THE DEFINITIVE GUIDE TO DATA QUALITY

talend

CONTENTS

INTRODUCTION

MANAGING DATA IN A DIGITAL ERA

talend

MORE DATA THAN EVER

Humanity is living through the Data Age. There is more data being produced today than there has been in the previous 5000 years of human history — roughly 2.5 quintillion bytes of data per day. Every time someone emails your business, downloads an app, sends a text message, checks the weather, or does a thousand other quotidian things, data is created, and those thousands of interactions by millions of people create an explosion of information.

Bytes of data created every day:

2,500,000,000,000,000,000

(2.5 quintillion)

of revenue is spent on bad data.

Most CEOs are concerned about data quality.

The insights that a business can extract out of data are only so good as the data itself. Poor data can lead to difficulties in extracting insights and ultimately poor decision-making. This is something that many executives are worried about. According to the Forbes Insights and KPMG "2016 Global CEO Outlook", data and analytics capability ranks the highest of the top five investment priorities, but despite this, CEOs do not have much faith in the data they rely on. 84% are concerned about the quality of the data they're using for business intelligence.

The reasons for their concerns are numerous: integrating new sources of data, particularly unstructured data, with their existing systems; the financial investment and competitive pressure needed to capitalize on all available enterprise data: and the difficulty of extracting data from the silos in which it resides, among others. And their concerns are not unfounded. Harvard Business School released a study which revealed that 47% of newly created data records contain at least one critical error. An astonishing study conducted by MIT Sloan notes that bad data can cost as much as 15-25% of total revenue.

DATA IS A TEAM SPORT

Mitigate rogue and inconsistent data

A major issue for companies in ensuring data quality is the fact that data economics are broken. The sources and volume of data are growing, and so are the number of data professionals who want to work with it. The impact of this data proliferation across a growing number of clouds and digital channels, and across an increasing number and variety of people, put the enterprise at risk for data leaks, data breaches, rogue and inconsistent data.

As an example, 62% of end users admit they have access to data they should not. This becomes crucial as new data governance regulations are being implemented, which may have concrete impact on business; for example, the fine for violating the European Union's General Data Protection Regulation (GDPR) is 4% of the organization's worldwide turnover.

There is a way to solve this problem. Organizations need to make sure that good quality data is available to everyone who needs it. They shouldn't have to rely on a small IT team or a couple of rockstar data personnel to make that happen. Data is a team sport; everyone from IT to data scientists to application integrators to business analysts should be able to participate and extract valuable insights out of constantly available , good quality data.

In this Definitive Guide, we're going to talk about the ingredients that go into creating quality data, how enterprises can ensure that all of their data is of good quality, and how to make that quality data available to anyone who needs it in a secure and governed fashion. You no longer have to let bad data cost your organization time and money.

Organizations need to make sure that good quality data is available to everyone who needs it and shouldn't have to rely on a small IT team or a couple of rockstar data personnel to make that happen.

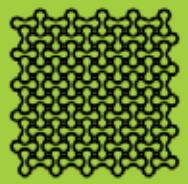
CHAPTER 1

WHAT IS DATA QUALITY?

talend

ACCURACY

COMPLETENESS



CONSISTENCY

TIMELINESS

UNIQUENESS

VALIDITY

NOT ALL DATA IS CREATED EQUAL

Data quality is the process of conditioning data to meet the specific needs of business users. Accuracy, completeness, consistency, timeliness, uniqueness and validity are the chief measures of data quality.

Data quality is measured on six critical factors, each of which is equally important.

NOT ALL DATA IS CREATED EQUAL

Bad data can come from every area of your organization under diverse forms from business departments, sales, marketing or engineering.

Improving the six data metrics

Data quality initiatives are generally centered on improving these metrics so that data will promote optimal performance of business systems and support user faith in the systems' reliability.

Regardless of an organization's size, function, or market, every organization needs to pay attention to data quality in order to understand its business and to make sound business decisions. The effectiveness of decisions is unquestionably impacted by the quality of data being used.

But, as we have noted, the kinds and sources of data are extremely numerous, and its quality will have different impacts on the business based on what it's used for and why. Data's value comes primarily when it underpins a business process or decision-making based on business intelligence. Therefore, the agreed data quality rules should take account of the value that data can provide to an organization. If it is identified that data has a very high value in a certain context, then this may indicate that more rigorous data quality rules are required in this context.

Companies therefore must agree on data quality standards based not only on the dimensions themselves — and, of course, any external standards that data quality must meet — but also on the impact of not meeting them.

BAD DATA IS...

Inaccurate

Data that contains misspellings, wrong numbers, missing information, blank fields

Noncompliant

Data that doesn't meet regulatory standards

Uncontrolled

Data left without continuous monitoring becomes polluted over time

Unsecured

Data left without controls, becoming vulnerable to access by hackers

Static

Data that is not updated, thus becoming obsolete and useless

Dormant

Data that is left inactive and unused in a repository loses its value as it's neither updated nor shared

CHAPTER 2

WHY IS DATA QUALITY SO HOT RIGHT NOW?

BAD DATA IS ON THE RISE

Bad data has never been such a big deal. The old adage – garbage in, garbage out (GIGO) – holds true. Poor data quality adversely affects all organizations on many levels, while good data quality is a strategic asset and a competitive advantage to the organization.

If data fuels your business strategy, bad data can kill it.

As we've noted, the amount of data that exists today is truly staggering. According to IDC's latest 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 the next seven years, the global data sphere is expected to grow to 10 times the 2016 data sphere. 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.

According to Gartner, poor data quality costs rose by 50% in 2017, reaching 15 million dollars per year for every company. You can imagine this cost will explode in the upcoming years if nothing is done.

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

EXAMPLES OF ORGANIZATIONAL IMPACTS INCLUDE:

Incorrect email addresses

Would have a significant impact on any marketing campaigns

Inaccurate personal details

May lead to missed sales opportunities or a rise in customer complaints

Incorrect shipping addresses

Goods can get shipped to the wrong locations

Incorrect product measurements

Can lead to significant transportation issues

i.e. the product will not fit into a truck, alternatively too many trucks may have been ordered for the size of the actual load.

CHAPTER 2: **WHY IS DATA QUALITY SO HOT RIGHT NOW?**

BAD DATA AFFECTS YOUR DAILY LIFE

Bad data in the news

Let's take a look at a recent "bad data" example from the news. A man followed his GPS application as he was driving, but because of some bad data, he wound up driving directly into a lake rather than the destination he intended.

Data quality can be a threat but also an opportunity

Now let's visualize a future where your car will be powered by machine learning capabilities. It will be fully autonomous and will choose directions and optimize routes on its own. If the car drives you into the lake because of poor geo-positioning data. This will end up costing the carmaker quite a bit in repairs and even more to brand reputation.

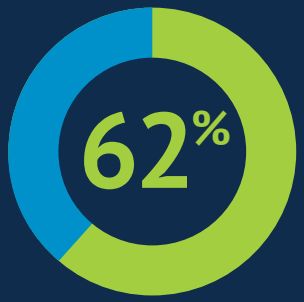
It's then very likely that poor data quality will continue to affect brand reputation if nothing is done. It will then certainly become a corporate responsibility to deliver trusted and governed data. That means everybody is the company should be aware of their individual responsibility.

At the same time, by improving data quality through modern collaborative approach and putting data into the hands of data citizens, you have the unique opportunity to add value to your brand and gain competitive advantage.

talend

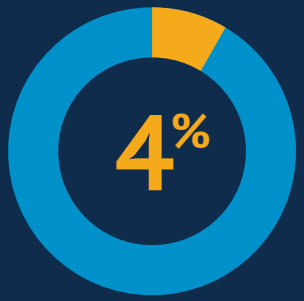
Extra cost to complete unit of work with flawed data

Exploding data volumes causes huge data quality issues.



