GOVERNING YOUR DATA LAKE:

WHAT EVERY ORGANIZATION SHOULD KNOW

collibra®

“WATER, WATER, EVERY WHERE, NOR ANY DROP TO DRINK.”

In “The Rime of the Ancient Mariner,” Coleridge famously describes the agony of sailors in a becalmed sea surrounded by water they’re unable to drink.

To be sure, data lakes hold an ocean of possibility for organizations eager to put analytics to work. But for business users thirsty for information and frustrated by a data lake overflowing with ungoverned data, Coleridge's words might sound eerily prophetic.

Today, data is a differentiator. Organizations that can gather, manage, and use reports, dashboards, metrics, and models to create real value for their customers and stakeholders are thriving.

At a time when the quantity, type, and sources of data are proliferating, data lakes make a lot of sense. Organizations can capture vast amounts of data from disparate platforms – including social and IoT – and put it to use faster and with less expense. That's good news for organizations pursuing transformative new data initiatives.

But the truth is, not every data lake delivers on its potential and many organizations are still struggling to find value in their investments.

WHEN IT COMES TO YOUR DATA LAKE, IS YOUR ORGANIZATION DYING OF THIRST?

You have data – lots of it. And some of it is accessible, meaningful, and trustworthy. But generating accurate reports seems to get more difficult every quarter.

Sometimes business users can't find the data they need. Or the data they do find is gibberish. Sometimes conflicting data definitions make finding the "right" data impossible. The data you need to keep your organization healthy and hydrated remains out of reach, hidden in the depths of the data lake.

If you recognize these symptoms, it might be time to ask some questions about the state of your data lake.

Do you know what's in your data lake?

Data lakes are tremendously scalable and flexible, accommodating all types of data and lots of it. That's great – until it isn't. These same attributes make it very easy for organizations to lose track of their data. Are you able to identify what data, reports, metrics, and algorithms are in your lake and index them a meaningful way?

Do you know what should be in your data lake?

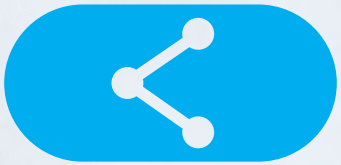
Getting data into a data lake isn't difficult. But building a data lake is not a data strategy. A data lake is one (very useful) tool that can help organizations pursue a data strategy. But without data definitions, even the best tool becomes ineffective. Does your organization have a larger vision driving the data lake? If not, it will quickly become a data dumping ground.

Do you know where your data comes from and where it's been?

We might think of the data lake as a still pond where data simply sits until we're ready to use it. But data is dynamic and understanding its lineage is critical to quality, context, and use.

Does your organization have visibility into the data journey?

How will they know if it's (still) certified?



Must users keep it internal – or can they share it with partners?

Do users know what happens if the data, report, or definition they are using changes?

Are the appropriate data sharing agreements in place to answer all of these questions – and more?

Do you know who has access to what data?

Lots of data likely means that lots of people will want to discover, curate, and prepare data sets to glean new insights. Can you provide policy-based, granular access to your data lake, however your organization defines it?

Are people regularly using the data?

Adoption problems often plague data lakes – largely because finding good, trustworthy data is just too hard. Are the people in your organization still thirsty for information, but hesitating to jump into the data lake? Are there other, different data sets that you should bring into the lake to encourage them to dive deeper? And are users searching for – but not finding – what they need?

Does your data lake encourage collaboration?

One reason organizations typically implement a data lake is to overcome data silos. How well is your data lake helping people across your organization – and your organization's data chain – to solve problems?

WARNING, DATA SWAMP AHEAD

If you can't answer these questions, you're not alone. As early as 2014, Gartner raised **a red flag**, cautioning that "without at least some semblance of information governance, the lake will end up being a collection of disconnected data pools or information silos all in one place."

A recent **survey** conducted by TDWI found that a lack of data governance was the **biggest roadblock** to data lake deployments. Without governance, businesses are throwing away the potential of their data to solve pressing business problems.

