White paper

talend

Healthy data,
healthy business.

Definitive Guide to Data Quality

Table of contents

Introduction

Humanity is living in the Data Age. Every day, 2.5 quintillion bytes of data are created, and that amount will only continue to grow. In fact, 90 percent of the world's data was generated in the last two years alone. Every time someone emails your business, downloads an app, sends a text message, checks the weather, or does a thousand other everyday things, data is created, and those thousands of interactions by millions of people create an explosion of information. On top of that, Internet of Things (IoT) devices are an increasingly large contributor to data volume, responsible for 13.6 zettabytes of data in 2019. This number is forecast to grow to 79.4 zettabytes by 2025.

For the enterprise, all of this data creates an extraordinary opportunity; instead of making business decisions by gut feel or by instinct, they can be based on proven fact, observed and tested with what customers are thinking or 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 health makes it impossible to generate trustworthy insights and consequently leads to poor decisions. This is a reality that many executives are facing today. According to our recent survey, 78% of executives face challenges using data effectively at their company, and less than half rate their ability to deliver on data accuracy, consistency, accessibility, or completeness as "very good."

Organizations need to make sure that healthy 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 healthy data, so they can extract valuable insights out of it.

There is a way out of this conundrum.

As it turns out, data health and personal health have a lot in common. To stay healthy, we must develop good habits like exercising regularly, eating a balanced diet, and scheduling regular check-ups to ensure our bodies are working the way we expect them to work.

Similarly, organizations must have consistent and routine practices that reduce risk and exposure while ensuring business data is always available, usable, secure, and trustworthy. To build a system that produces healthy data, organizations need a combination of preventive measures, effective treatments, and a supportive culture.

This Definitive Guide focuses on one of the key components for healthy data: data quality. This guide will cover the ingredients that go into creating quality data, explore how enterprises can ensure that all of their data is of good quality, and show you how to make that quality data available to anyone who needs it in a secure and governed fashion.

talend

Chapter 1

What is data quality and why does it matter?

If data fuels your business strategy, poor quality data could kill it

Just like professional athletes don't fuel their body with junk food, your business cannot function properly on a diet of unhealthy data. With poor data, results can be disastrous and cost you millions.

Time and time again, we've seen companies large and small fail because they put their trust in bad data. For example, in 2013, Target began its expansion to Canada with more than favorable market conditions. But the company's inventory data was incomplete and inaccurate, which led to severely understocked shelves and a customer experience that fell far from expectations. Eventually, the situation became so dire that the entire merchandising division was shut down for one week so that employees could manually review and confirm every piece of data in the system to ensure accuracy – hardly a scalable solution. Two years later, Target closed all its 133 locations in Canada, losing 2.5 billion dollars and laying off 17,000 employees.

Of course, poor data quality leads to more than just poor decisions. It causes companies to waste resources, miss opportunities, and spend far too much time fixing their 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 5 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?

Chapter 2

4 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 achieving better data quality was too lengthy and complicated.

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

Myth N°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 take care of 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 systems. 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 and natural language processing capabilities to ease up your work and separate the wheat from the chaff. And 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 N°2

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

Just like data does not come all at once to a company, improving data health is not a one-time operation. 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.

10x

cheaper to fix data quality issues at the beginning of the chain than at the end

[1] The 1-10-100 rule is a quality management concept developed by G. Loabovitz and Y. 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.

talend

Myth No.3

"Data quality is IT's responsibility."

Gone is the time when maintaining healthy, trustworthy data was simply an IT function. Data is the whole company's priority as well as a shared responsibility. 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 to business users because they're the data owners. Business users can then become data stewards 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 modern data strategy.

Myth No.4

"Data quality software is complicated."

As companies are starting 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 well-known 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 & Salesforce, these solutions will dramatically improve your daily productivity and your data flows.

It's time for you to take care of your organization's data health.

Chapter 3

The enterprise challenge: eliminating bad data with pervasive data quality

$100 or more

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

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 just 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, so that tasks can 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 solutions like Talend Data Fabric can simplify these processes even further because data integration, quality, and stewardship capabilities are all part of the same unified platform. Quality, governance, and stewardship can be easily embedded into data integration flows, MDM initiatives, and matching processes to manage and quickly resolve any data integrity issues.

