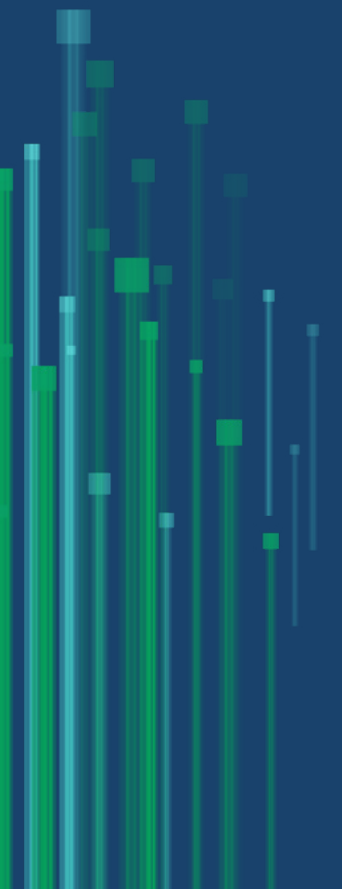
Qlik

Definitive Guide to Data Quality

Content



Introduction

An astonishing amount of data is being generated at an estimated pace of **2.5 quintillion bytes of data per day** — and this amount is expected to increase exponentially in the coming years. In fact, a staggering **90% of the world's data** has been generated in just the last two years, highlighting the rapid pace at which data is being created.

For the enterprise, all of this data creates an extraordinary opportunity. Instead of making business decisions by gut feel or instinct, they can be based on proven fact, observed and tested against what customers are thinking and doing. Being able to collect data, integrate it, interrogate it, and extract insights out of it has become a significant competitive differentiator in today's business environment. You have only to look at the success of Amazon, Netflix, or Google to realize the extraordinary power of data.

But the insights that a business can extract out of data are only as good as the data itself. Poor data quality makes it impossible to generate trustworthy insights and consequently leads to poor decision-making.

Organizations need to make sure that high-quality data is available to everyone who needs it, and they shouldn't have to rely exclusively on a small IT team or a couple of "rock star" data personnel to make that happen. Beyond IT, everyone from data scientists to application integrators to business analysts should be able to participate in maintaining trusted data so they can extract valuable insights out of it.

In this Definitive Guide, we explore the factors that contribute to creating quality data and offer guidance on how enterprises can ensure that all their data meets high standards. We also provide insights on making quality data available to everyone who needs it in a secure and governed manner. By prioritizing the quality of their data, organizations can unlock its full potential.

90% of the world's data has been generated in just the last two years, highlighting the rapid pace at which data is being created.

CHAPTER 1

What is data quality and why does it matter?

If data fuels your business strategy, poor-quality data could kill it. The results can be disastrous and cost you millions.

Time and time again, we've seen organizations large and small fail because they put their trust in bad data. In 2020, a study published in the Journal of the American Medical Association (JAMA) found that poor data quality in electronic health record (EHR) systems was a major issue in the fight against the COVID-19 pandemic.

The study found that many EHR systems were not designed to capture and track important data related to COVID-19, such as test results, symptoms, and patient outcomes. As a result, healthcare providers were often unable to access the information they needed to effectively diagnose and treat patients — leading to delays in care and potentially worse health outcomes.

Furthermore, the study found that poor data quality also hindered efforts to track the spread of the virus and to develop effective public health interventions. Inaccurate and incomplete data made it difficult to accurately identify outbreaks and hotspots, and to measure the effectiveness of different interventions.

This example illustrates how poor data quality can have serious consequences for public health and safety, as well as for the effectiveness of healthcare systems. By prioritizing data quality and investing in improvement initiatives, healthcare providers and public health organizations can ensure that their data is accurate, consistent, and reliable, ultimately leading to better health outcomes and a more effective response to public health crises.

