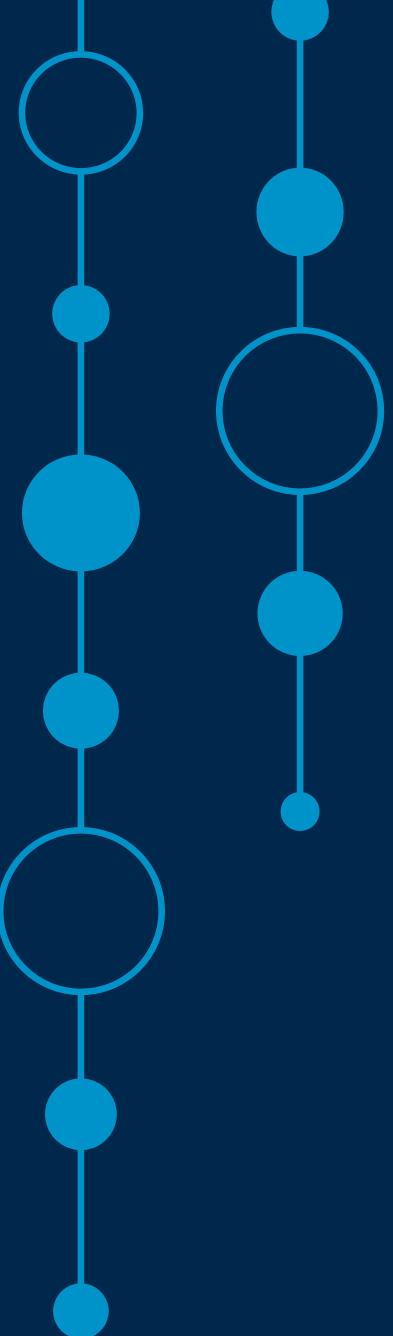


DEFINITIVE GUIDE TO

# CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES





## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES

### TABLE OF CONTENTS

#### INTRODUCTION

**CHAPTER 1:** WHAT IS A CLOUD DATA WAREHOUSE?

**CHAPTER 2:** CLOUD DATA WAREHOUSES VS CLOUD DATA LAKES:  
WHAT IS THE RIGHT SOLUTION?

**CHAPTER 3:** HOW DO I SET UP MY NEW DATA WAREHOUSE  
OR CLOUD DATA LAKE?

**CHAPTER 4:** HOW TO OPTIMIZE YOUR CLOUD DATA WAREHOUSE

**CHAPTER 5:** ENSURING DATA QUALITY BEFORE IT ENTERS YOUR SYSTEMS

**CHAPTER 6:** ENDING YOUR DATA SILOS

**CHAPTER 7:** TRENDS IN CLOUD DATA WAREHOUSING

**CHAPTER 8:** YOUR THREE-STEP SOLUTION TO MAKING YOUR CLOUD DATA  
WAREHOUSE OR CLOUD DATA LAKE WORK

**CHAPTER 9:** CASE STUDIES

**CHAPTER 10:** INTEGRATION SOLUTIONS CHECKLIST

**CHAPTER 11:** WHAT'S THE TAKEAWAY?

## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: INTRODUCTION

Right now, the world is all about data. The companies that can use their data most effectively are the ones who are winning in the marketplace. Data is a strategic asset; it is one of the most critical drivers of business. The reasons are simple. Data tells you which products to build, which business models to pursue, which customer experiences to create based on your successes and the successes of others.

It's no wonder, then, that the companies that base their business on data have a competitive edge. But the actual process of becoming a data-driven company isn't easy. Turning raw data into insights is a complex challenge to solve. And that challenge is exacerbated by the fact that the pace of change is changing. In fact, the pace of doing business is accelerating. Things that you used to expect to do in a week or overnight or an hour – now have to be done in real time. And at the same time, there's more and more innovation coming at a faster and faster speed.

The challenge for most enterprises is: how can I take advantage of that change? How can I see all the change and innovation as an opportunity that I can unleash?

The answer lies in collecting and analyzing data, but not merely as a technical exercise to be performed by IT and then placed in some repository somewhere never to be seen again. Becoming a data-driven company has to become a business strategy, believed in and implemented not just by technical teams, but by employees all throughout the business. Which is why a data warehouse strategy or a data lake strategy isn't just a technical project; rather, it's a business strategy vital to the continued thriving of the entire organization.

***Forrester reveals  
that insights-driven  
businesses are growing  
at an average  
of more than  
30% each year.***



*Enterprises have  
hundreds of systems  
without master data.*

*A cloud data warehouse  
does not fix a data structure  
problem.*

### **A data warehouse strategy is a critical business strategy**

When we look at most organizations today, we see a complex data landscape.

There is on-premises data. There is data in the cloud. There is data that's being exchanged with customers and suppliers. There is data coming from new sources like sensors and IoT devices. Sometimes these things are added one at a time, sometimes they're added in bulk, and you might end up with an architecture that scales out beyond your control.

Companies end up lacking the ability to get all of the data into one place where they can analyze it and actually get the appropriate value out of it. But it's not just being able to accommodate new sources of data; there are now further demands on the data being used by the enterprise. Data repositories have to deal with how fast the data is coming in. More governance needs to be applied to data; it has to be compliant with data protection regulations like GDPR and CCPA.

If companies really want to base mission-critical decisions on their data, that data needs to be clean and of good quality. These demands are becoming more and more important and at the same time, companies are trying to move ever faster to keep up with the demands of the market and just stay competitive.

## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: INTRODUCTION

But simply building a data warehouse is not the end of the problem. You might end up with a data mart here or a data warehouse over there. There might be different projects that were begun but perhaps never got finished, or perhaps they weren't properly maintained over time. Also, as the data volumes and the complexity increase, the cost of storage continues to go up, further increasing costs.

Plus, organizations have to do all this capacity planning. How much hardware am I going to need? How many servers am I going to need to support this new volume of data? And then, is that setup actually going to meet performance demands as the data scales up? What do I do if I don't need all of that capacity anymore? All of that capacity planning and the cost of the software and all of the hardware is a lot just to get something basic to function.

There's a great deal of evidence that the way the market is approaching this challenge is changing.

A recent survey by the [Data Warehouse Institute](#) noted that almost half of the respondents said that they were going to do a replacement project for their data warehouse product. These would be rip and replace data warehouse initiatives to replace on-premises data warehouses with cloud ones – not a small undertaking.

**Nearly half  
of surveyed businesses  
said they were planning  
to replace their  
on-premises data  
warehouses  
with cloud ones.**



# DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES:

## INTRODUCTION

### Why data warehouses must move to the cloud

The on-premises data warehouse business is inexorably shrinking. Most new customer data warehouses being built today are being built in the cloud (commonly Snowflake, AWS Redshift, Azure SQL Data Warehouse, or Google BigQuery). Putting your data repository in the cloud is simply better. It's faster, more scalable, with zero install time, you can go live in minutes, and it's always up to date. Nearly every single company looking for a new data warehouse or a new data lake will choose a cloud-based data repository.

Companies are making the shift to these warehouses because they immediately get access to all the hardware resources they might need to scale that solution. And as they add data, compute resources, or memory – whatever it is that they need to scale – they're really paying only for what they use. Plus, they can easily add best-of-breed SaaS applications to their cloud-based data repositories, so they can run their whole analytics or business intelligence stack as a service. They can scale up and down as desired with very little installation or maintenance costs. It is clear that cloud-based data warehousing is where this market will be going.

According to a [TDWI and Talend survey](#), the top reasons companies migrate to a cloud data warehouse are:

- ▶ A flexible cost model
- ▶ To take advantage of cloud features
- ▶ Faster performance
- ▶ To migrate existing products to cloud

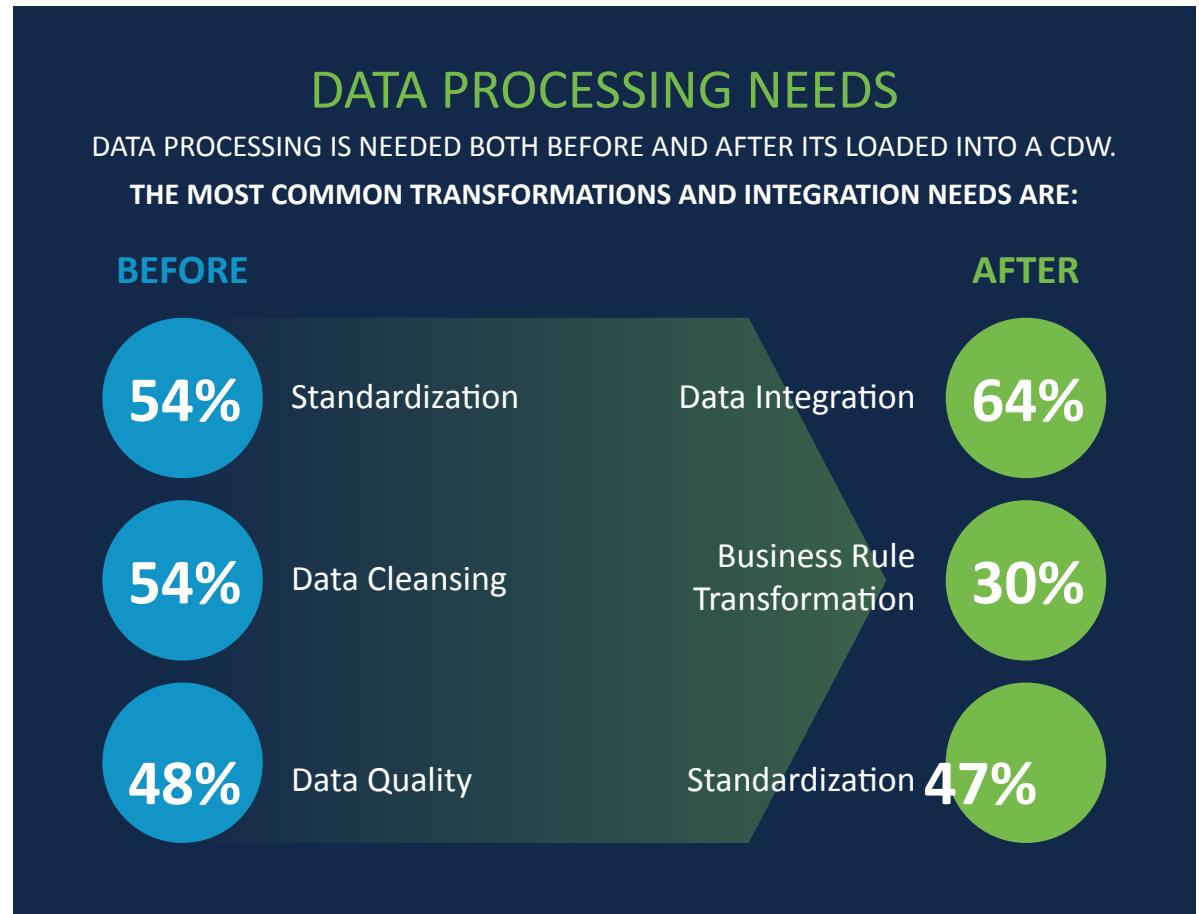
# DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: INTRODUCTION