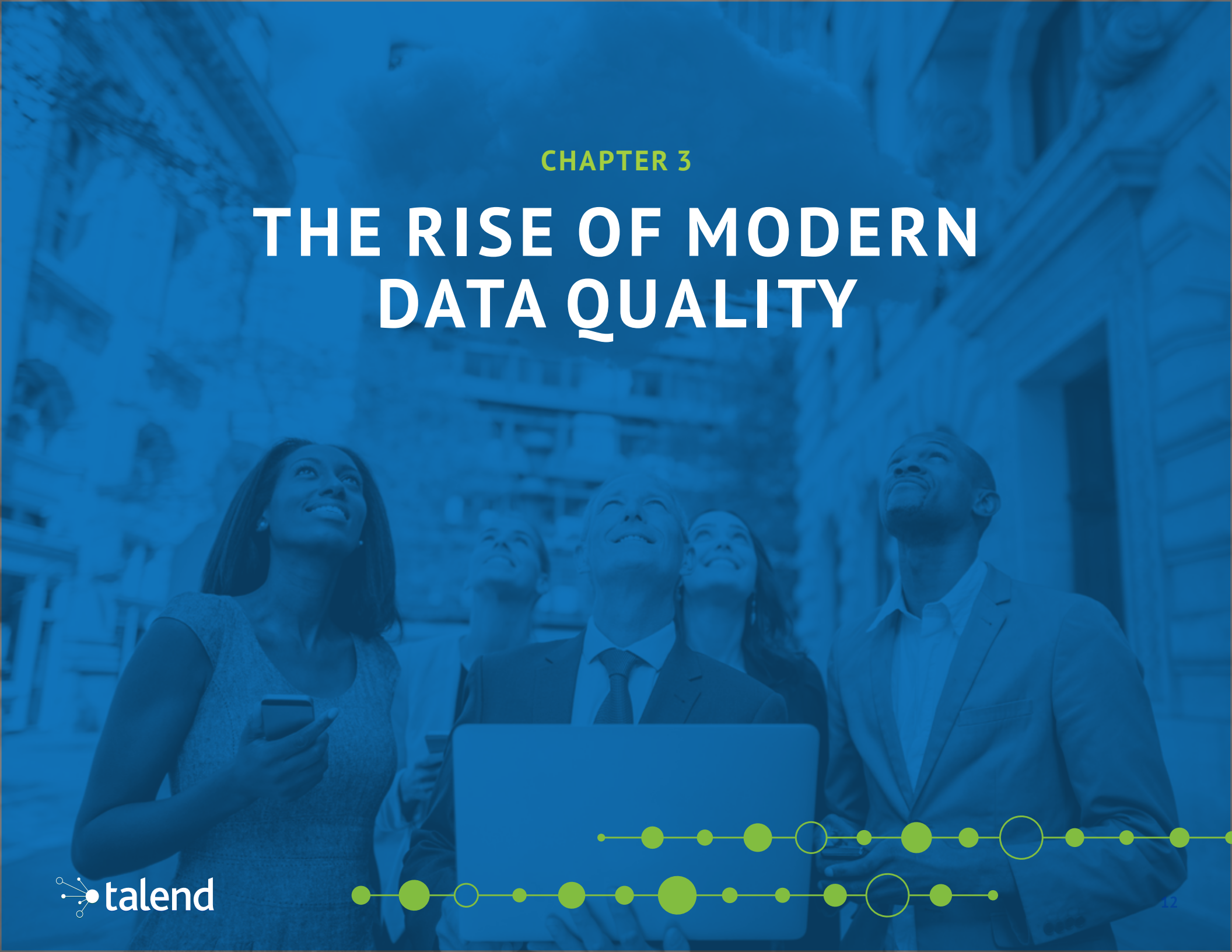
Have access to data they should not

Employees need immediate access to data to meet their goals.



GDPR fines can be up to 4% of worldwide turnover

New regulations penalize companies that fail to manage



CHAPTER 3

THE RISE OF MODERN DATA QUALITY

talend

PERVASIVE DATA QUALITY

Embed real-time data quality in all business processes and applications.

The cost of doing nothing explodes over time.

Poor data quality can be mitigated much more easily if caught before it is used — at its point of origin. If you verify or standardize data at the point of entry, before it makes it into your back-end systems, we can say that it costs about $1 to standardize it. If you cleanse that data later, going through the match and cleanse in all the different places, then it would cost $10 in comparison to the first dollar in terms of time and effort expended.

And just leaving that bad quality data to sit in your system and continually give you degraded information to make decisions on, or to send out to customers, or present to your company, would cost you $100 compared to the $1 it would've cost to actually deal with that data at the point of entry, before it gets in. The cost gets greater the longer bad data sits in the system.

Organizations need to take a proactive approach to their data quality process.

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.

PERVASIVE DATA QUALITY

Ensure, analyze, and monitor data quality at its source

A proactive approach to data quality allows you to check and measure that level of quality before it even really gets into your core systems. Accessing and monitoring that data across internal, cloud, web, and mobile applications is a pretty big task. The only way to scale that kind of monitoring across all of those systems is through data integration. It therefore becomes necessary to control data quality in real time.

Of course, avoiding the propagation of erroneous data by inserting control rules into your data integration processes is key. With the right tools and integrated data, you can create whistleblowers that detect some of the root causes of overall data quality problems.

Then you will need to track data across your landscape of applications and systems. That allows you to parse, standardize, and match the data in real time. You can organize the activity to check the correct data whenever needed.

Data stewardship: delegate the data to the people who know it best.

More than a tool just for data stewards with specific data expertise, IT can empower business users to use a point-and-click, Excel-like tool to curate their data. With Talend Data Stewardship you can manage and quickly resolve any data integrity issue to achieve trusted data across the enterprise.

With the tool, you define common data models, semantics, and rules needed to cleanse and validate data, then define user roles, workflows, and priorities, and delegate tasks to the people who know the data best. Productivity is improved in your data curation tasks by matching and merging data, resolving data errors, certifying, or arbitrating on content.

Because it is fully integrated with the Talend Platform, it can be associated to any data flow and integration style that Talend can manage, so you can embed governance and stewardship into data integration flows, MDM initiatives, and matching processes.

talend

TALEND TIPS

Use built-in smart features to accelerate your data controls. Today, almost everyone has big data, machine learning, and cloud at the top of their IT "to-do" list. The importance of these technologies can't be overemphasized, as all three are opening up innovation, uncovering opportunities, and optimizing businesses.

Tasks that used to be done by data professionals, such as data experts, now done by operational workers that know the data best, is called self-service. It requires workflow-driven, easy-to-use tools with an Excel-like UI and smart guidance.

"Machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; These analytical models allow researchers, data scientists, engineers, and analysts to produce reliable, repeatable decisions and results."

Machine learning isn't a brand new concept. Simple machine learning algorithms actually date back to the 1950s, though today they are subject to large-scale data sets and applications.

INTELLIGENT DATA QUALITY

Smarter tools make people smarter.

Because of large unstructured data volumes, there is a need to have automated help and assistance when proceeding with data analytics. This clearly allows users to accelerate their time-to-insights when dealing with various data sources and data formats.

Machine learning plays a key role in this collaborative process. Some people think that AI might replace people and kill jobs. But for data collaboration, it is rather about augmenting the way that people can work together with data.