Of course, this example is just one of many ways that relying on bad data can harm an organization — and anyone who relies on that organization. It leads to wasted resources, missed opportunities, and far too much time spent fixing data — time that could be better spent on other areas of the business. And all of this translates into increased costs. In fact according to Gartner, poor data quality costs organizations $12.8 million dollars per year on average. With the exponential overall growth of data, the cost of poor data quality will also grow exponentially if not addressed quickly.

That's why it's crucial to spot and fix that data in your organization.

How to spot bad data

Bad data can come from every area of your organization, from sales to engineering. But there is a common framework to assess data quality. The five most critical dimensions are:

1. **COMPLETENESS**
   Is the data sufficiently complete for its intended use?

2. **ACCURACY**
   Is the data correct, reliable, and/or certified by some governance body? Data provenance and lineage — where data originates and how it has been used — may also fall in this dimension, as certain sources are deemed more accurate or trustworthy than others.

3. **TIMELINESS**
   Is this the most recent data? Is it recent enough to be relevant for its intended use?

4. **CONSISTENCY**
   Does the data maintain a consistent format throughout the dataset? Does it stay the same between updates and versions? Is it sufficiently consistent with the other datasets to allow joins or enrichments?

5. **ACCESSIBILITY**
   Is the data easily retrievable by the people who need it?

Data observability: the key to continuous data quality

Since data can be spread across and moved to different locations within an organization, most organizations can't see the state of their data until something goes wrong. As a result, low-quality data can hinder critical decision-making, and the "garbage in, garbage out" dynamic that typically builds up as data flows through the system can increase data handling costs.

Data observability is the practice of being able to continuously evaluate your data and provide insights into how it is evolving. This means more than just data dashboards — it is the what, where, why, and how behind your data systems. It requires a holistic view of many different aspects, including lineage, monitoring, and notifications, in addition to quality. This gives you a 360-degree view of your entire data lifecycle, from data ingestion and transformation to data access and disposal. An effective data observability solution can provide insights into how the different dimensions of data quality evolve over time, which can be useful for assessing the effectiveness of data quality remediations. Ultimately, this enables consumers to trust the data they receive and to use it confidently, while data stewards can identify bad data and remediate it quickly.

Five myths about data quality

Results from the sixth annual Gartner Chief Data Officer (CDO) survey show that data quality initiatives are the top objective for data and analytics leaders. But the truth is that little has been done to solve the issue. Data quality has always been perceived by organizations as difficult to achieve. In the past, the general opinion was that the process for achieving better data quality was too lengthy and complicated.

Let's take a closer look at a few common data quality misconceptions.

MYTH 1

"Data quality is just for traditional data warehouses."

Today, there are more data sources than ever, and data quality tools are evolving. They are now expanding to handle any dataset — whatever its type, its format, and its source. It can be on-premises data or cloud data, data coming from traditional systems, and data coming from IoT devices. Faced with data complexity and growing data volumes, modern data quality solutions can increase efficiency and reduce risks by fixing bad data at multiple points along the data journey, rather than only improving data stored in a traditional data warehouse. These data quality solutions use machine learning (ML) and natural language processing (NLP) capabilities to ease your work and separate the wheat from the chaff. The earlier you can implement these solutions to fix your data, the better. Solving data quality downstream at the edge 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.[1]

MYTH 2

"Once you solve your data quality, you're done."

Just as data does not come all at once to a company, improving data quality is not a onetime operation. For example, data quality issues can arise from changes in business processes, changes in data sources, or changes in regulatory requirements. Data quality must be 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. This is where data observability capabilities can provide immense value, by allowing you to track how data changes over time and implement controls where necessary to improve its quality.

[1] The 1-10-100 rule is a quality management concept developed by G. Labovitz and Y.S. Chang that is used to quantify the hidden costs of poor quality. Labovitz, G., Chang, Y.S., and Rosansky, V., 1992. Making Quality Work: A Leadership Guide for the Results-Driven Manager. John Wiley & Sons, Hoboken, NJ.

MYTH 3

“Data quality is IT’s responsibility.”