## The challenges of cloud-based data warehouses

Cloud data warehouses are clearly part of the future of enterprise data. But that isn't where the story ends. A cloud data warehouse, as useful as it is, is an empty shell. How are you going to get your enterprise data in there, and once it's in there, available to be used by any employee or system that needs it? How will you account for all of the different types of data coming in at incredibly fast speeds? How will you be able to add new data sources with the same instantaneous availability as all the others? How will you be able to make sure that your data is good quality and compliant with data protection regulations?

You might be tempted to think about legacy data integration tools. But many of these tools are simply not compatible with nor optimized for a lot of the new data platforms. They might not have a connector for Redshift or Azure SQL Data

Warehouse or Snowflake. Tools that weren't built in the cloud-based world won't be able to get started with, easily interoperate with, or scale with the kinds of data warehouses companies want to migrate to. A number of these legacy tools have stopped development on connectors – or don't release connectors quickly enough – to these new platforms. Or, if you're hand-coding connections from your systems to these warehouses, your team might have to learn all over again how to hand code to this new platform. Hybrid platforms are becoming increasingly important for many companies, as regulatory or other restrictions, still require their legacy on-premises systems. A number of legacy data integration tools won't work in a hybrid infrastructure.



## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: INTRODUCTION



And, of course, you will want to introduce data quality or data governance capabilities. As your projects get more complex, perhaps for machine learning or other types of more advanced analytics, you will want to offload some of that processing and large volumes of data to automated systems.

### Optimizing your cloud data warehouse for great business outcomes

Businesses are investing time, money, and resources into ripping and replacing their old data repositories with new cloud data warehouses. It's important to get the most out of that investment.

This guide will outline:

- The advantages and disadvantages of cloud data warehouses vs cloud data lakes, and which one your company needs.
- How to set up your cloud data lake or cloud data warehouse
- How to move your data from on-premises systems to the cloud safely and securely
- How to make sure you're getting the best data – and therefore the best insights – from your data in the cloud

**Companies need to turn raw data into insight-ready data to stay ahead of the competition. This guide will help you overcome the challenges of this process and put your company on the path to become truly data-driven.**

*“Cloud data warehouses are uniquely poised as a key repository for digital transformation. Their innate value is driven by their data, but the data must be comprehensive, trustworthy, and consumable.”*

*– Vincent Lam, director of cloud product marketing, Talend*



# CHAPTER 1

## WHAT IS A CLOUD DATA WAREHOUSE?

---



## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES:

### WHAT IS A CLOUD DATA WAREHOUSE? CHAPTER 1



Data drives businesses. That's been true for some time. There are many applications such as ERPs, CRMs, and Supply Chain Management systems that various lines of business use to manage their business processes such as finance, human resources, or customer relationship management (to name a few).

While using best-of-breed SaaS applications has many benefits for the business, this approach creates a number of problems when it comes to storing and accessing data:

- These applications create silos of unconnected data that do not talk to each other.
- There is no single version of the truth which combines data in a standard way; people try to collaborate with one another with different key metrics.
- There is limited ability to read the large amount of data needed for analytics; transactional systems are good at writing data but not good at reading lots of data

This problem was solved by traditional data warehouses and data lakes providing a means of consolidating data into repositories that were once unthinkable or untenable. Data warehouses have made the ability to aggregate data from different sources into a high performance central trusted repository. Data lakes have provided a low-cost means of using traditional off-the-shelf hardware to amass large amounts of data across a wide variety of formats.

**Why is it important for these data warehouses and data lakes to be in the cloud?**

While data warehouses and data lakes are meant to be simple in concept, their implementations are not. Setting up any of these has been extremely difficult and in the domain of experts. The promise and the reality don't necessarily match. In the data warehouse world, simplification came in the form of "appliances" because data warehouse implementations were so complicated. Just buy a box, vendors promised, and don't worry about installation.

These platforms hardware their physical presence. To increase capacity, you had to upgrade to a bigger box. In fact, you often had to buy an oversized box since you needed to have capacity in case your needs grew.

There were other limitations to traditional data warehouses as well. How do you back them up? What about downtime when your expensive hardware breaks? How do you stay agile when the infrastructure is so fixed and costly? How do you keep up with innovation when procurement is so slow?



*In 2020,  
there will be  
40 times more  
bytes of data  
than there will be  
observable stars  
in the sky.*

## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: WHAT IS A CLOUD DATA WAREHOUSE? CHAPTER 1

Cloud data warehouses and cloud data lakes promise insight-ready data without complexities or constraints. You don't have to worry about any of the challenges above. With a couple of clicks, you're ready to go. You can build as big or small as you want. You can pay for what you need. You're able to upgrade or downgrade instantly; the platform automatically upgrades itself.

The ecosystem that feeds and consumes the data in your warehouse or lake is going to the cloud as well. In today's market, the expectation is that anybody can set up shop in the cloud, whether that's a project, department, or a whole company. If your ecosystem is in the cloud, the cloud benefits apply to the entire organization. If your data warehouse is on-premises, the warehouse would destroy the benefits cloud offers to your organization.



**The ever-larger appetite for data means that the use cases for enterprise cloud data warehouses is expanding:**



Machine learning



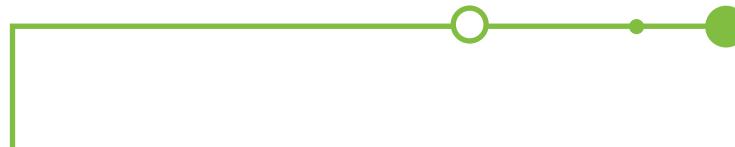
Advanced analytics



Artificial intelligence



Applications



# CHAPTER 2

## CLOUD DATA WAREHOUSES VS CLOUD DATA LAKES: WHAT'S THE RIGHT SOLUTION?

---





*In 2025,  
most people can  
expect to have 5000 digital  
interactions per day,  
up from 700-800  
in 2019-2020.*

Data lakes and data warehouses are both widely used for storing big data, but they are not interchangeable terms. Traditionally, data warehouses are more focused on structured data and data lakes can accommodate both structured and unstructured data.

The two types of data storage are often confused, but do serve different functions. The main similarity between them is their high-level purpose of storing data. While a data lake works for one company or use case, a data warehouse will be a better fit for another. Typically, a data warehouse is used for more structured purposes like reporting. Data lakes are more of a giant inbox where you eventually need to pull out the relevant data and info to make it useful. A common use case is feeding the data warehouse from a data lake.

It's important to note that the lines are blurring between the two more and more as each platform becomes more capable. There is, in fact, a gray area where either will suffice. In fact, data warehouses are practical replacements for data lakes in some cases now. But there are still certain determinations you can make to discover whether a cloud data lake or a cloud data warehouse is right for you.

# DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: CLOUD DATA WAREHOUSES VS CLOUD DATA LAKES CHAPTER 2

## FOUR KEY DIFFERENCES BETWEEN A DATA LAKE AND A DATA WAREHOUSE

There are several differences between a data lake and a data warehouse. [Data structure](#), ideal users, processing methods, and the overall purpose of the data are the key differentiators.



### Data Structure



### Purpose of Data



### Users



### Accessibility

	Data Lake	Data Warehouse
<b>Data Structure</b>	Raw	Processed
<b>Purpose of Data</b>	Not yet determined	Currently in use
<b>Users</b>	Data scientists	Business professionals
<b>Accessibility</b>	Easier to ingest and update unstructured data	More difficult to ingest and update data due to the need for uniformity and structure, but easier to consume

### Preventing your data lake from turning into a data swamp

One of the challenges with creating a data lake is that the unstructured data within it can become difficult to find and use, like a data swamp.

To prevent this, here are the [three rules and three phases of implementing your data lake](#):

- 1 Properly implementing your ecosystem, data models, architecture, and methodologies
- 2 Incorporating exceptional data processing, governance, and security
- 3 Deliberately using job design patterns and best practices

When implementing your data lake, you must manage the three phases of its lifecycle:

- Ingestion
- Adaptation
- Consumption



### Data structure: raw vs. processed

Raw data is data that has not yet been processed for a purpose. Perhaps the greatest difference between data lakes and data warehouses is the varying structure of raw vs. processed data. Data lakes primarily store raw, unprocessed data, while data warehouses store processed and refined data.

Because of this, data lakes typically require much larger storage capacity than data warehouses. Additionally, raw, unprocessed data is malleable, can be quickly analyzed for any purpose, and is ideal for machine learning. The risk of all that raw data, however, is that data lakes sometimes become data swamps without appropriate data quality and data governance measures in place.

Data warehouses, by storing only processed data, save on pricey storage space by not maintaining data that may never be used. Additionally, processed data can be more easily understood by a larger audience.

### Purpose: undetermined vs in-use

The purpose of individual data pieces in a data lake is not fixed. Raw data flows into a data lake, sometimes with a specific future use in mind and sometimes just to have on hand. This means that data lakes have less organization and less filtration of data than their counterpart.