[2] The 1-10-100 rule is a quality management concept developed by G. Loabovitz and Y. 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

5 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 5 key steps for delivering quality data. And 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 5 steps for better data quality are:

Step 01 Profiling

The first step is to really understand what your data looks like. Profiling your data will help you discover data quality issues, risks, and overall trends. Analyzing and reporting on your data in this manner will give you a clear picture of where to focus your data quality improvement efforts. And as time goes on, continued profiling can provide valuable insights into how and where your data quality is improving.

Step 02 Standardizing and matching

Many of the data quality issues uncovered during the profiling process can be fixed through standardization and matching processes. Standardizing data is an essential step when getting data ready for analysis so that all the data being examined is in the same format. Matching lets you associate different records within different systems, and you can even embed matching into real-time processing and make those associations on the fly.

Step 03 Enriching

At this stage, you really start to see the data you've been working with come together. This is where federation can come into play. For example, you may want to use an API to share a particular piece of information, but from your profiling and matching exercises you know that additional related data exists in other locations. Because you've standardized your data and know that it's formatted correctly, you can confidently enrich and augment the data you want to share with the additional related data, so that data users can get a more complete understanding of the information they're consuming.

Step 04 Monitoring

Data quality is not a "one and done" operation. It needs to be a continuous, ongoing practice because the data in your organization is constantly transforming and shifting, and those changes need to be monitored to ensure quality is maintained. When any new quality issues are discovered, you can go back to the previous steps to standardize, match, and enrich that data to get it back on track.

Step 05 Operationalizing

The final step is to operationalize data quality. This is where you really get to see data quality in action. By automating the checks and rules created in the previous steps and embedding them in your data pipelines, you can see significant gains in efficiency and drastically cut down on the amount of bad data requiring manual interventions to fix.

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.

talend

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 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, healthy data can thrive. And the best way to do this is through a secure, unified, cloud-based platform that provides better accessibility, scalability, and reliability, so users can share, operate, and transfer data, actions, and models together.

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

Take control of your data pipelines

Data profiling—the process of gauging the character and condition of data stored across the enterprise—is a vital first step toward gaining control over organizational data.

Choosing a solution that delivers rich functionality and gives you broad and deep visibility into your organization's data can help you:

- Jumpstart your data profiling project with built-in data connectors to easily access a wide range of databases, file types, and applications, all from the same graphical console
- Drill down into individual data sources and view specific records
- Perform statistical data profiling on your organization's data, ranging from simple record counts by category to analyses of specific text or numeric fields to advanced indexing based on phonetics and sounds
- Apply custom business rules to your data to identify records that cross certain thresholds, or that fall inside or outside of defined ranges
- Identify data that fails to conform to specified internal standards such as SKU or part number formats, or external reference standards such as email address format or international postal codes
- Identify non-duplicates or defer to an expert the decision to merge or unmerge potential duplicates

Share quality data without unauthorized exposure

With advanced data quality tools, you can selectively share quality data using on-premises or cloud-based applications without exposing personally identifiable information (PII) to unauthorized people. This not only gives you tools to comply with data privacy regulations like GDPR or CCPA, but it can also protect you from incoming threats and data breaches, as data can be anonymized all along the data lifecycle.

Fix errors with a little help from stewards

Data stewardship is the process of managing the data lifecycle from curation to retirement. It includes defining and maintaining data models, documenting data, cleansing data, and defining rules and policies. It encompasses data governance processes such as monitoring, reconciliation, refining, deduplication, cleansing, and aggregation.

Data stewardship is critical for delivering data-driven insight across the enterprise. And cleaner data contributes to a healthy data environment, leading to more data use while reducing the costs associated with bad data quality.

In addition to improved data integrity, data stewardship helps ensure that data is being used consistently through the organization and reduces data ambiguity through metadata and semantics. Simply put, data stewardship reduces bad data, which translates to better decision-making and reduced costs.

Next-generation data stewardship tools deliver:

- **Self-service –** so that people who know data best are accountable of its quality
- **Team collaboration –** so groups can take advantage of workflows and task orchestration
- **Manual interaction –** so people can certify and validate datasets
- **Integration with data integration and MDM –** to orchestrate human intervention into an automated data pipeline
- **Built-in privacy –** to empower data protection officers to address industry regulations for privacy, such as GDPR

Harness the power of self service

Anyone can be data-driven with self-service data preparation. Self-service is the way to scale data quality standards. Data scientists spend 60% of their time cleaning data and getting it ready to use. Reducing that time and effort means getting more value and more insight from data.