Gone is the time when maintaining trustworthy data was simply an IT function. Data should be the whole company’s priority as well as a shared responsibility — from business users who enter data into systems to managers who oversee data operations to data stewards who are responsible for data governance. No central organization, whether it’s IT, compliance, or the office of the CDO, can magically cleanse and qualify all organizational data. It’s better to delegate some data quality operations, and roles can evolve over time. For example, business users can become data stewards and play an active role in the data management process. It’s only by moving from an authoritative model to a more collaborative approach that you will succeed in your modern data strategy.

MYTH 4

“Data quality software is complicated.”

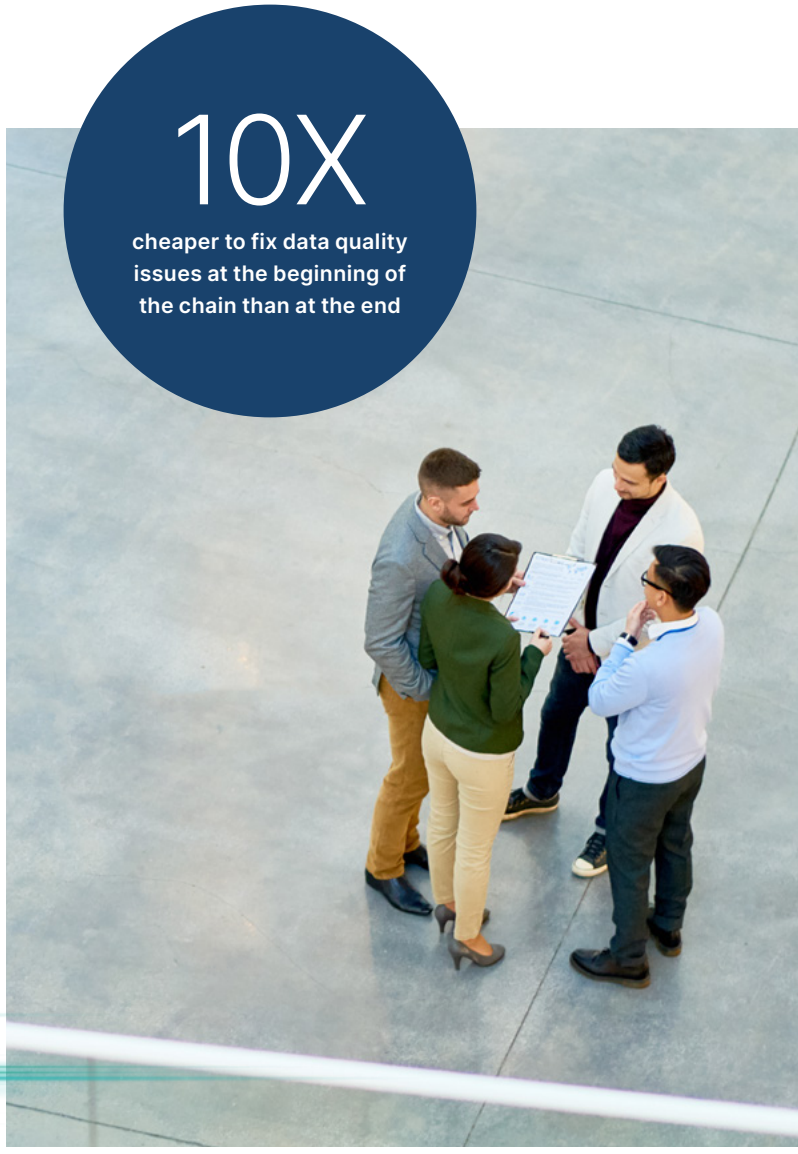
As companies start to rely on data citizens and data has become a shared responsibility, data quality tools have also evolved. Many data quality solutions are now designed as self-service applications so that anyone in an organization can combat bad data. With an interface that is familiar to users who spend their time using popular data programs like Excel, a non-technical user can easily manipulate big datasets while keeping the company’s raw data intact. Line of business users can enrich and cleanse data without requiring any help from IT. Connected with your apps like Marketo and Salesforce, these solutions will dramatically improve your daily productivity and your data flows.