Processed data is raw data that has been put to a specific use. Since data warehouses only house processed data, all of the data in a data warehouse has been used for a specific purpose within the organization. This means that storage space is not wasted on data that may never be used.





### Users: data scientists vs business professionals

Data lakes are often difficult to navigate by those unfamiliar with unprocessed data. Raw, unstructured data usually requires a data scientist and specialized tools to understand and translate it for any specific business use.

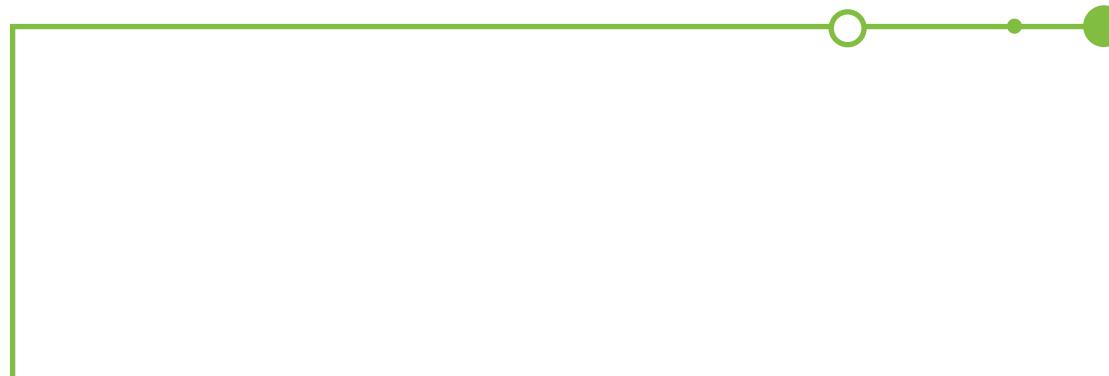
Alternatively, there is growing momentum behind data preparation tools that create self-service access to the information stored in data lakes.

Processed data is more consumable by tools for analytics and BI because of its well-defined structure.

### Accessibility: flexible vs secure

Accessibility and ease of use refers to the use of the data repository as a whole, not the data within it. Data lakes have no structure and are therefore easy to access and easy to change. Plus, any changes that are made to the data can be done quickly since data lakes have very few limitations.

Data warehouses are, by design, more structured. One major benefit of data warehouses is that the processing and structure of data makes the data itself easier to decipher. However, the limitations of structure make data warehouses difficult and costly to manipulate.



# DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES:

## CLOUD DATA WAREHOUSES VS CLOUD DATA LAKES CHAPTER 2

### Data lake vs data warehouse: which one is right for me?

Organizations often need both. Data lakes were born out of the need to harness big data and benefit from the raw, granular, structured and unstructured data for machine learning. But there is still a need to create data warehouses for business users to perform analytics. Here are three typical use cases which determine whether a cloud data warehouse or cloud data lake is necessary:

#### Healthcare



Data lakes store unstructured information

Data warehouses have been used for many years in the healthcare industry, but they have never been hugely successful. Because of the unstructured nature of much of the data in healthcare (physicians' notes, clinical data, images and scans, etc..) and the need for real-time insights, data warehouses are generally not an ideal model.

Data lakes allow for a combination of structured and unstructured data, which tends to be a better fit for healthcare companies.

#### Finance



Data warehouses appeal to the masses

In finance, as well as other business settings, a data warehouse is often the best storage model because it can be structured for access by the entire company rather than a small data science team. Big data has helped the financial services industry make big strides, and data warehouses have been an important player in those strides.

#### Transportation & Logistics



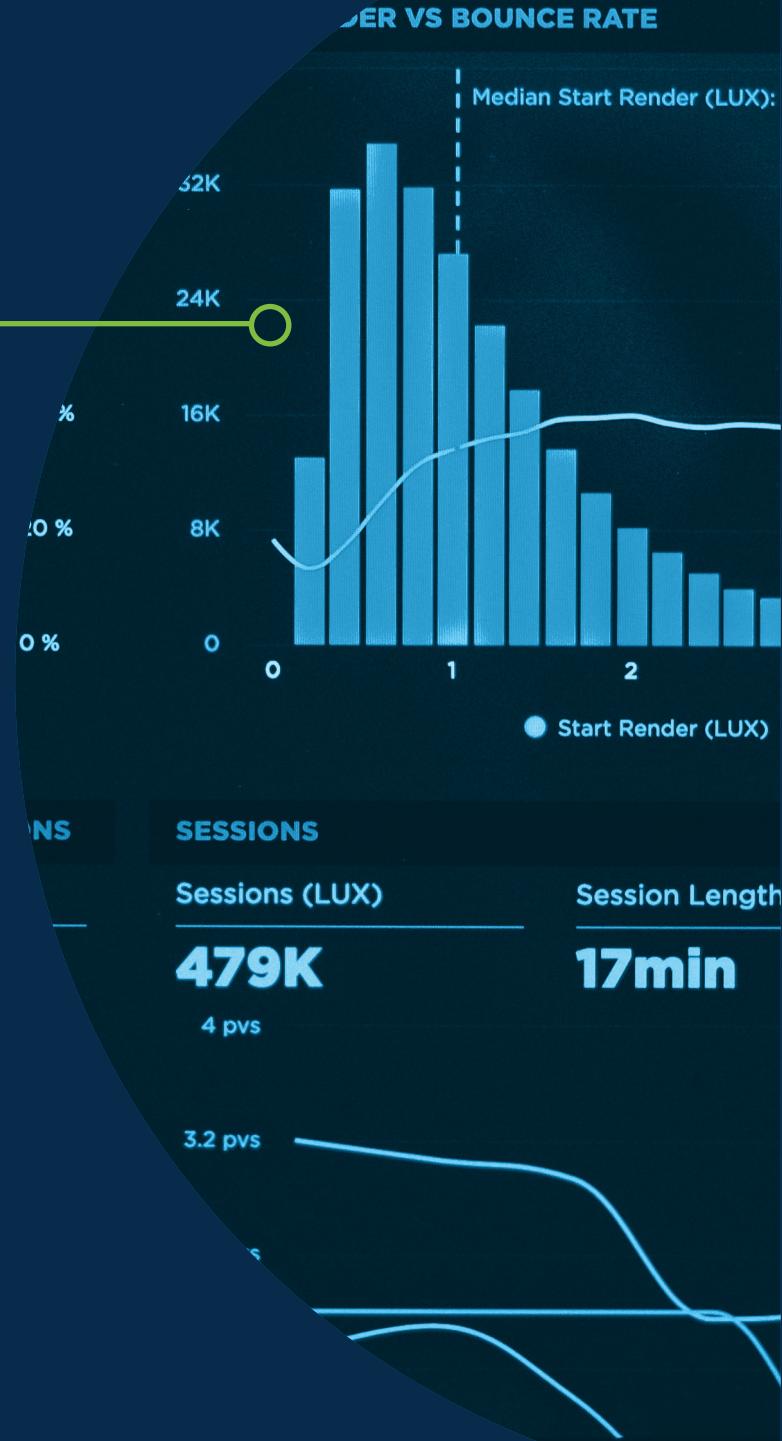
Data lakes help make predictions

Much of the benefit of [data lake insight](#) lies in the ability to make predictions.

In transportation or logistics companies, especially in supply chain management, the prediction capability that comes from flexible data in a data lake can have huge benefits, e.g. cost-cutting benefits realized by examining data from forms within the transport pipeline.

# CHAPTER 3

## HOW DO I SET UP MY NEW DATA WAREHOUSE OR DATA LAKE?



## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: HOW DO I SET UP MY NEW DATA WAREHOUSE OR DATA LAKE? CHAPTER 3

The great thing about setting up a cloud data warehouse or cloud data lake is that getting the actual infrastructure in place is very easy. All you need is a web browser and a credit card, and you can just purchase the cloud infrastructure as a service.

But before you buy yourself a brand-new data warehouse or data lake, you have to ask yourself some key questions:

- Who is going to use the data? And what for?
- How are you going to get data into it?
- How are you going to protect the data in your repository and make sure it's good quality while still making it available for everyone?



IT decision makers surveyed by Talend and TDWI say their top three challenges in kicking off their cloud data warehouse projects are:

- Data governance
- Integrating data from multiple sources
- Data security



*“Data governance enables benefits at every management level by enabling and improving the processes around the creation, management, and use of data. Strategic benefits include aligning business needs with technology and data, better customer outcomes, and a better understanding of the organization’s competitive ecosystem.”*

*Jeff Tyzzer,  
customer success architect, Talend*

# DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: HOW DO I SET UP MY NEW DATA WAREHOUSE OR DATA LAKE? CHAPTER 3

## Who will be using the data? And what for?

If you want to become a data-driven company, you have to provision data to everyone who wants to use it for business intelligence, not just a select few. One issue that many organizations have discovered on their journey to becoming more data-driven is that they have to deal with specialists to extract information out of their data platform. This might be because their data infrastructure was built on-premises, or even as an early consequence of cloud migration.

But getting to the data through a specialist team became a bottleneck, no matter what the function was: line of business developers, data scientists, data business analysts, financial analysts, even executives. If they wanted a cut of the data suited to their function, they would have to wait for the data team to get to it. And that creates a delay that most organizations simply don't have time for. Business today has to accelerate time-to-value and speed up the entire value chain.



*"We don't want IT to be a bottleneck—we want it to be an enabler for self-service. We have a mountain of data, and we want people to be able to retrieve and use it themselves."*

- Rene Grenier, Vice President for Data Integration, Uniper SE.

## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: HOW DO I SET UP MY NEW DATA WAREHOUSE OR DATA LAKE? CHAPTER 3

That's why a modern cloud data warehouse or cloud data lake approach will often provide self-service access to data. Cloud data warehouses offer direct access into your data platform by all functions. Your data team is still there to orchestrate your warehouse, perhaps setting up policies which can be individualized per your enterprise needs. However, once the infrastructure is established and a policy is put in place, any function can go about the business of looking at their data, with the proper security and access controls that will be employed. Self-service access means that you're able to achieve faster time to value, you're able to eliminate bottlenecks from the value chain of looking at or ingesting and looking at data and creating insights.