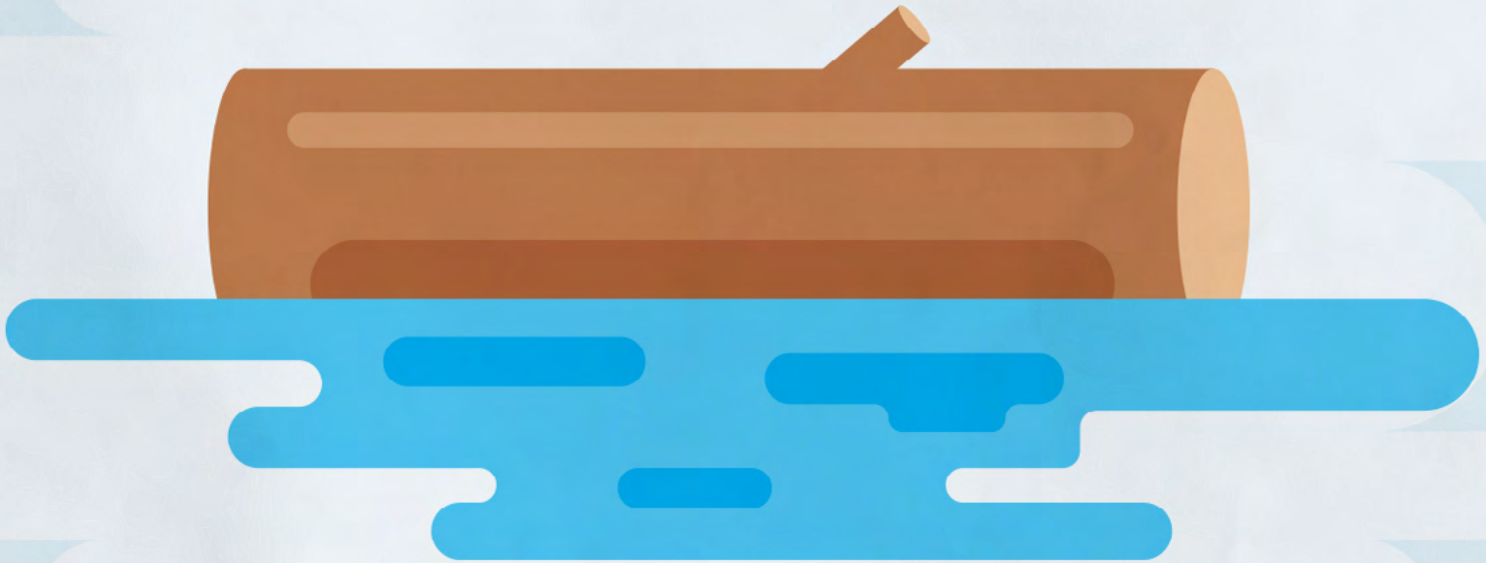
And while some in your organization may be able to maneuver through the data lake today, that tribal knowledge will deteriorate as your organization grows, as employees move on to other projects or to other enterprises, and as the sheer volume of your data increases.

Without governance, and a governed data catalog, your data lake can get murky fast.

SEEING BELOW THE SURFACE

Data lakes allow organizations to collect and store remarkable amounts of data – both structured and unstructured – from within and beyond the enterprise. But questions still remain. How do we make this data meaningful? How do we use it with consistency and clarity? How do we build integrity into the conclusions we draw from the data we now have available to us? And how do we make it consumable to the users who need it most?

A good data governance framework, complete with a data catalog, can help you keep your data lake pristine by avoiding these four common logjams that too often muddy the waters.



1

Data without context

A data catalog helps people understand the data they find by providing information about that data – its relationship to other data; its origin, format, and use; its lineage; and information about how it is organized, classified, and connected. It provides data users with a common language for understanding their data in business terms. And it uses ontology to define the relationships and associations through an enterprise-wide business glossary.

2

Data that can’t be found

A data catalog organizes and structures data to help people find the information they need to solve business problems.

3

Data that can't be trusted

A data catalog can help data users find the best data for their purposes, understand the quality of that data, and know whether it's appropriate to join data from disparate sources.

4

Data that can't be shared

A data catalog makes it easier for people to work collaboratively with transparency and trust, enriching data sets and driving value across – and beyond – the enterprise.

Simply put, a governed data catalog can save you from drowning in your data lake.