MYTH 5

“Investing in data quality improvements is too expensive and time-consuming.”

In reality, the costs of poor data quality can far outweigh the costs of investing in data quality improvement. For example, inaccurate data can lead to missed business opportunities, increased risk, and legal and regulatory compliance issues. Investing in data quality improvement can ultimately lead to improved business outcomes and a stronger competitive advantage.

It’s time for you to take care of your organization’s data quality.



CHAPTER 3

Fixing bad data with proactive data quality

It's 10x more expensive to fix bad data at the end of the chain than it is to cleanse it when it enters your system. But the costs don't stop there. If that data is acted upon to make decisions, or sent out to your customers, or otherwise damages your company or its image, you could be looking at a cost of $100 or more compared to the $1 it would've cost to deal with that data at the point of entry. The cost gets greater the longer bad data sits in the system.[2]

Pervasive data quality can ensure, analyze, and monitor data quality from end to end. This proactive approach allows you to check and measure data quality before the data gets into your systems. Accessing and monitoring data across internal, cloud, web, and mobile applications is a huge undertaking. The only way to scale that kind of monitoring across those types of systems is by embedding data quality processes and controls throughout the entire data journey.

With the right tools, you can create whistleblowers that detect and surface some of the root causes of poor data quality. Once a problem has been flagged, you need to be able to track the data involved across your landscape of applications and systems and parse, standardize, and match the data in real time.

This is where data stewardship comes in. Many modern solutions feature point-and-click, Excel-like tools so business users can easily curate their data. These tools allow users to define common data models, semantics, and rules needed to cleanse and validate data and then define user roles, workflows, and priorities. Tasks can then be delegated to the people who know the data best. Those users can curate the data by matching and merging it, resolving data errors, and certifying or arbitrating on content.

Holistic platforms like Qlik Talend can simplify these processes even further because data integration, preparation, stewardship, and governance capabilities are all part of the same unified platform. They can be easily embedded into data integration flows, MDM initiatives, and matching processes to manage and quickly resolve any data integrity issues.

$100+

in potential costs once bad data is acted upon compared to $1 to fix it at the point of entry

[2] The 1-10-100 rule is a quality management concept developed by G. Labovitz and Y.S. Chang that is used to quantify the hidden costs of poor quality. Labovitz, G., Chang, Y.S., and Rosansky, V., 1992. Making Quality Work: A Leadership Guide for the Results-Driven Manager. John Wiley & Sons, Hoboken, NJ.

CHAPTER 4

Six steps for better data quality

With pervasive data quality embedded at every step of the data journey, organizations can close the gap on ensuring that trusted data is available everywhere in the enterprise. But what does the data quality process actually look like?

There are six key steps for delivering quality data. While the specifics may vary when looking at different data sources and formats, the overall process remains remarkably consistent. In fact, that highlights another benefit of using a single, unified platform across your entire data infrastructure — you don't have to build everything from the ground up every time you add a source or target.

In this case, when quality rules are created, they can be reused across both on-premises and cloud implementations, with batch and real-time processing, and in the form of data services that can automate data quality processes.

The six steps for better data quality are:

STEP 1 **OBSERVE**
Continuously track data sources and evaluate their quality and integrity.