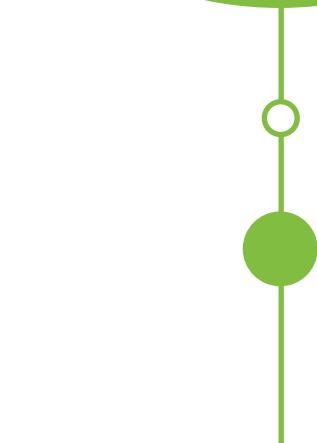
In parallel with this, it's not only about self-service access to data, it's also about the data infrastructure being expected to adopt to multiple skill levels. Every professional who needs data cannot be expected to write scripts to extract data and build insights. That would slow the business down. With a self-service type of approach, you must accommodate all functional and skill levels; from the basic program manager all the way up to perhaps an advanced technical practitioner. If you imagine that access to data should be enterprise-wide, and should accommodate all skill levels, you'll find that you'll be able to accelerate results within your environment.

A data lake, because of its use of unstructured data, may not have the same possibilities for self-service access. Extracting value out of a data lake then becomes a matter of thinking about how to process and manage its data at scale so it can be used easily.



***“Talend helps visualize the data flow with little or no experience in coding.”***

- IT Architect, US Federal Government



## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: HOW DO I SET UP MY NEW DATA WAREHOUSE OR DATA LAKE? CHAPTER 3

**How are you going to get data into your data lake or data warehouse?**

**And what will you use it for?**

Changes in the velocity and variety of data have fundamentally changed the nature of how data actually gets into data lakes and data warehouses. Once upon a time, getting data into a data warehouse involved queuing jobs. If you had a master job, perhaps you might schedule it at night. If you had data loading, sometimes that would have to be scheduled in off hours because you didn't want to have disruption with a BI environment that was active in the day. Or perhaps you might have batching from your corporate OLTP system early in the morning. You've shut down that job and then some time in the morning you have resources available to do the executive dashboards, which tend to run very actively first thing in the morning once everyone arrives. Then throughout the day you might orchestrate your resources to ensure that you can handle queries from BI teams and analysts and so forth. You might push off large jobs towards the end of the day if you're a data scientist and the streaming and loads can be at midnight.

The problem with this is that today everyone wants access to their data faster. People want to be able to get to their data sooner and then, in turn, they're able to evaluate more data at their fingertips. And it's important to be able to customize workloads for your specific business needs. A modern approach to data warehousing can have a variety of different workloads allocated to different warehouses, and then those workloads can be overlapping. You don't have contention for resources because each of the workloads are isolated – they have their own resources – yet at the same time they're sharing and accessing the same data with full integrity and consistency. This allows you to enable people to query the data when they need it, to get access to the data as soon as it's available and it avoids long delays. The sky's the limit for the most part, because with cloud data warehousing, you can spin up as many warehouses as you can dream up and this will still apply.

*With cloud data warehousing, you can spin up as many warehouses as you can dream up, and still allow customized access for anyone who needs it.*

## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: HOW DO I SET UP MY NEW DATA WAREHOUSE OR DATA LAKE? CHAPTER 3



When you have this kind of flexibility in terms of when the data is available, then it's important to not limit yourself to thinking about what the use cases for your data are. Obviously, analytics is a huge one and people certainly leverage data warehousing for that, but if you take a look at the inherent technology and the benefits of the cloud, there is no reason to keep your use case tied to one thing. When you have a flexible, modern approach to cloud data warehousing, the most important capability you need along with it is the ability to integrate and pull in all types of data you probably never would have thought of before. Everything from enterprise apps and mobile, to the lab, to APIs, the internet of things and having them all go through in a very consistent and governed way. This gives you unprecedented data power. There are so many more things you can do beyond analytics; for example, you can work with real-time applications, or self service applications for your customers.

***1 in 5 surveyed  
IT decision makers  
confirmed there are  
more than 100 sources  
of data in their  
organizations.***

Then, as you start adding more and more data sources and you take advantage of all that activity, there is no limit to what you can do. One of the things that should happen with this modern approach to cloud data warehousing is that this process of ingesting data and making it available should be extremely easy, and that is a big advantage of cloud. So even though you need to incorporate different types of data, it's important to let your data integration solution take care of that so you don't have to think about it.

You can use a cloud data warehouse for all its great benefits without having to worry about how that stuff gets in there in the first place. A good data integration solution should have hundreds of connectors and components that make it really simple to connect any application or source of data – SaaS, on-premises, or hybrid. A modern approach to data warehousing, with a robust integration solution alongside it, allows you to broaden your horizons and think about tying together things that you might not have even considered it possible to integrate, taking advantage of your complete data infrastructure.

### How will you ensure good quality data in your data warehouse or data lake?

Data quality is probably the most important – and most underrated – factor in your data stack. Data quality absolutely matters because inherently it's the difference between right and wrong. Your data must be trustworthy because you're going to leverage all this data later on; you want to definitely make sure that it is clean and correct.

As previously mentioned, you're pulling in different types of data into your data warehouse through your integration solution, you need to best leverage all that data. But what happens in between? Two things must happen:

#### **1) You need to be able to govern your data.**

Governance in this context means trying to isolate and find data that's important to the organization. It helps identify the key data in the organization that we want to be pulling in in the first place and then define who will know if this data is right or wrong.

#### **2) You need to ensure that it's clean and accurate.**

Once you've defined the standards, you know exactly what good data actually looks like. Now it can be automated. Automated data quality systems can make it absolutely clean so that everything that goes into your cloud data warehouse or cloud data lake is completely trustworthy and usable.

For more information about exactly how this can be done, take a look at the

[Definitive Guide to Data Quality](#).

*What is the best way to migrate to a cloud data warehouse or cloud data lake?*

A typical migration scenario involves lift and shift – everything you have and moving it to the cloud in a 1 to 1 replacement. This is the cleanest migration and leaves you with no legacy. The only drawback is everything that touches your warehouse has to change at the same time.

Sometimes ripping off the Band-Aid is difficult. So, it's also common to run data repositories on-premises and in the cloud simultaneously. By running projects both in the cloud and on-premises, you can try the cloud and see how it works for you. A practical way to do this is with new projects.

They're relatively self-contained so they are easier to implement. Over time, the goal is to move more existing use cases over so that eventually you're 100% in the cloud.

# CHAPTER 4

## HOW TO OPTIMIZE YOUR CLOUD DATA WAREHOUSE

---



## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: HOW TO OPTIMIZE YOUR CLOUD DATA WAREHOUSE CHAPTER 4

The phrase “data warehouse optimization” sounds like such a complicated undertaking. Data warehouses can often be cumbersome and complex systems that can store terabytes and even petabytes of data that people depend on to make important decisions on the way their business is run. The thought of any type of tinkering with such an integral part of a modern business would make even the most seasoned CIO break out into a cold sweat.

However, the value of optimizing a data warehouse isn’t in dispute. Minimizing costs and increasing performance are mainstays on the to-do lists of all Chief Information Officers. But that is just the tip of the proverbial iceberg. You’ve got to maximize availability, increase data quality, limit data anomalies, and eliminate depreciating overhead. These are the challenges that become increasingly more difficult to achieve when you’re stuck with unadaptable technologies and confined by rigid hardware specifications.

**According to a Talend and TDWI survey, the top three capabilities needed by companies switching to a cloud data warehouse are:**

- Supporting ANSI SQL
- Supporting structured and unstructured data
- Supporting in-memory processing

### The data warehouse of the past

Not long ago many of today's technologies (i.e. Big Data analytics, Spark engines for processing and cloud computing and storage) didn't exist, yet the reality of balancing the availability of quality data with the efforts required to cleanse and load the latest information proved a constant challenge. Every month, IT was burdened with loading the latest data into the data warehouse for the business to analyze. However, often the loading itself took days to complete and if the load failed, or worse, the data warehouse became corrupted, recovery efforts could take weeks. By the time last month's errors were corrected, this month's data needed to be loaded.

It was an endless cycle that produced little value. Not only was the warehouse out-of-date with its information, but it was also tied up in data loading and data recovery processes, thus making it unavailable to the end user. With the added challenges of today's continuously increasing data volumes, a wide array of data sources and more demands from the business for real-time data in their analysis, the data warehouse needs to be a nimble and flexible repository of information, rather than a workhorse of processing power.



*If your company is not yet a data-driven company, it's likely transforming into one by relying on a diverse set of data sources: NoSQL and relational SQL data stores, SaaS and legacy applications, and even IoT data.*



## Today's data warehouse needs

In this day and age, tech executives can rest easy knowing that optimizing a data warehouse doesn't have to be so daunting. With the availability of Big Data Analytics, lightning-quick [processing with Apache Spark](#), and the seemingly limitless and instantaneous scalability of the cloud, there are surely many approaches one can take to address the optimization conundrum.

The most effective approach to simplifying data warehouse optimization (and providing the biggest return on investment) is to remove unnecessary processing (i.e. data processing, transformation and cleansing) from the warehouse itself as well as the integration solution, which is where Hadoop and Spark come in. By removing the inherent burden of ETL processes, the warehouse has nearly instantaneously increased availability and performance. This is commonly referred to as "offloading ETL".

Another approach is to push the workload to the CDW using an ELT approach. Also known as "ELT pushdown", this approach leverages the compute scalability of the cloud warehouse to speed up the time data is in flight while allowing you to preserve the data at the source.



*A cloud data warehouse must support a variety of technical use cases. These include: accessing data in the CDW for analytics, ingesting data from the cloud and on-premises to a CDW, and transforming and processing data.*

## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: HOW TO OPTIMIZE YOUR CLOUD DATA WAREHOUSE CHAPTER 4

This isn't to say that the data doesn't need to be processed, transformed and cleaned. On the contrary, [data quality is of utmost importance](#). But relying on the same systems that serve up the data to be responsible for processing and transforming the data is robbing the warehouse of its sole purpose, which is to provide accurate, reliable and up-to-date analysis to end-users in a timely fashion, with minimal downtime. By utilizing Spark and its in-memory processing architecture, you can shift the burden of ETL onto other in-house servers designed for such workloads. Or better yet, you can shift the processing to the cloud's scalable infrastructure and not only optimize your data warehouse, but ultimately cut IT spend by eliminating the capital overhead of unnecessary hardware.