Self-service applications can allow anyone to access a data set and cleanse, standardize, transform, or enrich the data. It's easy to use, so don't have to spend time crunching data in Excel or expect colleagues to do that on their behalf.

Data preparation shouldn't be a separate discipline to make lines of business more autonomous with data; it's a core tool for data quality and integration that drives collaboration between business and IT.

Spend less time scrubbing and more time analyzing

What if you could slash data prep time with a browser-based, point-and-click tool? Look for data preparation tools that can use machine-learning-based smart guides and sampling to quickly identify errors and apply changes to any size data set from any source for export into any target in minutes.

With data preparation tools, you can accelerate your time-to-insight by preparing data as a team, and can share your preparations and datasets or embed data preparations into batch, bulk, and live data integration scenarios. Combined, data integration and data preparation allow business and IT to work together to create a single source of healthy data in the cloud, on premises, or in a hybrid configuration.

Chapter 6

4 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, SSQ Insurance 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 Talend. "Talend offers a complete solution, from data integration to data enhancement, to API-based applications," explains Robert Beauregard, BI Architect.

The company knew that to truly understand its customers, it would need to put healthy, 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 Talend Data Quality, we were able to standardize and clean our data. Talend 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."

Talend has also helped resolve one of the most difficult issues in a master data project: data stewardship, where a human has to take back control from a machine. "Without 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 Talend, combined with the Data Vault 2.0 methodology, we enter production in agile mode every three weeks."

SSQ Insurance now has a unified Customer Center portal, where its contracts display is consolidated in a single location. Customers' operations are automatically registered on the portal, and call centers have access to more comprehensive data. Marketing can now customize its campaigns by using consolidated data to run predictive models. SSQ Insurance has tripled its conversion rates in terms of customer win-back actions. "We now have the necessary foundations to achieve our marketing strategy based on the next best action," says Pelletier. "We can use Talend to build up a 360-degree view of the customer so that we are able to send the right offer to the right customer, via the right channel and with the right message."

SSQ Insurance has tripled its conversion rates in terms of customer win-back actions.

Energy

Developing Customer 360 on a global scale

A global leader in low-carbon energy and services has made 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 of which 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, which, according to the company's Director of Business Acceleration, 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."

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

The organization selected Talend to help it standardize data from 70 entities and in multiple languages, and improve its overall quality. According to the company, "Talend is second to none for trusted data. Prior to its implementation, more than 7% of our data was being rejected. With 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 data 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."

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 healthy asset, so it partnered with 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 healthy 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, 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 life-saving medicines and vaccines to the people who desperately need them.

"For delivering value to our business and delivering medicines to our patients, 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 Talend. "We chose 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," says Hammond.

Travelodge uses 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. Talend's data quality capabilities were used to surface quality issues within Travelodge's internal property data and mediate them, ensuring that the business is only run on healthy data.

"With 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 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,."

Niall Hammond
Data Architect, Travelodge

Chapter 7

Conclusion

Data quality should be a company-wide strategic priority involving professionals from every corner of the business. But to set your organization up for success, you need the right solution that will allow everyone in your organization to contribute to maintaining and improving data quality.

A solution like Talend Data Fabric offers numerous capabilities to promote a healthy data environment. It combines data integration, integrity, and governance in a single, unified platform, with pervasive data quality embedded into every step. And discovering which data needs improvement is easier than ever with the Talend Trust Score™, which instantly assesses the reliability of your data.

Contact us today to see how Talend's solutions can help you meet the data quality goals at your organization.

About Talend

Talend, a leader in data integration and data integrity, is changing the way the world makes decisions.

In order to compete and win, IT and business leaders need data that they can trust and understand instantly. Talend Data Fabric is the only platform that seamlessly combines an extensive range of data integration and governance capabilities to actively manage the health of corporate information. This unified approach is unique and essential to delivering complete, clean, and uncompromised data in real-time to all employees. It has made it possible to create innovations like the Talend Trust Score™, an industry-first assessment that instantly quantifies the reliability of any data set.

Over 6,500 customers have chosen Talend to run their businesses on healthy data. Talend is recognized as a leader in its field by leading analyst firms and industry media.

Talend is Nasdaq listed (TLND) and based in Redwood City, California.

For more information, visit www.talend.com