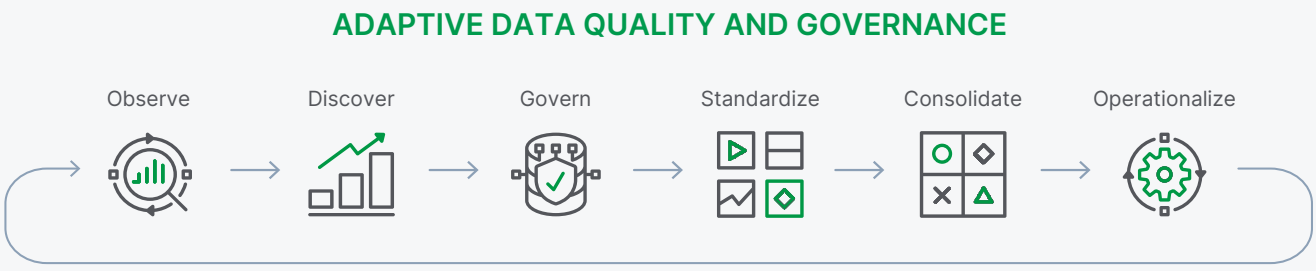
STEP 2 **DISCOVER**
Search, find, and profile data to understand structure and content.

STEP 3 **GOVERN**
Define data owners, roles, and data-related policies.

STEP 4 **STANDARDIZE**
Establish application of consistent formats, definitions, and structures.

STEP 5 **CONSOLIDATE**
Eliminate data redundancies, inconsistencies, and create a source of truth.

STEP 6 **OPERATIONALIZE**
Integrate data quality and governance practices into day-to-day workflows.



Now that you know what goes into improving data quality, let's take a look at how to choose the best solution for your organization.

CHAPTER 5

Choosing the right data quality solution

Run any quick search and you'll discover plenty of data preparation and stewardship tools designed to fight bad data — but only a few of them cover data quality for all.

Many specialized data quality tools require deep expertise for successful deployment and require in-depth training for users. Their sometimes-confusing user interfaces may not be suitable for business users, so typically only IT uses them.

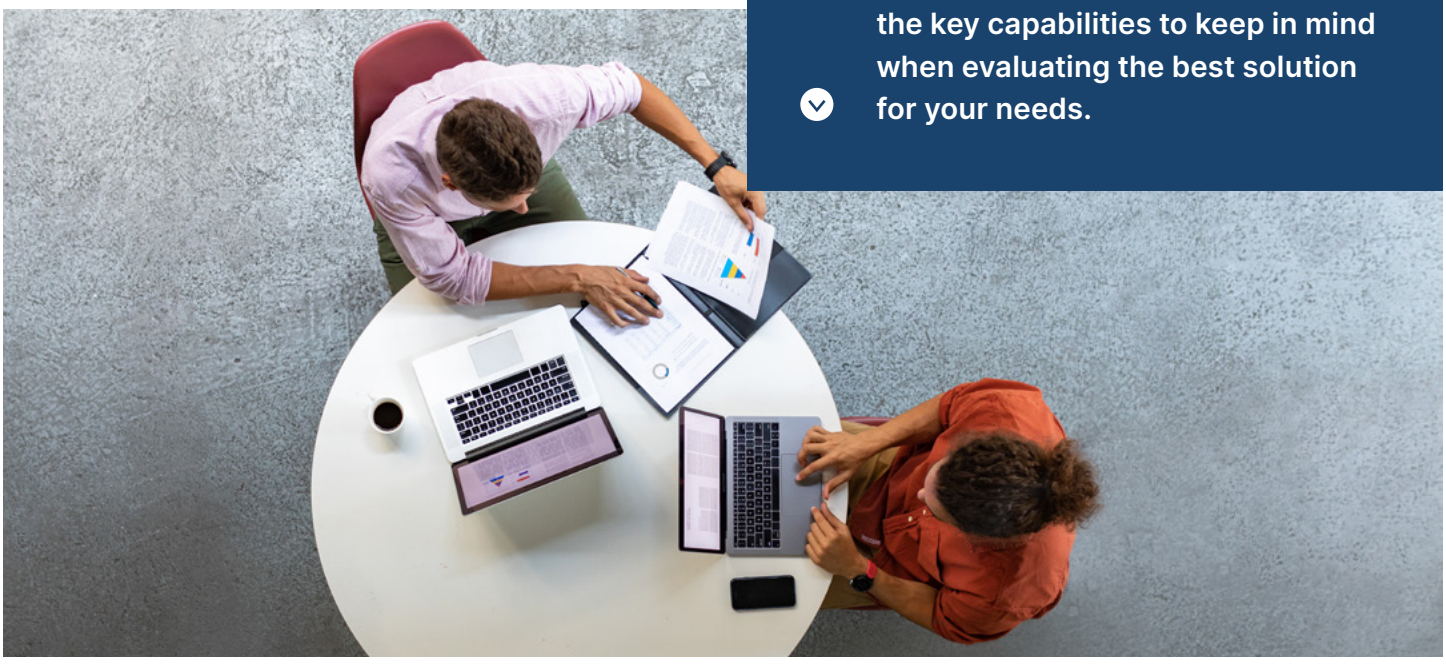
While these data quality tools can be powerful, their complexity is also their downfall. Deploying them in a collaborative environment is like asking a casual runner to participate in a marathon. The runner will not have the knowledge or experience to compete effectively, and things won't go well.