Optimizing a data warehouse can surely produce a fair share of challenges. But sometimes the best solution doesn't have to be the most complicated. Data integration tools need to be as nimble as the systems they are integrating. Therefore, leveraging a future-proof architecture means you will never be out of style with the latest technology trends, giving you peace of mind that today's solutions won't become tomorrow's problems.



***“With Talend, data is much easier to validate, clean, and transform before it enters the data warehouse.”***

- Senior IT architect,  
*financial services company*



## 5 steps to successful data storage and management

### 1) Scale for tomorrow's data volumes

The amount of data available is vast, and it's only growing by the day. You'll need to consider how your data lake will handle current as well as future data projects. That means ensuring you have enough developers, as well as processes in place, to manage, cleanse, and govern hundreds or thousands of new data sources efficiently and cost-effectively, without affecting performance.

### 2) Focus on business outcomes

You can't transform your enterprise if you don't understand what's most important to the business. Understanding the organization's core business initiatives is the key to identifying the questions, use cases, analytics, data, and underlying architecture and technology requirements for your data lake.

### 3) Expand the data team

Data quality is increasingly becoming a company-wide strategic priority involving individuals from different departments, rather than merely the IT team. With bad data often impacting business analysts, involving business users in your data quality process makes sense. Business analysts have the domain knowledge and skills to choose the right data for business needs, and by providing them with self-service access, you help ensure your data lake fulfills some of its key objectives.

### 4) Future-proof your infrastructure

Business needs are constantly changing, so your data lake will likely need to run on other platforms. Since different teams within the same organization often use different cloud providers based on their needs and resources, most companies operate in a multi-cloud infrastructure.

If this is the case in your organization, you'll need to make sure your data infrastructure can handle that by opting for a flexible strategy that allows you to maintain agility as your technology choices change. A [data vault methodology](#) that gives you the flexibility to continuously onboard new types of data is often a sound approach.

### 5) Create a data governance strategy

Don't wait until after your data lake is built to think about data quality. Having a well-crafted [data governance](#) strategy in place from the start is a fundamental practice for any big data project, helping to ensure consistent, common processes and responsibilities. Start by identifying business drivers for data that needs to be carefully controlled and the benefits expected from this effort. This [strategy](#) will be the basis of your data governance framework.

# CHAPTER 5

## ENSURING DATA QUALITY BEFORE IT ENTERS YOUR SYSTEMS

---



Bad data has never been such a big deal. Why? According to IDC's report, "Data Age 2025", the projected size of the global data sphere in 2025 would be the equivalent of watching the entire Netflix catalog 489 million times (or 163 ZB of data). In a nutshell, the global data sphere is expected to be 10 times the 2016 data sphere volume by the year 2025. As the total volume of data continues to increase, we can also infer that the volume of bad data will increase as well unless something is done about it.

### How to spot bad data

Bad data can come from every area of your organization under diverse forms from business departments, sales, marketing or engineering. Let's take a look at a few common categories of bad data:

- **Inaccurate:** data that contain a misspelling, wrong numbers, missing information, blank fields
- **Non-compliant:** data not meeting regulatory standards
- **Uncontrolled:** data left without continuous monitoring becomes polluted over time
- **Unsecured:** data left without control and vulnerable to access by hackers
- **Static:** data that is not updated and becomes obsolete and useless
- **Dormant:** data that is left inactive and unused in a repository lose its value as it's neither updated nor shared

**If data fuels your business strategy, bad data could kill it**

If data is the gasoline that fuels your business strategy, bad data can be compared to cheap, poor quality gas. There is no chance you'll go far and fast if you fill the tank with the bad stuff. This same logic applies to your organization. With poor data, results can be disastrous and cost millions.

According to [Gartner](#), poor data quality cost rose by 50%, reaching 15 million dollars per year, for every company. You can imagine this cost will explode in the upcoming years if nothing is done.

**Time for a wake-up call**

Results from the third Gartner Chief Data Officer (CDO) survey show that the data quality role is again ranked as the top full-time role staffed in the office of the CDO<sup>4</sup>. But the truth is that little has been done to solve the issue. Data quality has always been perceived by organizations as a difficult play. In the past, the general opinion is that achieving better data quality is “too lengthy” and “complicated.” Over the last two years, data quality tooling and procedures have dramatically changed. It’s time for you to take the data bull by the horns.

*The average financial cost of bad data is \$15 million per year.*



**Let's take a closer look at a few misconceptions about data quality:**

***“Empower people for data curation, and remediation.”***

Today, data is coming from everywhere, and data quality tools are evolving. They are now expanding to cover any type of data no matter their type, their nature, or their source. Faced with data complexity and growing data volume, modern data quality tooling uses machine learning and natural language processing capabilities to ease up your work and separate the wheat from the chaff. Solving data quality downstream, at the end of the information chain, is difficult and expensive. It's 10x cheaper to fix data quality issues at the beginning of the chain than at the end.

***“Once you solve your data quality, you're done.”***

Data management is not a one-time operation. To illustrate, let's look at the example of social networks. The number of social media posts, video, tweets, and pictures added per day is in excess of several billion entries. This rate only continues to increase at lightning speed. It's also true for business operations. Since data is becoming more and more real time, you need “in-flight data quality”. Data quality is becoming an always-on operation, a continuous and iterative process where you constantly control, validate and enrich your data, smooth your data flows and get better insights. You also simplify your work if you link all your data operations together in a single managed data platform.





***“Data quality is IT’s responsibility”***

Gone is the time when data was simply an IT function. As a matter of fact, data is now a major business priority across all lines of business. A security breach, data loss or data mismanagement may lead your company to bankruptcy. Data is the entire company’s priority as well as a shared responsibility. No central organization - whether it’s IT, infosec, or the office of the CDO - can magically cleanse and qualify all the data.

Bad data management has immediate negative business consequences – penalties, a bad reputation, or lost revenue. Good data management requires company-wide accountability.

***“It’s hard to control data quality”***

Data management isn’t just a matter of control anymore, but a matter of governance. IT should understand that it’s better to delegate some data quality operations to the lines of business because they’re the data owners. Business users then become data stewards. They feel engaged and play an active role in the whole data management process. It’s only by moving from an authoritative mode to a more collaborative role that you will succeed in your data strategy.



# CHAPTER 6

## ENDING YOUR DATA SILOS

---



## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: ENDING YOUR DATA SILOS **CHAPTER 6**

True data-driven organizations seek to extract all the insight from all their data to optimize every aspect of their business and better serve their customers. They collect and analyze more and more data from traditional sources, such as ERP, CRM, and point of sale systems, as well as from newer data sources such as logs, web applications, Internet of things (IoT) devices and more.

However, that's only possible with a single repository to easily and efficiently store all your data and make it useful. But it's not feasible or practical to load all data into a traditional data warehouse.

Data from newer sources often arrives in semi-structured formats; requiring additional transformation and processing before loading. Further, the cost and complexity of storing large quantities of raw, unrefined data in a traditional data warehouse from an increasing number of sources would be prohibitive.



*by 2025, 49%  
of all the world's  
stored data will  
be in public cloud  
environments.*

## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: ENDING YOUR DATA SILOS **CHAPTER 6**



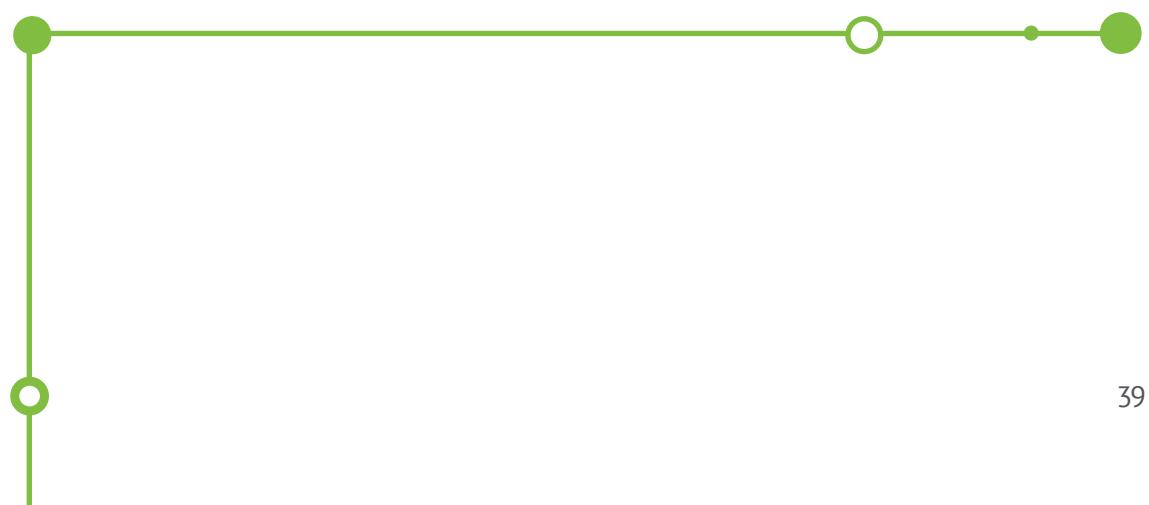
The data lake emerged more than a decade ago to solve this problem: how to create a scalable, low-cost data repository for storing raw data from a diverse set of sources to explore and refine that data. Then, move subsets of the refined data to other systems, including a data warehouse, to support high-performance analytics and reporting.

But that simple idea is not so simple in practice. Without the right technology and without proper data quality and data governance, a data lake can all too easily become a data swamp—an isolated pool of data difficult to use, hard to understand, and practically inaccessible to most of the organization. The greater and more diverse the data put into the data lake, the more significant the problem becomes, making it harder and harder to get meaningful insights and value from that data.

Your company relies on a diverse set of data sources: NoSQL and relational SQL data stores, SaaS and legacy applications, and even IoT data. These data sources have different formats, data models and structures. You'll need modern platforms and tools that make managing and consolidating all this data as simple as possible.

