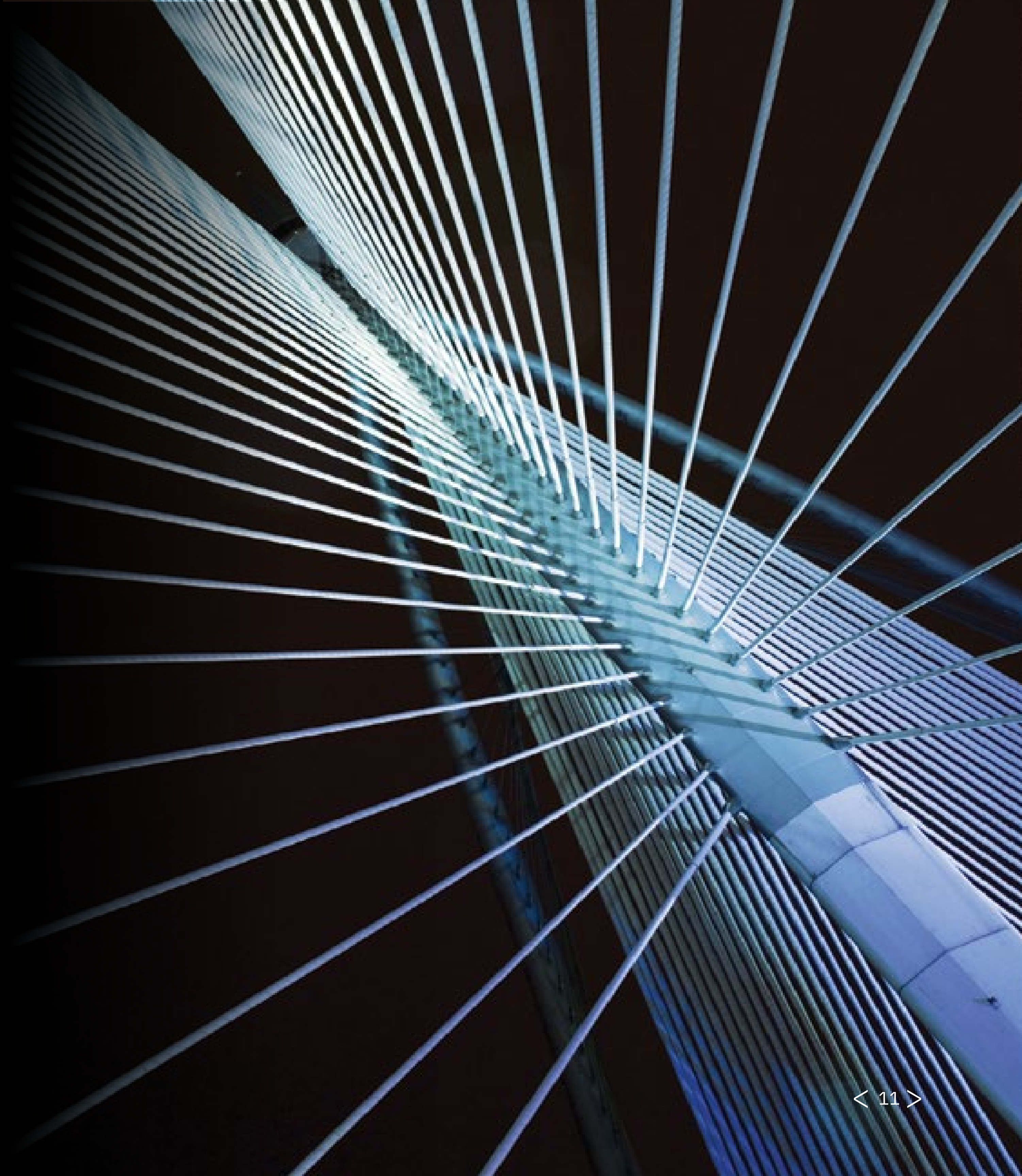
You will need budget approval to acquire any new equipment, hire new people and so on. But you also need to get buy-in from the teams who will use and manage your data lake. Presenting a clear explanation of potential use cases, anticipated results and current resource gaps can help you convince teams to alter their processes so they can make the most of the data lake.