On the other hand, more basic programs may be too limited to be used in a comprehensive data quality process. Even if they successfully cater to business users with a simple UI, they may miss the important part — collaborative data management, which applies to both users and the technologies they deploy. When it comes to data quality, success relies not only on the tools and their capabilities, but also on their ability to talk to each other.

It's impossible for a single person or team to manage an entire organization's data successfully.

Instead, a solution that enables IT and business users to work together throughout the data lifecycle can help build a collaborative culture where high-quality data can thrive. The best way to do this is through a secure, unified, cloud-based platform that provides better accessibility, scalability, and reliability. This allows users to share, operate, and transfer data, actions, and models together.

To meet these objectives, here are the key capabilities to keep in mind when evaluating the best solution for your needs.



Key capabilities for data quality solutions

OBSERVE

Quality monitoring so you can identify and address data quality issues that affect the validity of the data analysis

Quality evolution so you can monitor data quality changes for validity and relevance

Quality alerts so you can notify stakeholders to promptly address data quality concerns

Pipeline monitoring so you can track data flows and identify failures or delays in the pipeline

Pipeline troubleshooting so you can identify the root cause of data loss, inconsistencies, and pipeline bottlenecks

DISCOVER

Dataset searching so you can quickly locate the datasets needed for analysis and model building

Dataset crawling so you can fetch and integrate newer datasets to discover fresh insights and run exploratory or comprehensive analysis

Dataset tagging and labeling so you can enable efficient organization and retrieval of datasets as well as promote stakeholder collaboration and insight sharing

GOVERN

Data modeling so you can create logical and physical data models to represent the organization's assets

Data lineage so you can visualize and trace data's journey from its source to various transformations and destinations

Data impact analysis so you can modify analysis processes and communicate changes to stakeholders by identifying potential impacts on existing reports, dashboards, and analytical models

CONSOLIDATE

Data matching so you can create a comprehensive dataset by finding data records representing the same entity

Data merging so you can create a unified view of data for everyone

Data survivorship so you can create a golden record

OPERATIONALIZE

Operationalize data preparations so you can build and share preparations with other users

Operationalize data campaigns so you can create data campaigns and share them with different governance stakeholders

STANDARDIZE

Data formatting so you can accurately compare and combine diverse datasets

Data validating so you can catch discrepancies, errors, and inconsistencies before data is used for analytics and model building

Data certifying so you can give users faith in the accuracy and consistency of data for analysis, modeling, and insights

Data enriching so you can leverage reference data or additional third-party data to expand the scope and coverage of a dataset

Data blending so you can look up and bring data from another dataset into an existing dataset for deeper, more complex analysis

Data cleansing so you can improve overall data quality and eliminate "noise" by fixing incorrect, duplicate, or otherwise erroneous data in a dataset

CHAPTER 6

Four examples of data quality in the real world

INSURANCE

Maximizing business efficiency at scale

With over three million customers and C$13 billion in assets under management, Beneva is one of Canada's largest financial institutions, offering auto, home, life, travel, group, health, and credit insurance, along with investment products. After a merger in 2020, it became the largest mutual insurance company in Canada. The company needed to create a unified view of its customers and personalize its customer relationships — but after 75 years of operation, its data systems had become complex, siloed, and unable to be used effectively.