Utilize smart semantics.

Talend combines smart semantics and machine learning to turn data into insights faster. Smart semantics automatically capture data footprints in a data pipeline to accelerate data discovery, data linking, and quality management.

Machine learning helps to suggest the next best action to apply to the data pipeline or capture tacit knowledge from the users of the Talend platform (such as a developer in Talend Studio, or a steward in Talend Data Stewardship) and run it at scale through automation.

Put your best foot forward.

Talend leverages machine learning and advanced analytics to guide the users in their data journey by suggesting next best actions. It improves developers' productivity and empowers non-data experts to work with data as well. An example might be to engage lines of business in data quality initiatives.

Machine learning also allows the capture of knowledge from business users and data professionals. One typical use case is data error resolution and matching. By using self-service tools such as Talend Data Stewardship for applying manual error resolution on a data sample and then applying machine learning to the whole data set into a fully automated process, Talend turns low value and time-consuming tasks into an automated process you can apply at scale to millions of records.

CHAPTER 3: **THE RISE OF MODERN DATA QUALITY**

COLLABORATIVE DATA QUALITY

Data quality is a collaborative effort

As we now know, data isn't the responsibility of one department or another. To ensure its quality and its value, it has to be owned by the whole organization and managed collaboratively. And it must also be noted that data leaders cannot master the opportunities and challenges of digital transformation with yesterday's roles and organizations. Data is a team sport that spans across roles and lines of business.

The cloud drastically extends the boundaries of data. Lines of business bring their own applications, and products, people, and assets create their own data pipelines through the web and the Internet of Things. Data can also be exchanged seamlessly between business partners and data providers.

This brings huge opportunities but challenges as well. When data comes from everywhere, and when everyone wants to access it, you need to control its quality and usage.

According to Harvard Business Review, it costs 10x as much to complete a unit of work when the data are flawed in any way as it does when they are perfect.

Data quality has always been a challenge in system integration. It's also been highlighted as one of the top 3 challenges to establishing data warehouses, ERP, or CRM systems. Not only haven't we fully solved the issue, but it is getting worse with Big Data and cloud.

The data quality solution isn't just about how it's approached or the tools you use; data quality has to be collaborative and the responsibility has to be distributed across teams.

talend

THERE'S A BETTER WAY TO HANDLE DATA

How can we provision access to data for a large and diverse consumer community in a controlled but flexible fashion?

New roles willing to access data.

There is a significant increase in the people demanding access to data to perform their job roles.

We had the IT developers, business analysts and business users. Then came data engineers and data scientists. Now, regulations like GDPR are mandating accountability on data protection with new roles like data stewards and data protection officers.

When the value of data increases, so does the risk.

When data proliferates across an increasing variety of people, the risks of data leaks, data breaches, fake news, and rogue and inconsistent data increases. And, when you lose control of your data, you can generate negative press, lose your job, or even kill your business.

We believe there is a better way to manage data quality collaboratively and safely.

The solution: governed data access.

The answer lies in a unified platform as a place to orchestrate collaboration between data professionals in a governed way.

EXAMPLE

USAGE OF WEATHER DATA

Think about a company that wants to consider weather data to improve the precision of their sales forecast. This could start with a data scientist in a data lab. The data scientist might provision weather data in the lab to refine its forecasting model. Once he confirmed that weather data would impact his model, this new dataset might be checked for quality, compliance and copyrights by a data steward, and then finally, an IT professional would automate data integration, so that weather data flows close to real time in the corporate data lake and documented, into the data catalog so that it can be consumed by business analysts and business users.

You can have a more siloed approach considering that IT should bring the data using data integration tools. Then the data scientists in their data lab would use a data science platform, while the office of the CDO would use a data governance framework. But how could they work as a team with a siloed approach? And who could control this disparate set of practices, tools and datasets?

This is what collaborative data management is all about — allowing people to work as a team to reap all the benefits of your data.

BENEFITS OF A COLLABORATIVE APPROACH

Operating your data team well means a **reduction in cost and time to market**.

Less time. Fewer costs. Better data.

The main goal is to reduce the time it takes to deliver data in and making it twice as valuable at half the cost. Collaborative data management can cut the cost for developing new apps and managing change. This is done through having lines of business self-serve access to data, while having developers focusing on more complex tasks.

Although collaborative data management can reap benefits in on-premises projects, we see its value even increasing when combined with cloud data management, as the cloud helps to cut the time and costs associated with data operations. In addition, managing data in the cloud accelerates change due to the speed at which cloud systems can be changed and updated.

CHAPTER 4

MAKE DATA A TEAM SPORT TO TAKE UP YOUR DATA CHALLENGES

talend

TIMES ARE CHANGING

Data quality is often perceived as an individual task of the data engineer. As a matter of fact, nothing could be further from the truth.

People close to the business are eager to work and resolve data quality issues as they're the first to be impacted by bad data. But they are often reluctant to update data as data quality apps are not really made for them or they are not allowed to use them. That's one of the reasons bad data keeps on increasing. According to Gartner, poor data quality costs rose by 50% in 2017, reaching 15 million dollars per year for every company. This cost will explode in the upcoming years if nothing is done.

It's time to huddle up and make a game plan.

Data quality is now increasingly becoming a company-wide strategic priority involving professionals from every corner of the business. To succeed, working like a sports team is a way to illustrate the key ingredients needed to overcome any data quality challenge.

As in team sports, you will hardly succeed if you just train and practice alone. You have to practice together to make the team successful.

Also, just as in team sports, Business/IT teams require having the right tools, taking the right approach and asking committed people to go beyond their daily tasks to tackle the data quality challenge one step at a time.

It's all about strengthening your data quality muscles by challenging IT and the rest of the business to work together. For that, you need to proceed with the right model, the right process and the right solution for the right people.

It'll be easier than you think to drive alignment now that the links between good data and good performance are so well-documented.

CHAPTER 4: **MAKE DATA A TEAM SPORT TO TAKE UP YOUR DATA CHALLENGES**

CHOOSING THE BEST GOVERNANCE MODEL FOR YOU

Migrate from an Encyclopedia Universalis to a Wikipedia model

Being data-driven is not optional for enterprises today, but there are tough challenges to overcome in order to get the most value out of data. There is exponential growth in the volume of data companies deal with, which doubles every 2 years. At the same time, there is more variety of data, such as new streaming data coming from the IoT, sensors, web logs, click streams, etc.

And there is a multiplication of new data-driven roles within your organization. Back in the early 2000s, we had IT developers and business analysts. And then came new data professionals like data stewards, data scientists, data curators, data protection officers, and data engineers. Today, even business users have become data-savvy and want to join in to turn their data into insights in an autonomous way.