The legacy of decades of different types of software, both on-line and on-premises, has led to so many proprietary data silos. There's no standard way to use everything. Some are better than others. However, the reality is that you will encounter difficulty in getting the data you want into your warehouse or lake.

This is why connectivity is such a critical – and often overlooked – necessity of an effective cloud data warehouse or data lake. Cloud data lakes and warehouses have very limited connectivity out of the box, and usually that is in a mere handful of formats like AVRO or JSON, etc. Lakes and warehouses don't expect to be interfaced directly by the consumer, but the volume and variety of data means that there needs to be a technology that feeds the data in and something that pulls the data out into applications so it can be consumed. As it comes in, data may need to undergo the same steps of being profiled, normalized, aggregated and cleansed. Accelerating data loading is an essential step in uniting diverse data sources.



The best practices for creating a solid platform for data ingestion include:

- Connecting to various data sources easily, without major programmatic scripting, so you can collect all your data from wherever it's located.
- Batch and streaming ubiquity—handle historical and real-time data loading and process data pipelines as they come in.
- Scale with volume and variety—Quickly onboard new data sources, such as data from third-party data providers, web clickstreams, social media, and smart devices
- Flexible with new technology – You might want to apply machine learning or AI to your data as you're loading it in. A good integration solution will allow you to seamlessly couple these steps so that you can use something like Spark for heavy duty data manipulation and also be smart enough to provide ELT to pushdown compute to a data warehouse.

## 4 questions you must ask about your cloud data lake or cloud data warehouse's connectivity capabilities:

### 1 Does your integration tool have ETL and ELT?

Sometimes data needs to stay raw, and other times you need more processing. Therefore, your data integration solution must be able to cope with both ETL and ELT capabilities, where data transformation can be handled either before the data is loaded to your final target, e.g. a cloud data warehouse, or after data has landed there. ELT is more often leveraged when the speed of data ingestion is key to your project, or when you want to keep more intel about your data.

### 2 Can your data lake or data warehouse handle both simple ETL tasks and complex big data ones?

Not all of your data lake usage will be complex, requiring advanced processing and transformation. Many use cases can be simple activities such as ingesting new data into your data lake. Often, the tasks go beyond the capabilities of the data engineering or IT teams. Ideally, therefore, the tool of your choice should be able to handle simple tasks quickly and easily, but also scale to the complexity to meet the requirements of advanced use cases. An integration tool that can cope with both can help you make your data lake more consumable and practical for various types of users and different purposes.

### 3 Can your data warehouse or lake handle both streaming and batch data?

Streaming data has become a part of our everyday lives whether you realize it or not. If your business is using social media, IoT, or sensor data, that is streaming data. In IDC's 2018 Data Integration and Integrity End User Survey, 93% of the respondents indicate the plan to use streaming technology by 2020. Real-time and streaming analytics have become a must for modern businesses today to create a competitive edge. Can your data lake handle both your batch and streaming needs? Do you have the technology and people to work with streaming, which is fundamentally different from typical batch needs?

### 4 Can your data lake strategy accommodate a collaborative data culture?

The workflow in your data lake should be able to be reused and leveraged among data engineers. A self-service culture of access to data means there will be less re-creation of work and operationalization can be much faster. A modern approach to a cloud data lake or cloud data warehouse can help improve the collaboration between IT and business teams. For example, your line of business teams are the experts on their own data and they know the meaning and the context of data better than anyone else. Data quality can be much improved if the business team can work on the data for business rule transformations, within parameters set by the IT team.

# CHAPTER 7

## TRENDS IN CLOUD DATA WAREHOUSING

---



## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: TRENDS IN CLOUD DATA WAREHOUSING CHAPTER 7

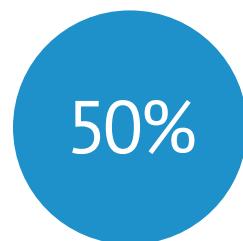
TDWI and Talend asked over 200 architects, IT and Analytics managers, directors and VPs, and a mix of data professionals about their cloud data warehouse strategy. We wanted to understand how cloud data warehouses fit into a cloud migration strategy:

- Are cloud data warehouses (CDW) seen as a key driver of digital transformation?
- Which use cases are driving CDW adoption?
- Does a cloud data warehouse help companies become more data-driven?

Survey respondents noted that their cloud data warehouses need to do complicated work. They have to cross hybrid environments as well as accommodate a larger organizational shift to the cloud. In addition to all that, respondents wanted their cloud data warehouse to work for functions throughout the company, not just a select few technical teams.

The cloud data warehouse environment is getting more complex. The majority (36%) of the respondents indicated that they would be deploying their CDW in a hybrid environment.

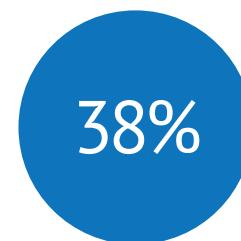
However, there are still major road blocks for organizations to adopt CDWs successfully that go beyond CDWs themselves. Getting data into the CDW is only the beginning. Besides the complexity of getting various data ingested into a CDW, there are many more major challenges. The top challenges indicated by the survey respondents are:



Governance



Integrating data across  
multiple sources



Getting data into  
the warehouse

## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: TRENDS IN CLOUD DATA WAREHOUSING **CHAPTER 7**



Meanwhile, the needs organizations have to perform data analytics in a CDW are increasingly complex. Companies require a number of additional processing types and methodologies, the top 3 including in-memory processing, supporting structured and unstructured data, and integration with 3rd party analytics tools.

Organizations are fully on board with adopting a cloud data warehouse, but recognize that what a cloud data warehouse must do has evolved with changes in cloud computing,

automation, machine learning, and other important trends. A CDW is no longer seen as an end in itself, but rather a stage in the data-driven journey, which has to involve managing a data lifecycle, ensuring data quality, providing a data governance framework, among other considerations.

Take a look at [more results](#) on survey respondents' experience with adopting cloud data warehouses.



*CDW is no longer seen as an end in itself, but rather a stage in the data-driven journey, which has to involve managing a data lifecycle, ensuring data quality, and providing a data governance framework.*

# CHAPTER 8

## YOUR THREE-STEP SOLUTION TO MAKING YOUR CLOUD DATA WAREHOUSE OR CLOUD DATA LAKE WORK

---



# DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: YOUR THREE-STEP SOLUTION TO MAKING YOUR CLOUD DATA WAREHOUSE OR CLOUD DATA LAKE WORK **CHAPTER 8**

## STEP 1: DEFINE YOUR DATA INTEGRATION AND DATA CURATION STRATEGY.

Many people fail to see that their data integration strategy spans beyond a simple integration project. Instead, your data integration strategy is incredibly important because it will determine your ability to grow as a data-driven organization, which often determines your ability to act on data-driven insights before your competitors. As your company's needs grow, your data needs will grow as well—and you need to make sure that your strategy takes that growth into account.

**While putting together your data integration strategy, here are a few questions to ask yourself:**

### **1) What are the long-term goals of your department and your company beyond your initial data integration project?**

Many companies see a data integration project as just the first step on the way to something much bigger. For example, you may be looking to move your Salesforce data into a cloud data warehouse today, but the end goal may be to have a master data management system that maintains the “golden record” of all of your customers. Make sure that the people choosing and implementing your integration technology know enough to make a choice that will be smart for the business today and tomorrow.

### **2) Do you want to embrace emerging technologies (even the ones you haven't heard of yet)?**

Most people would like to think that they will take advantage of any new technology that will bring their business some tangible improvement. However, some data integration strategies—like manually building your integration stack—will be less capable of enabling you to use new technologies without an enormous amount of development time or procuring a new tool. If you know that you will want to have flexibility in your data environment and the technologies you leverage, you will need to have an integration strategy that is flexible enough to handle those changes easily. You will need to look at the underlying architecture of different integration platforms in order to see whether they are built to easily integrate with new technologies. Open source integration technologies, for example, tend to be able to work with new data technologies very easily—especially since many of the new innovative data technologies are based on open source projects themselves.

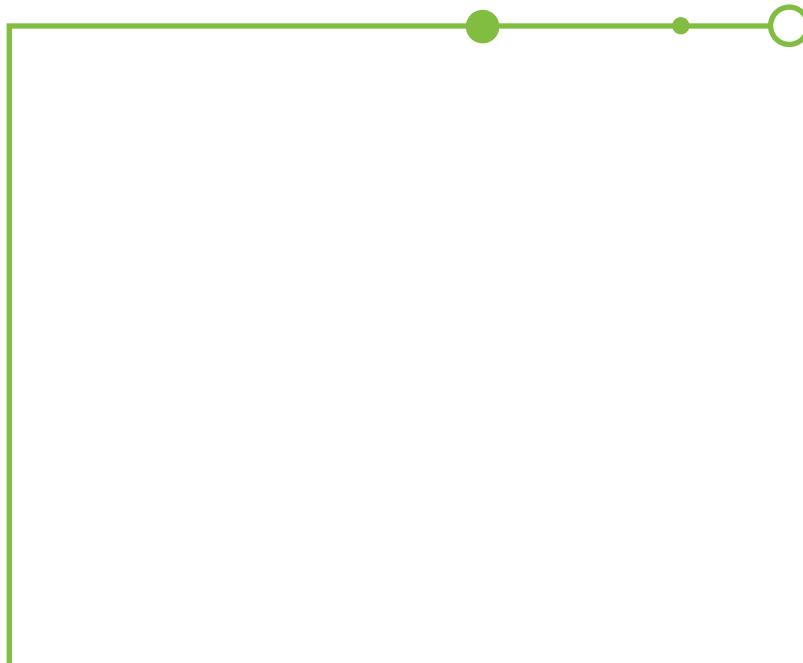
## **DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: YOUR THREE-STEP SOLUTION TO MAKING YOUR CLOUD DATA WAREHOUSE OR CLOUD DATA LAKE WORK** CHAPTER 8

### **3) What quantifiable business value is this data integration strategy supposed to bring my company?**

Though we provide many reasons why a data integration strategy is important to most businesses in this guide, it is essential to map out how an integration strategy will impact you specifically for a few reasons. First, it will help your technical teams get the funding that they need to bring your data together and to make it accessible for the rest of the business. Second, understanding the clear business value of an integration platform will help you prioritize what features should be most important to you when evaluating different integration methods and vendors.