"It was difficult to make relevant offers to customers without having a complete picture of their insurance coverage," says Annie Pelletier, Marketing and E-Business Director. To break down these data silos, the company quickly opted for Qlik Talend. "Qlik Talend offers a complete solution, from data integration to data enhancement to API-based applications," explains Robert Beauregard, BI Architect at Beneva.

The company knew that to truly understand its customers, it would need to put high-quality data at the center of its business. "We weren't prepared to make any compromises as far as data quality is concerned," says Simon Latouche, Director of Data Engineering. "By using Qlik Talend Data Quality, we were able to standardize and clean our data. Qlik Talend's data matching has enabled us to establish links between people's names that were similar but not identical, based on their phonetics, using an algorithm developed by Stanford University."

Qlik Talend has also helped resolve one of the most difficult issues in such a project: data stewardship, where a human has to take back control from a machine. "Without Qlik Talend, we would not have been able to establish a complete process with data stewardship as a bridge to the algorithm that was developed," notes Latouche. "Historically, our data projects used to take between nine and 12 months. Now, with Qlik Talend, combined with the Data Vault 2.0 methodology, we enter production in agile mode every three weeks."

ENERGY

Developing Customer 360 on a global scale

A global leader in low-carbon energy and services has made it its mission to accelerate the transition to a carbon-neutral world through more energy-efficient and environmentally friendly solutions, including advanced technologies like offshore wind, green gas, and geothermal energy. But the company is large and complex, with 24 geographical divisions and 70 country entities, each with multiple lines of business. This made it impossible for the company to develop a single, unified view of its customers on a global scale. Yet according to the company's Director of Business Acceleration, it was necessary for the company to "understand the consumption habits of [its] customers around the world so [it] could better assess how to transition them to zero carbon."

The organization selected Qlik Talend to help it standardize data from 70 entities in multiple languages and improve its overall quality. According to the company, "Qlik Talend is second to none for trusted data. Prior to its implementation, more than 7% of our data was being rejected. With Qlik Talend, this percentage has fallen considerably, which is significant given that more than 90,000 opportunities are added to the system each week. Each entity is now responsible for quality."

With its high-quality data, the company can now accurately measure business efficiency through multiple key indicators, including sales, number of opportunities won and lost, and sales volume. "This shared, comprehensive vision of our customers helps us meet our ambitious goals for a carbon-neutral world."

"Qlik Talend is second to none for trusted data. Prior to its implementation, more than 7% of our data was being rejected. With Qlik Talend, this percentage has fallen considerably, which is significant given that more than 90,000 opportunities are added to the system each week."

— Director of Business Acceleration

PHARMACEUTICALS

Developing more vaccines, faster, with high-quality data

One of the world's largest vaccine companies strives to protect patients by discovering and developing better vaccines faster, but until recently, complex data systems and too many data silos severely limited the organization's efficiency and collaboration. Without better, faster access to high-quality data, the company would not be able to innovate rapidly, threatening its ability to help patients, remain competitive as a company, and make decisions that drive business growth.

The company knew it needed to turn its data into a shared and trusted asset, so it partnered with Qlik Talend. In 2019, it started an ambitious project to migrate its systems to the cloud, improve its data quality, and make that data more easily accessible to everyone in the organization. The company took a "start small, fail fast, and win big" approach, quickly expanding its new cloud architectural backbone across multiple business units, including R&D, manufacturing, and commercial.