However, IT's budget and resources are relatively flat. There is a growing gap between business expectations and what IT can deliver. What this means is that the economics of data integration are broken. There is an exponential growth in the volume of data and more variety of data from modern data sources. In fact, there is a 35% CAGR for streaming analytics, as companies need to ingest and analyze real-time data, instead of reacting to day or week-old data.

talend

20 BILLION

IoT devices deployed by 2020

10x

Faster growth rate for human-generated data

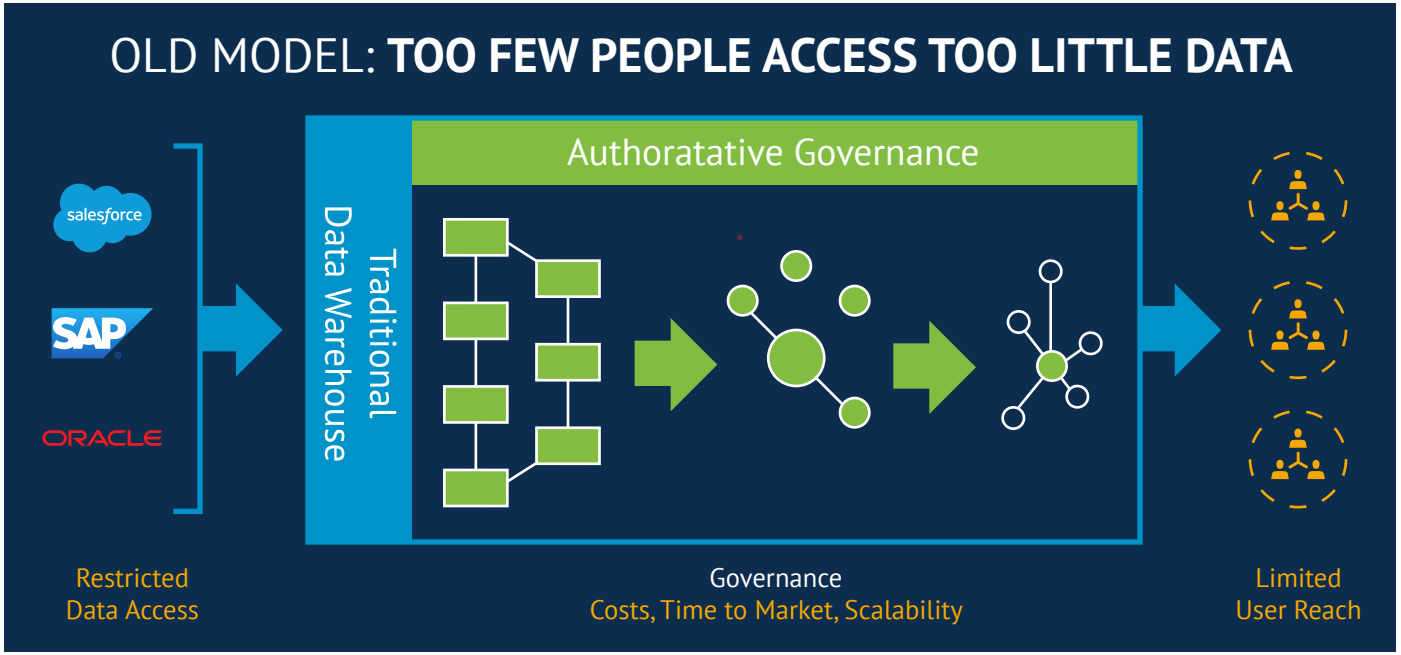
compared to traditional business data

50x

Faster growth rate for machine-generated data

compared to traditional business data

THE ENCYCLOPEDIA BRITANNICA MODEL



Too few people access too little data

The encyclopedia model fails to scale in this big data era, when multiple people demand immediate access to the right data from every source.

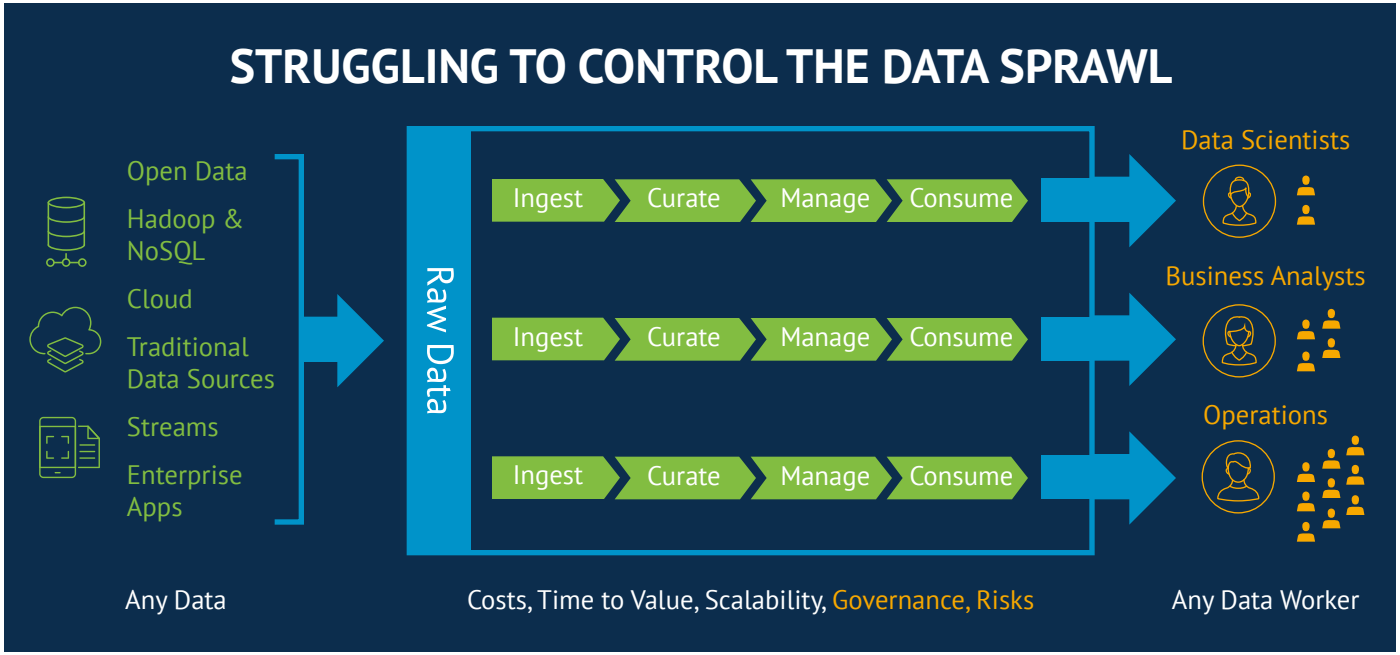
The old model was about allowing a few people to access a small amount of data. This model worked for many years to build data warehouses. The model relies on a team of experienced data professionals armed with well-defined methodologies and well-known best practices. They design an enterprise data warehouse, and then they create data marts, so the data can fit to a business domain. Finally, using a business intelligence tool, they define a semantic layer such as a "data catalog" and predefined reports. Only then can the data be consumed for analytics.

We can compare this model to the encyclopedia model. Before we entered the 21st century, we had encyclopedia such as Encyclopedia Britannica, or Microsoft Encarta. Only a handful of experts could author the encyclopedia.

Of course, the quality is great- for example, Encyclopedia Britannica is written by about 100 FTE editors together with around highly skilled 4000 contributors, including Nobel prize winners and former US presidents. But the encyclopedia model fails to scale in this big data era where you always want accurate articles on each and every topic, in your native language.

Your organization is facing the exact same issue with its data; you do not have enough resources to bring all this data in accurately, nor can you address the growing needs of the business users.

THE HYPER COLLABORATIVE & UNGOVERNED MODEL



Struggling to control the data sprawl

This more agile model has multiple advantages over the previous one. It scales across data sources, use cases and audiences. Raw data can be ingested as it comes with minimal upfront implementation costs, while changes are straightforward to implement.