05

How does a data lake fit into your overall data management strategy?

You could establish a data lake that handles all of your data. But because data is ingested “as is,” a data lake might not be as useful as a data warehouse for analyzing traditional structured data. Your existing data warehouse can still serve as a source of extracted, vetted data for typical organizational, historical and financial reporting.

A data lake can complement that data warehouse, enabling you to store, query and analyze additional data types in a more cost-effective environment. In addition, a data lake can provide a less-expensive repository for older data that you no longer want or need to keep in your data warehouse.



Self-sufficiency is essential for all users. By providing simple, fast access to data plus the right tools to query data, you can support ad hoc analyses and foster greater innovation—all while reducing the need for IT assistance.

Line-of-business users in marketing could take advantage of a data lake to develop targeted marketing campaigns, while finance users could identify ways to enhance internal efficiencies. LOB users are responsible for creating summaries and analytic reporting. They want simple tools to access and analyze data relevant to their projects.

Data scientists could run analytics to spot emerging business trends or use predictive analytics to help sales teams determine the next best action for customers. In general, data scientists are the ones who build models and algorithms, create data visualizations and collaborate with business teams to generate new insights from large data sets.

App developers could use a data lake as a test-and-development sandbox for new mobile applications. App developers need to run ad hoc and real-time queries, and integrate multiple data sources from across the organization. They want control over data access with minimal reliance on IT.

06

What technology should you use to build your data lake?

Weigh the pros and cons of Hadoop

Hadoop has become the platform of choice for building data lakes. This highly scalable framework enables processing of very large data sets across hundreds—or thousands—of computing nodes, all operating in parallel. As an open source technology, Hadoop is community-built and community-supported. By using commodity hardware, Hadoop can help drive down costs.

Enterprise-grade Hadoop solutions offered by Hortonworks and other vendors can address common complexity and management limitations and add key security capabilities. With the right solution, you could accelerate data ingestion and use governance tools and security features that help you meet rigorous policies and regulations while capitalizing on real-time data processing and streaming analytics.

Select the right integration tools

If you intend to integrate a data lake with existing data warehouse environments, you'll need the right tools and strategy. Traditional enterprise service bus (ESB) and ETL tools are designed to work with batch processes, not real-time ones, so they cannot handle the low-latency requirements of a data lake. Traditional techniques can result in issues with context, linking and visualizing data—all necessary when mining big data.

Traditional tools also require your staff to understand the tools, the data source and the target data store. But ensuring you have the right expertise on hand can be costly. The right tools facilitate interoperability with your existing environments and reduce the amount of time and effort required for integration. Employing automation capabilities can help control costs and keep data scientists focused on other tasks.

How will you handle data governance and security?

A data lake is a shared platform that can be accessed by many users in a variety of roles. To protect data and maintain compliance with regulations, you need sufficient governance, security and auditing capabilities. Strong governance strategy is critical for complying with regulations and ensuring that data can be readily found, understood and trusted by users. Employing metadata management tools as part of that strategy helps you take full advantage of your data lake.

The right metadata management tools will allow your team to build an index of data assets, add metadata to classify content and trace the lineage of data. This helps your users more easily find what they need and gain confidence that the insights they generate are accurate.