With nearly 100,000 employees across more than 90 countries, making trusted data accessible to everyone who needs it was no easy feat. But thanks to its new cloud approach to data, the company has improved production efficiency, production quality, and compliance in manufacturing — and boosted commercial opportunities to drive faster growth. "For delivering value to our business and delivering medicines to our patients, Qlik Talend is a key enabler for us," says the company's Chief Data & Analytics Officer. Now the company can more quickly and efficiently develop and distribute lifesaving medicines and vaccines to the people who desperately need them.

"For delivering value to our business and delivering medicines to our patients, Qlik Talend is a key enabler for us."

— Chief Data & Analytics Officer

HOSPITALITY

Optimizing hotel revenue through better customer understanding

Travelodge is the UK's largest independent, low-cost hotel brand, with more than 560 hotels and 40,000 guest bedrooms across the UK, Ireland, and Spain. The budget hotel industry is highly competitive, and new services like Airbnb have only increased the pressure on a company like Travelodge, which aims to be the favorite hotel for value seekers. To stay ahead of the competition, retain current customers, and attract new ones, the company needs to provide outstanding customer experiences. "Our goal was to provide personalized offers to our customers in order to maximize occupancy at our hotels," says Niall Hammond, Data Architect for Travelodge. "To do this, we needed to know and understand our customers better through data and combine this understanding with occupancy forecasting data."

But Travelodge faced significant challenges. As a small-to medium-sized organization, it had a small data architecture team and disparate data. And although the company had already migrated to the cloud to gain resiliency, scalability, security, and cost-effectiveness, it still suffered from core data integration structural problems.

To solve its data challenges, the company selected Qlik Talend. "We chose Qlik Talend for its data integration, data management, and data quality capabilities, and for its speed, flexibility, comprehensive features, mature platform, and need for little overhead," explains Hammond.

Travelodge uses Qlik Talend to manage its 180-gigabyte operational data store, which houses its customer information, and to process 2,000 executions daily. The customer data is also provided to a third-party marketing partner, which uses it to create personalized offers that have produced millions of pounds in additional revenue per year. Qlik Talend's data quality capabilities were used to surface quality issues within Travelodge's internal property data and remediate them, ensuring that the business is only run on trusted data.

"With Qlik Talend, we can put ourselves in our customers' shoes, understand what's most important to them, in a location where they need to be, at a price they are prepared to pay," says Hammond. "We can also understand how busy a hotel will be so rooms can be marketed at an appropriate price and ensure we have appropriate staffing to provide a good customer experience. This way we are operating in a cost-effective manner."

"With Qlik Talend, we can put ourselves in our customers' shoes, understand what's most important to them, in a location where they need to be, at a price they are prepared to pay."

— Chief Data & Analytics Officer

Conclusion

When data fuels your business strategy, poor-quality data can kill it. Relying on bad data can be disastrous and cost you millions. However, many successful organizations mitigate those risks by making data quality a strategic priority. When supported by the right technology, these efforts allow professionals from every corner of the business to collaborate on, improve, and maintain accurate data.

Qlik Talend provides an end-to-end, modern data management solution that enables teams to discover, transform, govern, and share data across their organization. Qlik Talend combines data integration, quality, governance, and API and application automation in an industry-leading platform that handles virtually any data source, destination, or architecture. Capabilities like the Talend Trust Score® make it easy to assess data reliability and identify improvements instantly.

Learn how Qlik Talend's solutions can help you meet the data quality goals at your organization.

About Qlik®

Qlik transforms complex data landscapes into actionable insights, driving strategic business outcomes. Serving over 40,000 global customers, our portfolio leverages advanced, enterprise-grade AI/ML and pervasive data quality. We excel in data integration and governance, offering comprehensive solutions that work with diverse data sources. Intuitive and real-time analytics from Qlik uncover hidden patterns, empowering teams to address complex challenges and seize new opportunities. Our AI/ML tools, both practical and scalable, lead to better decisions, faster. As strategic partners, our platform-agnostic technology and expertise make our customers more competitive.

qlik.com

2-WEB-61
24-DI-0001-1-WHM