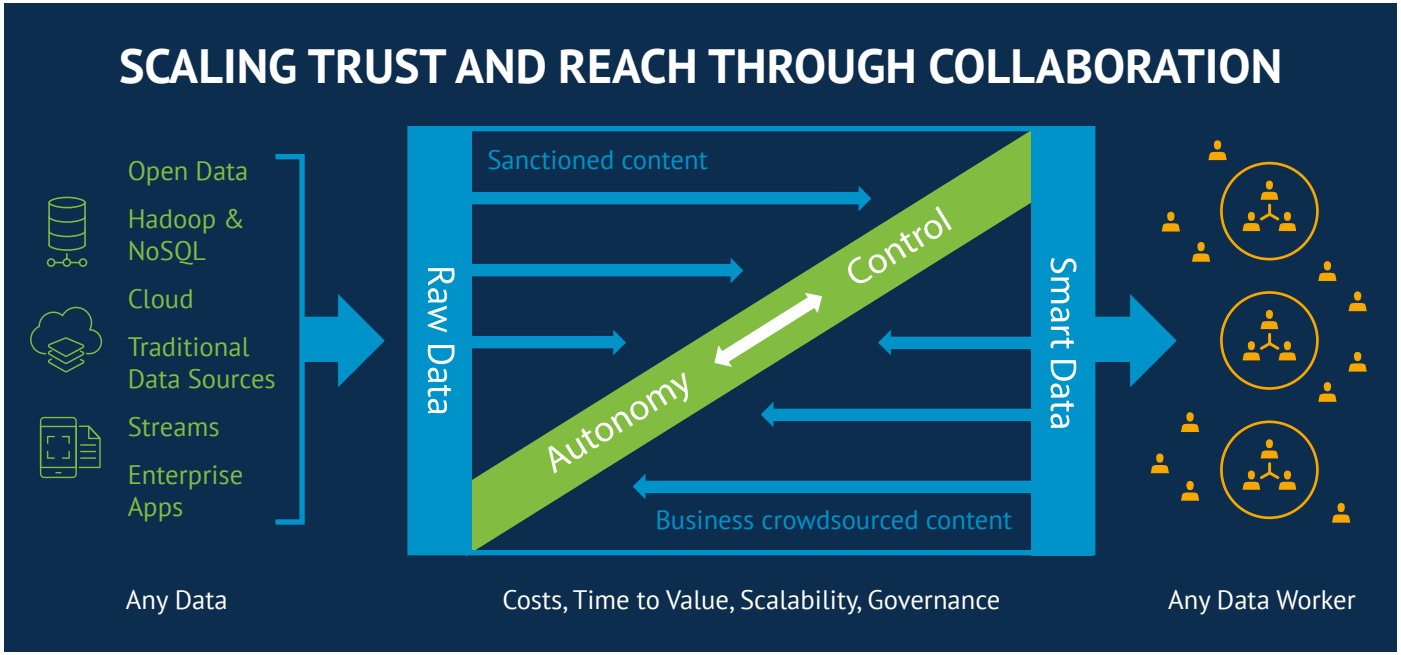
Data lakes then came to the rescue as an agile approach for provisioning data. You generally start with a data lab approach targeting a few data-savvy data scientists. Using cloud infrastructure and big data, you can drastically accelerate the data ingestion process with raw data. Using schema on read, data scientists can autonomously turn data into smart data.

The next step is to share this data with a wider audience, so you create a new data layer for analytics, targeting the business analyst community. Because you are targeting a wider audience with different roles, you then realize that you need to establish stronger governance. Then, the third next step is to deliver information and insights to the whole organization. Again, the prerequisite is to establish another layer of governance.

This more agile model has multiple advantages over the previous one. It scales across data sources, use cases, and audiences. Raw data can be ingested as it comes with minimal upfront implementation costs, while changes are straightforward to implement.

But through this approach, we created a big challenge, as data governance was not considered alongside this way of doing things. This is like what Facebook is ecurrently experiencing with their data practices. They created a platform that has no limits in terms of content it can ingest and number of users and communities it can serve. But because anyone can enter any data without control, control is almost impossible to establish. Facebook says it will hire 10,000 employees in its trust and safety unit and plans to double those headcounts in the future.

THE COLLABORATIVE & GOVERNED MODEL



A scalable, managed third way

You need to establish a more collaborative approach in parallel, so that the most knowledgable among your business users can become content providers and curators.

It's important to work with data as a team from the start. Otherwise, you may become overwhelmed by the amount of work needed to validate trusted data. By introducing a Wikipedia-like approach where anyone can potentially collaborate in data curation, there is an opportunity to engage the business in contributing to the process of turning raw data into something that is trusted, documented, and ready to be shared.

By leveraging smart and workflow-driven self-service tools with embedded data quality controls, you can implement a system of trust that scales.

IT and other support organizations such as the office of the CDO need to establish the rules and provide an authoritative approach for governance when it is required (for example for compliance, or data privacy.)

You need to establish a more collaborative approach in parallel, so that the most knowledgeable among your business users can become content providers and curators.

CHAPTER 4: **MAKE DATA A TEAM SPORT TO TAKE UP YOUR DATA CHALLENGES**

BUILDING YOUR DATA QUALITY TEAM

Identify the people who will become your data champions.

Do not underestimate the human factor once you want to deliver your data management project. Remember: your data quality project is the right combination of having the right people equipped with the right tools following the right approach. What is specific regarding Data Quality is that your team must come from different departments. You will not succeed if only IT is in charge. It's the same for business, who will need IT knowledge, skills and governance to avoid any shadow IT and low governance projects.

BUILDING YOUR DATA QUALITY TEAM

"In God we trust; all others bring data."

It's important to help everyone understand how important a corporate asset data is.

Identify your data disruptors to create a data culture. In each organization, department, or division, spot your soon-to-become data-driven champions. You will need them not only to instill a data-driven culture within your company but also to make people responsible for the data they use, manage, and own.

To succeed, you must work as a team

Whatever the size, the culture, and the success of the company, cross collaboration is one of the most difficult challenges. Data management is no exception. You need to spend extra time on transversal projects that go beyond your individual objectives. It will require executive sponsorship, individual and team efforts.

Find inspiration in Six Sigma and keep it simple

Six Sigma is a systematic approach to stamping out defects. Defects stem from what Six Sigma calls "variation", which is the bane of quality. We don't necessarily need to have a sophisticated Six Sigma driven approach for any Data Quality Project. But there are some useful tips to apply that would help any team wishing to successfully deploy data quality.

DEPLOYMENT LEADER & CORE TEAM

The Team Leader

Leaders must have the big picture of the project in mind. They are often considered as the relationship interface between departments, executives and the data quality team. They should support and encourage team visibility along the deployment process.

Team Members: Business & IT Together

Your core team must be composed of business, data protection and IT people. By putting people close to the business together with Data Specialists, you will make sure your project is natively business driven.

Make sure your team includes data governance professionals, as they will be key in documenting your data as well as identifying and mapping systems containing sensitive and confidential information that could put the enterprise at risk.