PLUMB THE DEPTHS OF EVEN THE DEEPEST DATA LAKE

Stop casting a wide net and trawling for data. A governed data lake provides a policy-driven process to help data users surface meaning with more precision.

A data catalog encourages data users across your organization to work together to understand the data's meaning and use. It defines rules and operating models for how data should be ingested. It provides context for how it should be understood.

It helps users determine which data is fit for purpose and which data needs to be thrown back because it's unusable, incomplete, or irrelevant.

And it provides a way for every user to find data, understand what it means, and trust that it's correct.

Think of what a governed data catalog might make possible for your data lake. What if you could...

Approve data sources	Flag sensitive data	Establish secure access protocols
Alert people when data relevant to their work is ingested	Index and tag data so that people will understand it	Preview sample data to determine its usefulness
Help your users understand the data journey, its source, lineage, and transformations	Determine how new data projects might impact downstream processes and reports	Easily “shop” for data, directly from the data lake
Help people discover other related data sets that they might not know about	Profile the data to assess if it’s fit for purpose	Ingest metadata, speeding up the process from ingestion to use
	Match columns in your data set to business concepts	

DIGITAL TRANSFORMATION ACROSS THE ORGANIZATION

By applying data governance principles to their data lake initiatives, leading organizations have improved supply chain management, human resources, and marketing functions. Let's take a look at how they did it.

Understanding supply chain disruptions

Supply chain disruptions are becoming increasingly common as supply chains themselves grow more complex. Identifying where – and why – the supply chain breaks down is key to improving quality and service. But understanding, predicting, and preventing damages across the supply chain requires quickly parsing vast amounts of data collected from disparate sources – most from outside of the enterprise.

One global company recognized the possibilities a data lake could provide, and used it to break down organizational silos. By pulling their diverse data sets into a data lake, they were able to analyze where damages occur and how they can be prevented. Data governance policies developed by a cross-functional team allow data users to find indexed data faster, understand its lineage, and see associated reports – improving analysis and speeding resolution.

Now, finding the point where complicated supply chains fail is much easier and reaction time is improved.

Managing talent in an evolving organization

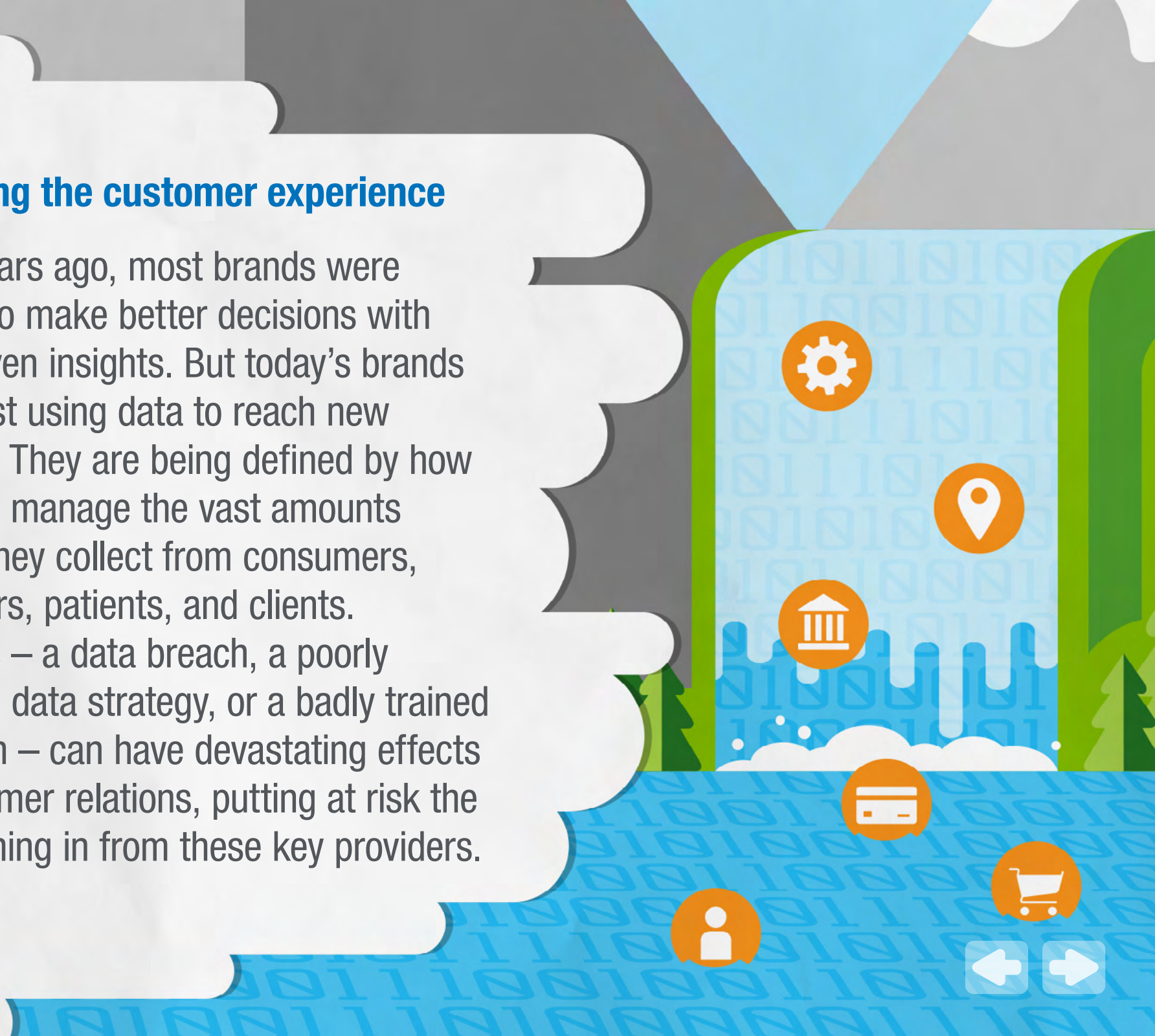
Managing human capital is hard enough. Managing those resources during one, or even multiple, mergers or acquisitions has its particular challenges: rationalizing job descriptions, determining salaries and benefits, and evaluating incentive plans. Even coming to a common understanding of how many employees the new organization employs is no longer a simple query. And, of course, these all are questions that need answers sooner rather than later.