### **4) Do you have the people resources in addition to the technical resources to effect the change you seek in your organization?**

Obviously, you will need technical resources to develop, maintain, and scale your integration initiatives. However, many integration strategies that enable a data-driven company go beyond a technical project. Instead, they require cultural and organizational change. If your integration project is meant to change the data landscape for your company, be sure to work with other stakeholders to envision what those changes mean to IT and the business units who are working with the data.



## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: YOUR THREE-STEP SOLUTION TO MAKING YOUR CLOUD DATA WAREHOUSE OR CLOUD DATA LAKE WORK **CHAPTER 8**

When considering how you're going to pull data into your systems for use, you need to also consider what data you're going to make available. Data curation may seem like an arcane topic, but with the risks of bad data getting greater than ever, and the need to comply with data protection regulations like GDPR and CCPA, it's becoming something IT professionals need to consider.

[IDC research has found](#) that today data professionals are spending 81% of their time searching, preparing, and protecting data with little time left to turn it into business outcomes. It has become crucial that organizations establish this same single source of access to their data to be in the winner's circle.

Although technology can help to fix the issue, enterprises need to set up a discipline to organize their data at scale. This discipline is called data governance. But traditional data governance must be re-invented with this data sprawl: according to [Gartner](#), "through 2022, only 20% of organizations investing in information will succeed in scaling governance for digital business." Given the sheer number of companies that are awash in data, that percentage is just too small.

Modern data governance is not only about minimizing data risks but also about maximizing data usage, which is why traditional authoritative data governance approaches are not enough. There is a need for a more agile, bottom-up approach. That strategy starts with the raw data, links it to its business context so that it becomes meaningful, takes control of its data quality and security, and fully organizes it for massive consumption.



*Through 2022,  
only 20% of organizations  
investing in information  
will succeed  
in scaling governance  
for digital business.*

## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: YOUR THREE-STEP SOLUTION TO MAKING YOUR CLOUD DATA WAREHOUSE OR CLOUD DATA LAKE WORK **CHAPTER 8**

Empowering this new discipline is the promise of data catalogs; leveraging modern technologies like smart semantics and machine learning to organize data at scale and makes data governance collaborative by engaging anyone for social curation. Comprehensive data integration solutions contain data cataloging capabilities. This is perfect for companies that modernize their data infrastructures with data lakes or cloud-based data warehouses, where thousands of raw data items can reside and can be accessed at scale. The catalog acts as the fish finder for that data lake, leveraging crawlers across different file systems (traditional, Hadoop, or cloud) and across typical file format. Then, data catalogs automatically extract metadata and profiling information, for referencing change management classification and accessibility.

Not only can a data catalog bring all of those metadata together in a single place, but it can also automatically draw the links between datasets and connect them to a business glossary. In a nutshell, this allows businesses to:

- **Automate the data inventory**
- **Leverage smart semantics for auto-profiling, relationships discovery and classification**
- **Document and drive usage now that the data has been enriched and becomes more meaningful**



# DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES: YOUR THREE-STEP SOLUTION TO MAKING YOUR CLOUD DATA WAREHOUSE OR CLOUD DATA LAKE WORK **CHAPTER 8**

## STEP 2: ENABLE SELF-SERVICE WITH A COLLABORATIVE DATA INTEGRATION SOLUTION.

It is widely proven and accepted that a data-driven organization optimizes its performance. For example, per the [McKinsey Datamatics Survey](#), data-driven companies have a 23x greater customer acquisition, 6x better customer retention, and 19x larger profits. But how does a company establish a culture of being data-driven?

The use of data is spreading rapidly at all levels of every organization and across all departments. As the volumes and varieties of data continue to rise, the use of data becomes more frequent. Subsequently, the demand for the availability of real-time data to make business decisions is high.

The aspiration created by the consumerization of data gives more power to business users. But concerns about security and compliance can sometimes lead IT to clamp down on access to data. Yet refusing to consider business users' demands for more agility, autonomy, and collaboration, can be dangerous as well. Whether they are experts in data or business users, if IT denies them access, business users will find a workaround. This practice is known as 'shadow IT' and it may jeopardize your company even further. The challenge is finding a way to make data available to employees in a safe, governed, secure way.

Because business demands are too numerous and varied to be satisfied by a single department, centralizing data management in the hands of IT lacks the ability to scale. Centralized IT will only result in the frustration of both IT side and business users. No one is better qualified than an accountant to clean up, enrich and reconcile billing and supplier portfolio data, or a marketing manager to do the same on leads generated by a given event. The competencies and the ability to make sense of data is – by definition - distributed across each line of business.

Today, it's possible for anybody to accomplish a task without knowing or caring about the complexities. For simple use cases, it's sufficient just to move the data. When use cases become more complicated and require governance, quality, or data transformation, the tools become more esoteric and capable. This is where IT can play a role – but even business users should be able to participate (e.g. data stewardship). Far from conflicting with each other, these varied audiences and tools are complimentary.



## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES:

### YOUR THREE-STEP SOLUTION TO MAKING YOUR CLOUD DATA WAREHOUSE OR CLOUD DATA LAKE WORK **CHAPTER 8**

For example, vendor billing files or enriched, new lead marketing files create value for the business and deserve to be shared and reused within the organization. Having an IT framework that allows for the collaboration of business and IT is important to have with the appropriate data management and security rules.

Because data preparations from individual employees will be even more useful if they can be shared and accessed by all the company's data sources and feed all its target applications, IT is justified in orchestrating these centralized repositories.

IT has a unique opportunity in front of it to keep (or regain) its role as a catalyst for change in the company. Take the initiative to set up and exploit collaborative enterprise repositories and rely on self-service data-processing solutions vs. those that are centrally managed.

The best self-service data platforms enable IT to deliver massive individual productivity gains for all employees, as well as collective insight through collaboration. They allow the integration of all data sources and targets. They ensure the consistency, compliance and security of data management.

Self-service data preparation tools represent an opportunity for all IT professionals to meet immediacy, autonomy and collaboration needs of employees, while maintaining data governance, and escaping the pitfalls of an overly prescriptive IT team.

*"In many organizations, the data governance team is seen as a roadblock to progress, but with the right mix of vision, people, process, policy, and technology, the data governance team can be transformed into a data enablement team."*

- Stewart Bond, director of Data Integration and Integrity Software research at IDC

## DEFINITIVE GUIDE TO CLOUD DATA WAREHOUSES AND CLOUD DATA LAKES:

YOUR THREE-STEP SOLUTION TO MAKING YOUR CLOUD DATA WAREHOUSE OR CLOUD DATA LAKE WORK

**CHAPTER 8**

### **STEP 3: EXPLORE PERVERSIVE DATA QUALITY IN YOUR CLOUD DATA INFRASTRUCTURE.**

There is a plethora of standalone data quality and data governance tools on the market. Register for any data tradeshow and you will discover plenty of data preparation and stewardship tools offering several benefits to fight bad data. Standalone tools can provide a quick fix but won't solve the problem in the long run. It's common to see specialized data quality tools requiring deep expertise for successful deployment. These tools are often complex and require in-depth training to be launched and used. Their user interface is not suitable for everyone so technical staff can manage them.

While these tools can be powerful, they won't serve either your short term or your long term data priorities. When the tools are too complex, you'll miss out on the speed benefits working in the cloud provides. On the other hand, you will find simple apps that anyone can use are often too siloed to be inserted into a comprehensive data quality process. You therefore need to have a comprehensive solution that shares, operates, and transfers data, actions, and models together.

Digital transformation depends on trusted data at speed. Data must be timely, because digital transformation is all about speed and accelerating time to market— whether that's providing real-time answers to business teams or delivering personalized customer experiences. But while speed is critical, it's not enough. For data to enable effective decision-making and deliver remarkable customer experiences, organizations need data they can trust.

The right approach to data quality and data governance is a pervasive one, where data quality issues are resolved upstream in the data flow.

Pervasiveness allows data quality tools to run anywhere and enables users to apply the same data quality sensors, controls, and metrics just in time and consistently across the data chain. And now, thanks to the cloud, data quality capabilities can be fully ubiquitous and access any data, anywhere in the data chain, and make it trustworthy. This applies to streaming data, real-time data or even data at rest; no matter if it is stored in an enterprise application, a cloud data warehouse, or in a data lake cluster. In addition, improvements in technology like pattern recognition, IA, and machine learning put data quality into anyone's hands. This encourages data literacy now that profiling morphs into data discovery, but also engages a wider audience to collaborate for better data.



# CHAPTER 9

## CASE STUDIES

---





**Uniper generates, trades, and markets energy on a large scale. With about 36 gigawatts of installed generation capacity, Uniper is among the largest global power generators in Europe. Uniper customers include large industrial customers and municipalities in Germany and neighboring countries.**

“We are in an increasingly complex world of ever-changing technologies and markets,” said René Greiner, Vice President for Data Integration, Uniper SE. “We produce energy. We buy and sell energy via marketplaces. How much coal and gas do we need to produce today and in the future? Is the market going to turn in a completely different direction? How shall we expand our market positions? How can we maximize our profit and loss?

“Before we embarked on our cloud journey, we didn’t have our data readily available to quickly make these decisions,” continues Greiner. “Once the idea of an organization-wide data strategy emerged, we decided to go with a public cloud solution for reasons of scalability and cost. We concluded that Talend would be the best software for such a cloud architecture,” said Greiner.

To create a new, cloud-based data infrastructure, dubbed the the Uniper Data Analytics Platform, Uniper selected Tableau along with Talend to integrate more than 120 internal and external data sources into a Snowflake central data lake in the Microsoft Azure Cloud. “Within 40 days,” said Greiner, “we reduced our integration costs by 80 percent.”

“Talend is essential to our cloud strategy because it can access any data from virtually any place and removes limits on data format and volume,” Greiner said. “Data governance, as provided by Talend, is essential to the success of the data lake. Talend Data Catalog provides those capabilities and helps us establish data lineage and the kind of security we need to comply with GDPR regulations.”