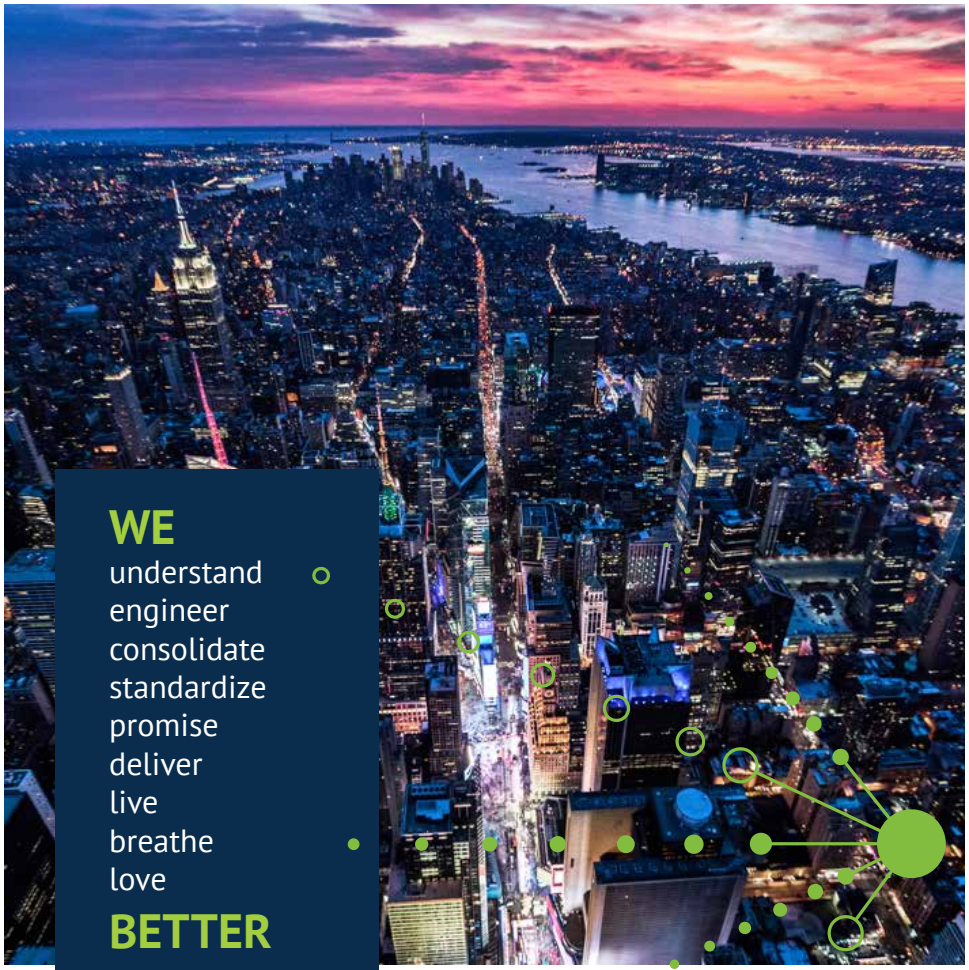
Some regulatory projects such as "portability" or the "right to be forgotten" in the context of the GDPR will require tight collaboration between business, marketing, IT and data protection. That may be your next project for your data quality team.

TALEND TIPS

If you are the leader, it's crucial that you are in control and that you introduce good governance to kick off your team-driven data quality project. It's also important to celebrate success each time the core team wins a victory. Be prepared to lead and facilitate collaboration.

If you're part of the core team, make sure you meet often. Set up regular meetings in both formal and informal places. Keep in mind the team members come with different technical backgrounds, a diverse knowledge base, and a variety of business cultures. A data privacy officer with a legal background will have a different point of view from a Data Architect with an IT background. One will know what data is at risk while the other will know how to operate the data. All this enriches any data quality project. Your success will then deeply rely on the tight collaboration between the two.

TRUST & RESPONSIBILITY WITH YOUR DATA



Keep fostering data responsibility.

Technology, products, sensors, and objects to be created in the coming years will heavily use data. The data they use needs to be trustworthy, so that each of your company's data professionals can feel committed and responsible for delivering good quality data that will help the business. This is a shared responsibility. Help data teams understand that quality data is an asset, and make sure they follow compliance training that will make them aware of your data's value as well as the consequences of data misuse.

Trusted data starts with trustworthy people.

Trust is at the heart of any data project. And this also applies to every team member. Trust is not a given but rather a continuous perception; team members must feel they can rely with confidence on someone else to delegate a task assigned to anyone on the team. Trusted data always relies on trustworthy people. This is what we mean when we say data is a team sport; trust is the essence of a successful data management project. To create that trust among team members, we suggest applying the RUMBA methodology. Taught in Six Sigma Training, RUMBA is a powerful acronym that stands for Reasonable, Understandable, Measurable, Believable, and Achievable.

By applying this principle within your team from the start, you will make sure you will get not only commitment from the top, but also from your team, as you won't discourage them from setting unreachable targets. Create lofty goals but make sure that your team has the tools to get there.

CHAPTER 5

CHOOSING THE RIGHT TOOLS

talend

FIGHT COMPLEXITY WITH A CLOUD UNIFIED PLATFORM

There is a plethora of standalone data quality tools on the market.

Register for any big data tradeshow and you will discover plenty of data preparation and stewardship tools offering several benefits to fight bad data. But only a few of them cover data quality for all.

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 only IT people can manage them. While these tools can be powerful, If you have short term data quality priorities, you will miss your deadline. Don't ask a rookie to pilot a jumbo jet. The flight instruments are obviously too sophisticated and it won't be successful.

On the other hand, you will find simple and often robust apps that can be too siloed to be injected into a comprehensive data quality process. Even if they successfully focus on the business people with a simple UI, they will miss the big part — collaborative data management. And that's precisely the challenge. Success relies not only in the tools and capabilities themselves, but in their ability to talk to each other. You therefore need to have a platform-based solution that shares, operates, and transfers data, actions, and models together. That's precisely what Talend provides.

You need a unified platform in the cloud.

You will confront multiple use cases where it will be impossible for one person or team to manage your data successfully. Working together with business users and empowering them on the data lifecycle will give you and your team superpowers to overcome traditional obstacles such as cleaning, reconciling, matching or resolving your data.

Click to learn more

TAKE CONTROL OF YOUR DATA PIPELINES WITH TALEND STUDIO

It's time to analyze your data environment.

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

Talend Studio delivers rich functionality that gives you broad and deep visibility into your organization's data:

- Jump-start 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.
- Use the Data Explorer to 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 forms, or external reference standards such as email address format or international postal codes.

Improve your data with standardization, cleansing and matching.

Data profiling—the process of gauging the character and condition of data stored in various forms across the enterprise—is commonly recognized as a vital first step toward gaining control over organizational data. It also allows you to identify non-duplicates or defer to an expert the decision to merge or unmerge potential duplicates.

Share quality qata without unauthorized exposure.

With Talend 's data quality tools, you can selectively share production quality data using on premises or cloud-based applications without exposing Personally Identifiable Information (PII) to unauthorized people. Not only you will have tools to comply with data privacy regulations such as GDPR, but this will prevent you from incoming threats and data breaches as data will be anonymized all along the data lifecycle.

Click to learn more

FIX ERRORS WITH A LITTLE HELP FROM STEWARDS

What is data stewardship?

As a critical component of data governance, data stewardship is the process of managing the data lifecycle from curation to retirement.

Data stewardship is about defining and maintaining data models, documenting the data, cleansing the data, and defining its rules and policies. It enables the implementation of well-defined data governance processes covering several activities including monitoring, reconciliation, refining, de-duplication, cleansing and aggregation to help deliver quality data to applications and end users.

Data stewardship is becoming a critical requirement for successful data-driven insight across the enterprise. And cleaner data will lead to more data use while reducing the costs associated with "bad data quality" such as decisions made using incorrect analytics.

Data stewardship benefits the organization.

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" in your company, which translates to better decision-making and redued costs.

Organizations need modern data stewardship.

With more data-driven projects being launched, "bring your own data" projects by the lines of business, and increased use of data by data professionals in new roles and in departments like marketing and operations, there presents a need to rethink data stewardship. Next-generation data stewardship tools such as Talend Data Stewardship deliver:

Self-service – so that people who know data best are accountable of its quality

Team collaboration – including workflow and task orchestration

Manual interaction – where people are required to certify and validate a dataset

Integration with data integration and MDM – orchestrating human intervention into an automated data pipeline

Built-in privacy – empowering data protection officers to address industry regulations for privacy such as GDPR (General Data Protection Regulation)]

Click to learn more

USE THE POWER OF SELF-SERVICE WITH TALEND DATA PREPARATION

Anyone can be data-driven with self-service data preparation.