Determine the sensitivity of data

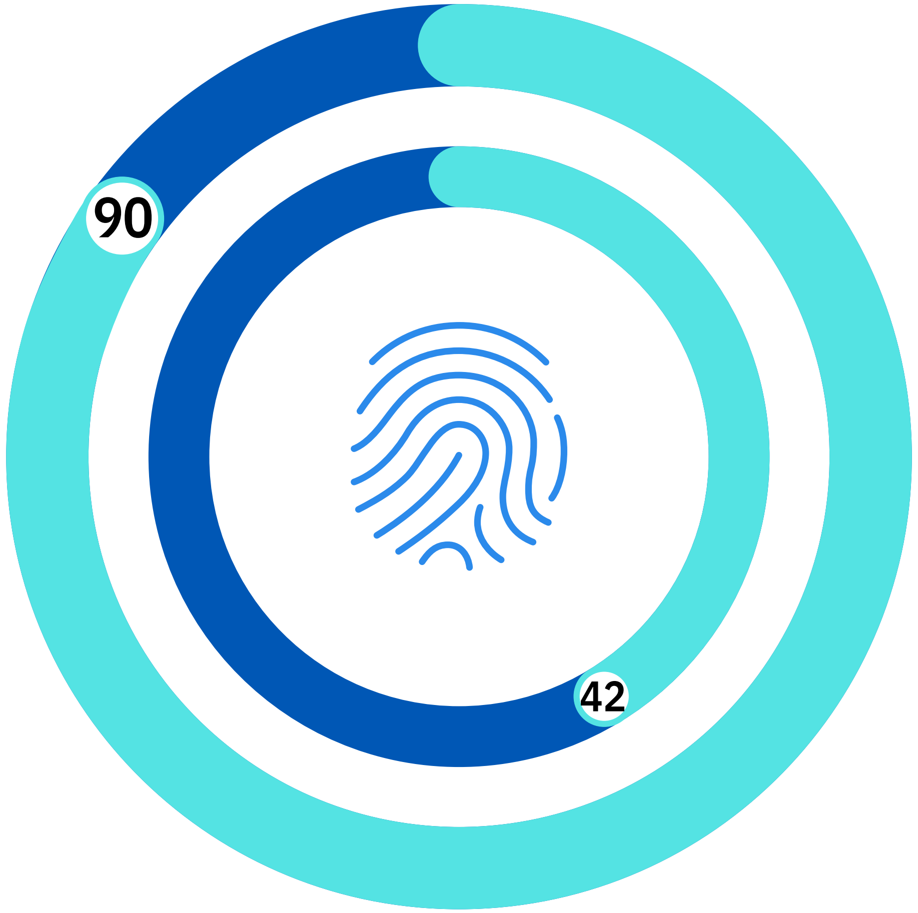
Do you intend to keep customer credit card data, patient healthcare information, corporate financial information, intellectual property or other sensitive data in your data lake? If so, you will need to make sure you safeguard data from loss or theft, and implement governance capabilities that help you comply with regulations.

Define access rights

Which users can access which data? You might decide to allow your marketing team members to query and analyze a wide range of customer data while restricting their access to the company’s financial information. In addition to implementing tools to protect against unauthorized access to certain information, plan to train data lake users on how to follow policies and regulations.

Assign data lake ownership

Who will oversee the data lake? You might decide to assign ownership to your centralized IT group or a smaller IT team that also manages your data warehouse or line-of-business groups. Clearly defining responsibilities will help ensure a smooth deployment and effective management of governance and security policies going forward.



By 2025,

created

in the global datasphere will require some level of security,

[5] IDC for Seagate, “Data Age 2025: The Evolution of Data to Life-Critical,” April 2017.

08

Build your data lake with IBM

A data lake lets you capitalize on large, fast-growing volumes of data—in particular, the unstructured and semi-structured data that is not well supported by traditional data warehouses. Adding a data lake can help you open analytics to new users, generate new insights and increase agility while also streamlining data preparation and controlling costs.

As you begin to evaluate data lake solutions, look for enterprise-grade capabilities that can help meet your rigorous technical and business requirements. Hortonworks and IBM have partnered to offer Hortonworks Data Platform (HDP) and Hortonworks Data Flow (HDF) with the addition of IBM® Db2® Big SQL. These solutions combine cost-effective, enterprise-grade open source technology with real-time analytic capabilities. You can tap into the tremendous potential for previously unanalyzed data and maximize the value of a data lake to make smarter, more agile, data-driven decisions.

Ready to learn more?

Visit ibm.com/analytics/data-management/data-lake

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
April 2018

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

16014716-USEN-00