Rather than trying to integrate incompatible HR systems, one organization decided to develop a new HR platform that sits on top of a common data lake. A data governance layer allows IT, HR, and its partners to collaborate on data policies and definitions, implement robust new data privacy rules, and work collaboratively to resolve issues.

Now everyone across the organization has a shared context for understanding conflicting data sets – an important step in managing talent more strategically.

Improving the customer experience

A few years ago, most brands were striving to make better decisions with data-driven insights. But today's brands aren't just using data to reach new markets. They are being defined by how well they manage the vast amounts of data they collect from consumers, customers, patients, and clients. Missteps – a data breach, a poorly executed data strategy, or a badly trained algorithm – can have devastating effects on customer relations, putting at risk the data coming in from these key providers.



For one leading organization, data governance is helping them shift the conversation from lock-down to access. Through better visibility and understanding of data usage, they are able to expertly use their data to improve the customer experience. They're embedding data privacy (GDPR) principles into their governance, which allows for certain sensitive data to be used in customer facing activities. Data that would otherwise be locked away due to a lack of clarity on its purpose and legal basis for using it in marketing and CRM.

Now, they are enriching their analytics with more external sources to improve marketing. And they are using governance to build confidence with their community of data providers. They're constantly expanding the pool of data and insights, resulting in a clear competitive advantage.

The key to achieving these results is trustworthy data. Without that trust, the data lake, no matter how abundant its sources, will eventually run dry.

REALIZING DATA'S POTENTIAL

Data, big or small, collected in a data lake or streaming through the Internet of Things, is simply raw material. Shaping it into a dynamic tool that can drive innovation and transform organizations can only be achieved when that data's meaning and use are clear and when it can be trusted by the people who use that data every day to drive knowledge.

Without governance,

data lakes, whether narrowly defined or broadly conceived, will fail to yield consistently trustworthy insights.

With governance,

a data lake can become more than a data repository. It can become a valuable business asset that delivers the kind of information your business users need to accelerate decision making, discover better business processes, and inspire ground-breaking discoveries.

When it comes to your data lake, which will you choose?

collibra®

©2018 Collibra