Self-service is the way to get data quality standards to scale. Data scientists spend 60% of their time cleaning data and getting it ready to use. Reduced time and effort means more value and more insight to be extracted from data.

Talend Data Preparation deals with this problem. It is a self - service application that allows potentially anyone to access a data set and then cleanse, standardize, transform, or enrich the data. Because it is easy to use, it solves a pain point in organizations where so many people are spending so much time crunching data in Excel or expecting their colleagues to do that on their behalf.

Data preparation is not just a separate discipline to make lines of business more autonomous with data; it's a core element for data quality and integration. And it will become a critical part of most data integration efforts. Although it improves personal productivity, the true value of Data Preparation is to drive collaboration between business and IT.

Spend less time scrubbing and more time analyzing.

What if you could slash that time with a browser-based, point-and- click tool? Data Preparation uses 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 instead of hours.

Accelerate data usage throughout the company.

Accelerate your time-to-insight by preparing data as a team with Data Preparation. You 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 trusted data in the cloud, on premises, or hybrid.

Click to learn more

CHAPTER 6

HOW OUR CUSTOMERS TACKLE DATA QUALITY CHALLENGES

talend

CHAPTER 6: **HOW OUR CUSTOMERS TACKLE DATA QUALITY CHALLENGES**

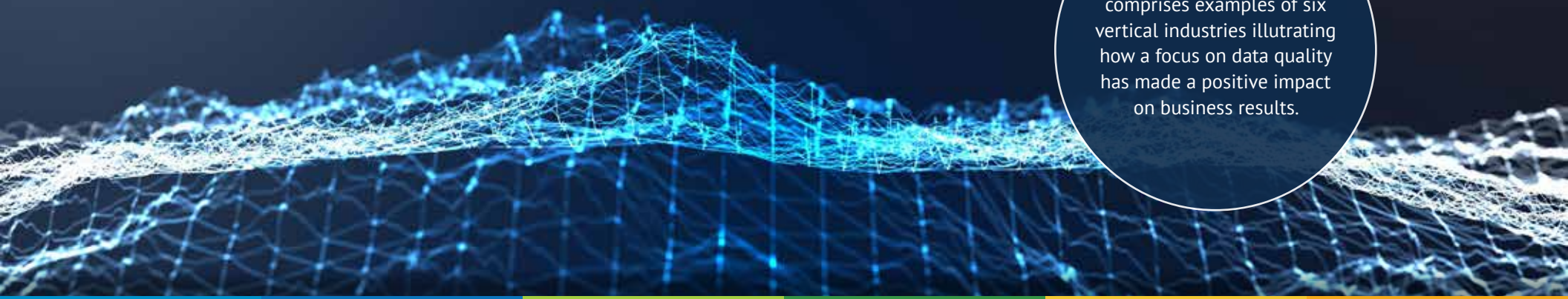
WEATHERING THE DATA TSUNAMI

As more and more businesses get ready to face the data tsunami, data quality is becoming an ever-more critical success factor. Yet, **84% of CEOs are concerned about the quality of the data** they're basing critical business decisions on, according to KPMG's "2016 Global CEO Outlook."

Making data quality a priority

Procedures and perceptions about data quality at many organizations—what it is, how to improve it, and how to institutionalize it—haven't kept pace with its growing importance. Many Talend customers, however, are the exception—they've made data quality a priority and are receiving the benefits.

The rest of this chapter comprises examples of six vertical industries illutrating how a focus on data quality has made a positive impact on business results.



RETAIL

TRANSPORTATION

CHARITY

CONSTRUCTION

MARKETING

MEDIA

TRANSPORTATION: BEING COMPLIANT WITH REGULATIONS

Air France-KLM relies on Talend to cater to each customer

A world leader in transporting passengers and cargo, Air France-KLM needs high-quality data to meet its goal of catering to every one of its customers. It needs accurate phone numbers and emails, which are essential for booking flights, and it needs to reconcile online and offline information, since visitors to these sites and applications are most often connected to their personal accounts. Data management was a challenge, and Air France-KLM resolved to get organized in order to ensure data quality, respect the privacy of its customers and offer customers and employees a clear benefit.

In addition, Air France-KLM collects and processes personal data (PII, or personally identifiable information) concerning passengers who use the services available on its website, its mobile site, and its mobile applications. The company is committed to respecting privacy protection regulations regarding its passengers, loyalty program members, prospects and website visitors. All personal data processing is therefore carried out with Talend Data Masking, which makes it possible to anonymize certain sensitive data and make it unidentifiable in order to prevent unauthorized access.

Every month, a million pieces of data are corrected with Talend Data Quality—proof of the essential role the Talend solution plays in ensuring Air France-KLM can deliver on its promise to cater to every customer. Their success also proves that it's no longer true that "data quality falls under IT responsibility"; rather, it's a business priority with such goals as respecting customer privacy.

Air France-KLM corrects **1 million** pieces of data each month with Talend.

CHARITY: MANAGING DONOR DATA, SAVING LIVES

Save the Children UK uses Talend to manage data

Save the Children UK (SCUK) saves lives by preparing for and responding to humanitarian emergencies. The charity has been using Talend Data Quality to dedupe data being imported into the database of donors, and to review existing CRM data.

By reducing duplicates and improving data quality, the charity ensures the information it has on an individual is as accurate as possible. That, in turn, aids in ensuring that donors receive only the information they ask for, and allows SCUK to manage the flow of messages to them in a truly relevant manner.

Improved data quality also helps SCUK avoid alienating its donors. For example, if the charity has three records for a J. Smith, John Smith, and J. Smyth with slight variations in the held addresses, and it's all the same person, they might mail him three times for the same campaign. That costs SCUK money, and may prompt the donor to say they do not wish to be contacted anymore. In addition, efficiently

importing higher-quality data supports the production of better, faster reporting by analysts and provides SCUK greater insight into donor behavior and motivations.

SCUK's commitment to an ongoing campaign to maintain a clean, accurate donor record shows that ensuring data quality is a process and not an event, and that "once you solve your data quality problem, you're done," is an outdated misperception.

"Talend is helping us quickly and accurately import data to our CRM in an automated way."

— Penny Kenyon,
CRM Import & Integrity Manager

RETAIL: RIGHT PRODUCT, RIGHT PLACE, RIGHT TIME

Travis Perkins manages product and customer data with Talend

When Travis Perkins started their data quality journey company data was siloed and not maintained or validated in any consistent way. As the company was moving into a multichannel world, focusing more on online sales, data quality was key. Relying on assorted employees and suppliers to enter product information resulted in incomplete and inconsistent details—and while data was supposed to be manually reviewed and approved, that didn't always happen.

Travis Perkins adopted Talend Data Quality to provide a data quality firewall that checks for duplicates, confirms that check digits are valid for barcodes, standardizes data, and maintains a consistent master list of values.

Since deploying the solution, 500,000 product descriptions have been checked for data quality by Talend. In addition, 10,000 updates to product entries were made in the first month after Talend went live, and Travis Perkins saw a 30 percent boost in website conversions, due in part to having consistent, accurate product descriptions.

Travis Perkins' success in standardizing and validating data quality helps dispel the general misperception that "it's hard to control data quality."

"We have increased sales conversions by at least 30% with Talend."