## INDUSTRY

- Utilities

## INFORMATION

- HQ: Germany
- 10,000+ employees

## USE CASE

- Operational efficiency

## CHALLENGE

- Providing self-service data and analytics in real time

## TECH STACK

- Microsoft Azure
- Snowflake
- Talend Cloud
- Talend Data Catalog

The results of the Uniper Data Analytics Platform have been impressive. To make informed decisions, Uniper relies on marketing analytics based on real-time information. Now that the relevant information is aggregated in the data lake, market analysis teams can access data faster and provide answers faster to questions they get every day from traders. Questions that previously required months of research can now be answered right away, or in just a few days. Speed is important in answering questions because the earlier trading teams can react, the earlier they can take a position and that can make a difference of millions of euros.

Talend has been a game changer by supplying data ten times faster and ten times cheaper. Greiner explained, “We have a mountain of data, and we want people to be able to retrieve and use it themselves, which Talend makes possible. Self-service gives us an advantage in speed-to-market.”

*“Talend is essential to our cloud strategy because it can access any data from virtually any place and removes limits on data format and volume.”*

- Rene Grenier,  
Vice President for Data Integration,  
Uniper SE



**Anheuser-Busch InBev SA/NV (AB InBev) is a Belgian publicly traded transnational beverage and brewing company, with a heritage that dates back more than 600 years, spanning continents and generations. It is considered one of the largest fast-moving consumer goods (FMCG) companies in the world.**

When companies grow via external acquisitions, integrating the systems and data from acquired companies is always a challenge. For AB InBev, that challenge included a hybrid environment with both on-premises and cloud systems such as Salesforce, 15 SAP instances, 27 ERP systems, 23 ETL tools, and a host of brewers operating as independent entities with their own internal systems. “As you can imagine, that made it extremely difficult to obtain a single, unified view of our business, and there was no single source of truth,” says Harinder Singh, Global Director of Data Strategy & Solution Architecture at AB InBev. “Also, we’re operating on six continents and we needed to become GDPR-compliant, and that required global visibility into all our data assets.”

Singh says he knew AB InBev had to take a different approach. “We recognized that we needed one central repository for our data assets,” he says. “Our internal customers—like data scientists, operations teams and business teams—were struggling to pull together data from over 100 source systems, analyze it, and make timely decisions on product development, supply chains, marketing campaigns and more.” We knew we wanted to embark on a cloud journey, and Talend was built in that world, enabling cloud and on-premises systems to talk to each other in a secure manner,” says Singh.

#### INDUSTRY

- Food and Beverage

#### INFORMATION

- HQ: Belgium
- 10,001+ employees

#### USE CASE

- Customer experience
- Sales and Marketing effectiveness
- Supply chain optimization

#### CHALLENGE

- Integrating systems and data from acquired companies

#### TECH STACK

- Microsoft Azure
- Hortonworks for Hadoop
- Talend Cloud
- Talend Data Preparation



In the architecture AB InBev has built, Talend extracts data from a range of sources—real-time and batch, cloud and on-premises,” ERP systems, data from IoT devices—and stores it in a landing zone, which is part of a data lake, or data hub, that resides in the cloud on Microsoft Azure. That data is then processed and archived before going into a golden layer, from where it’s consumed by AB InBev internal users.

“Among the biggest benefits of the new IT architecture are simplification of the infrastructure, and reusability of processes to rapidly extract and provide access to data,” says Singh. “Because we have reusable code, what used to take us six months now takes us six weeks. That translates into faster decisions and reduced time to market for decisions, campaigns, products and more.” Singh cites cost savings as another major benefit. Now, instead of paying for and managing 23 different ETL tools, we’re moving towards managing only one by standardizing on Talend.”

We’re building a company to last, brewing beer and building brands that will continue to bring people together for the next hundred years and beyond,” says Singh. “Data is one of the most important factors in making that happen.”

*“Among the biggest benefits of the new IT architecture are simplification of the infrastructure, and reusability of processes to rapidly extract and provide access to data. What used to take us six months now takes us six weeks.”*

*- Harinder Singh,  
Global Director of Data Strategy &  
Solution Architecture at AB InBev*





**AstraZeneca plc is a global, science-led biopharmaceutical company headquartered in Cambridge, United Kingdom. It is the world's seventh-largest pharmaceutical company and has operations in over 100 countries.**

AstraZeneca had data dispersed throughout the organization in a wide range of sources and repositories. Having to draw data from CRM, HR, Finance systems and several different versions of SAP ERP systems slowed down vital reporting and analysis projects. Says Simon Bradford, Senior Data & Analytics Engineer at AstraZeneca: "We knew we needed to put in place an architecture that could help with a mass consolidation and bring data together in a single source of the truth." In addition to causing inconsistencies in reporting, Bradford says silos of information prevented the company and his division, the Science and Enabling Unit, from finding insights hiding in unconnected data sources.

We wanted to consolidate everything and get a single set of global metrics so we could monitor activity across divisions and markets and do comparisons that were not previously possible." AstraZeneca resolved to build a data lake on AWS to hold the data from its wide range of source systems. To capture that data, they selected Talend. McPhee explains: "Talend is responsible for lifting, shifting, transforming and delivering our data into the cloud, extracting from multiple sources and then pushing that data into Amazon S3.

#### INDUSTRY

- Biopharmaceuticals

#### INFORMATION

- HQ: UK
- 10,001+ employees

#### USE CASE

- Operational efficiencies:  
Monitoring activities across Sciences and Enabling unit

#### CHALLENGE

- Twice the value for half the cost

#### TECH STACK

- Amazon Redshift
- Amazon S3
- Amazon Aurora
- Amazon Elastic Beanstalk
- Talend Cloud

AstraZeneca has deployed Talend as part of the orchestration layer in its architecture. In addition to extracting data from CRM, ERP, finance, document management, HR and other systems to load into the data lake, Talend serves to facilitate point-to-point connections, such as those between an Amazon Redshift analytic database and an SQL database in order to add data to what AstraZeneca calls a ‘conformed layer’.*“The data lake enables us to pull large volumes of valuable data from disparate systems and make our data discoverable across divisions,” says Bradford.*

The data lake aligns with AstraZeneca’s business priorities: there is a goldmine of data available to help the Sciences and Enabling unit to manage itself more efficiently, with a new level of visibility. *“We started using the data lake for certain projects and, each week another manager will come in with a new set of business questions,” says Bradford. “We’ve really only scratched the surface of what’s possible.”*



*“We started using the data lake for certain projects and, each week another manager will come in with a new set of business questions. We’ve really only scratched the surface of what’s possible.”*

- Simon Bradford, Senior Data & Analytics Engineer  
at AstraZeneca



# CHAPTER 10

## INTEGRATION SOLUTION CHECKLIST





## YOUR INTEGRATION SOLUTION CHECKLIST

Have you accomplished the tasks you need to succeed?

- Define the purpose of your cloud data warehouse and data lake.
- Define what data you need to achieve your goals.
- Define whether transformation needs to take place or you only need simple data processing tasks.
- Select a cloud-based integration solution to complement your cloud strategy.
- Ensure your integration solution is hybrid-capable, in order to future-proof your investments by moving data from on-premises to cloud, or even between cloud.
- Ensure your integration solution is flexible to accommodate new technologies, systems, and applications.
- Ensure your integration solution is scalable enough to grow with you.  
Ideally the integration solution you buy today is the last one you ever buy.

# CHAPTER 11

## WHAT'S THE TAKEAWAY?

---





### WHAT'S THE TAKEAWAY?

Moving your data to a cloud data lake or cloud data warehouse could have profound and lasting business benefits. But it's important to make sure that you make the most out of your cloud investment. Make sure that ingesting the data into your cloud data warehouse or cloud data lake and ensuring its quality is very easy, and make sure that your entire business has access to the trusted data you've collected.

Contact Talend today to learn more about our cloud data warehouse and cloud data lake solutions.



#### Contact us

<https://www.talend.com/contact/>



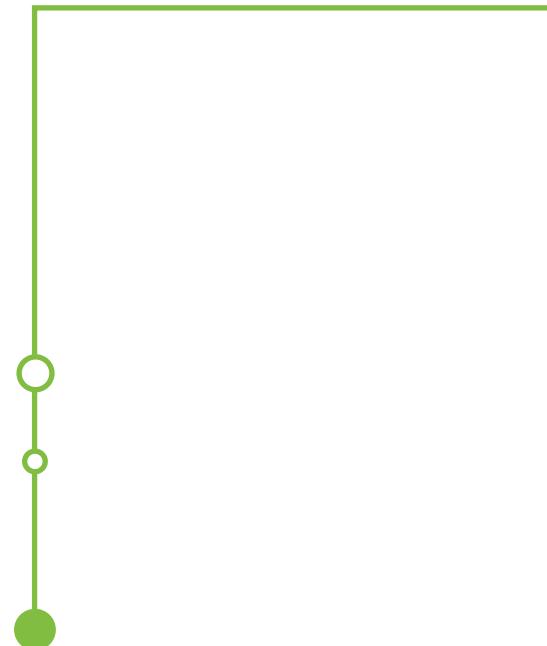
#### Customer Support

[Visit the Talend Community](#)



#### Learn More

[www.talend.com](http://www.talend.com)



Talend (Nasdaq: TLND), a leader in cloud integration solutions, liberates data from legacy infrastructure and puts more of the right data to work for your business, faster. Talend Cloud delivers a single platform for data integration across public, private, and hybrid cloud, as well as on-premises environments, and enables greater collaboration between IT and business teams. Combined with an open, native, and extensible architecture for rapidly embracing market innovations, Talend allows you to cost-effectively meet the demands of ever-increasing data volumes, users, and use cases.

Over 1,500 global enterprise customers have chosen Talend to put their data to work including GE, HP Inc., and Domino's. Talend has been recognized as a leader in its field by leading analyst firms and industry publications including Forbes, InfoWorld, and SD Times. For more information, please visit [www.talend.com](http://www.talend.com)