— David Todd,
Group Data Director

CONSTRUCTION: OPTIMIZING HUMAN CAPITAL MANAGEMENT

VINCI optimizes data from 157,000 employees with Talend

In any international group, communication with employees, collaboration, and identifying and sharing expertise are essential. But at VINCI, the global player in concessions and construction, managing employees turned out to be a matter of consolidating information contained in highly complex and diverse IT systems. With as many as 30% of internal emails failing to be sent, VINCI has become aware of the need to better manage and continually update data on its 157,000 employees.

VINCI selected Talend MDM to create a common language to be shared between all divisions as well as Talend Data Quality to correct these data and return them to all divisions and employees.

Since each division operates independently, it was important to make each division responsible for governing its own data. A support team was put in place to monitor the quality of the data and to identify errors in a monthly report. The group's HRMs regularly intervene to identify errors and alert employees. Today, the error rate is as low as 0.05%, compared to nearly 8% in the past and the employee information is always up-to-date.

VINCI 's successful centralized solution counters the misperception that "it's still hard to make all data operations work together."

"Talend provides us with a one-stop, instant overview of our employees."

— Jean Huot,
Director of Infrastructure and Services

MARKETING: DELIVERING CAMPAIGNS WITH DATA QUALITY

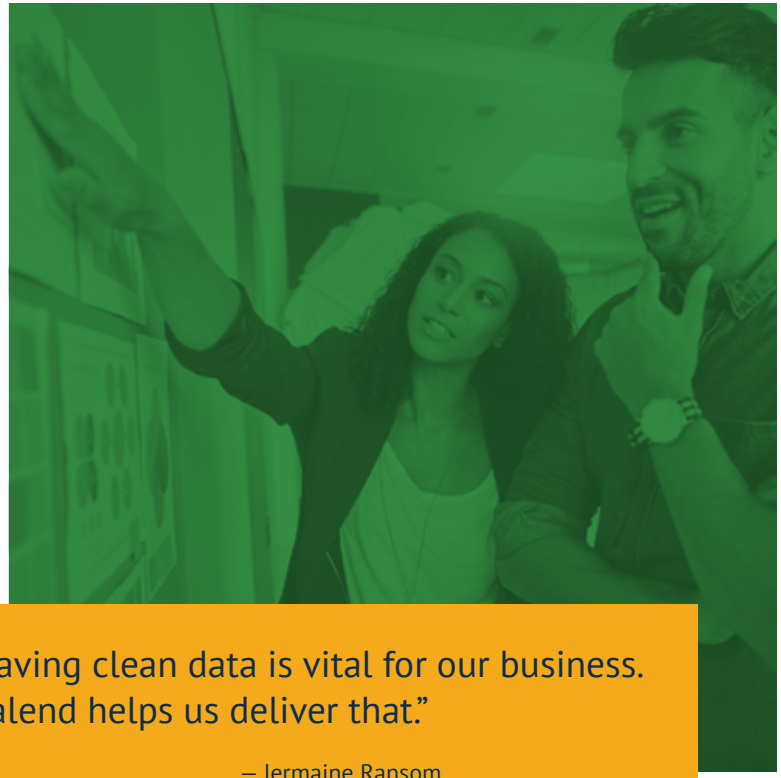
DMD Marketing delivers information for health providers with Talend

For DMD Marketing, a pioneer in healthcare digital communications and connectivity, data quality is a key differentiator. Because the principal service DMD provides—emails to health care professionals—is a commodity that can be supplied more cheaply by competitors, DMD needs to maintain its edge in data quality.

The company's client base needs to know they are targeting the proper healthcare professionals, so having clean data for names, addresses, titles and more is vital. DMD has deployed Talend Cloud Data Preparation and Data Quality to help deliver just that.

DMD also chose Talend for its data stewardship and self-service functionality. The company felt it was important to enable its internal users and its clients, to get in and see the email data and web tracking data on their own—without needing advanced technical skills. The company is also moving away from manual processes with manual data checks and is instead automating as much as possible, then providing human users access so they can augment and enhance the data.

The ROI for DMD Marketing includes raising the mail deliverability rate to a verified 95 percent, reducing email turnaround time from three days to one, and getting a 50 percent faster time to insight. DMD Marketing's success in empowering internal users and clients to monitor data quality proves it's not true that "data quality software is complicated and just for experts."



"Having clean data is vital for our business. Talend helps us deliver that."

— Jermaine Ransom,
Vice President of Data Services

MEDIA: EXPOSING DATA TO THE PUBLIC

Talend remains ICIJ's preferred data solution

In early 2016, The International Consortium of Investigative Journalists (ICIJ) published the Panama Papers –one of the biggest tax-related data leaks in recent history involving 2.6 Terabytes (TBs) of information. It exposed the widespread use of offshore tax havens and shell companies by thousands of wealthy individuals and political officials, including the British and Icelandic Prime Ministers. Now if that wasn't fascinating or mind-blowing enough, shortly after came the Paradise Papers –wherein

1.4 terabytes of documents were leaked to two reporters at the German newspaper Suddeutsche Zeitung.

To make public a database containing millions of documents, ICIJ raised its requirements for data quality and for documenting data integration procedures. Since millions of people would see the information, a mistake could be catastrophic for ICIJ in terms of reputation and lawsuits.

Talend Data Quality became ICIJ 's preferred solution when it came to cleaning, transforming, and integrating the data they received. It was key for ICIJ's data team to efficiently work remotely across two continents and have each step of the preparation process documented.

The Panama and Paradise Papers investigation has found an extraordinary global audience, which was unprecedented for ICIJ and their media partners. Within two months of Panama Papers publication, ICIJ's digital products received more than 70 million-page views from countries all around the world. In the six weeks after public disclosure of the Paradise Papers, Facebook users had viewed posts about the project a staggering 182 million times. On Twitter, people liked or retweeted content related to the Paradise Papers more than 1.5 million times.

These series of investigations in which the ICIJ and its partners used mass data to examine offshore-related matters — advances public knowledge to yet another level.

ICIJ's public database, with unimpeachable data quality, shows it's not true that "data quality is just for traditional data warehouses."

> "With Talend, we have been able to rapidly connect the dots between the corporate information for secret offshore companies and the people behind them.""
>
> — Mar Cabra, Head of the Data & Research Unit

CHAPTER 7

GUIDELINES FOR DATA QUALITY SUCCESS

talend

DO:

Build your interdisciplinary team

Recruit data architects, business people, data scientists, and data protection experts as a core data quality team. It should be managed by a deployment leader who should be both a team coach and a promoter of data quality projects.

Set your expectations from the start

Why data quality? Find your data quality answers among business people. Make sure you and your team know your finish line. Make sure you set goals with a high business impact.

Anticipate regulation changes and manage compliance

Use your data quality core team to confront short term compliance initiatives such as GDPR. You will then gain immediate short term value and strategic visibility.

Establish impactful and ambitious objectives

When establishing your data quality plan, don't hesitate to set bold business-driven objectives. Your plan will retain attention of the board and stretch people's capabilities.

Still deliver quick wins

Quick wins start by engaging the business in data management. Examples include onboarding data, migrating data faster to the cloud, or cleansing your Salesforce data.

Be realistic

Define and actively use measurable KPIs accepted and understood by everyone. data quality is tied to business so drive your projects using business driven indicators such as ROI or Cost-Saving Improvement Rate.

Celebrate success

When finishing a project with measurable results, make sure you take time to make it visible within key stakeholders. Know-how is good. It's better with good communication skills.

DON'T:

Try to boil the ocean

Do not pick a project that is too broad. Attack a piece at a time.

Over-train

More knowledge isn't always better. Your team must learn from their own team experience first.

Focus too heavily and risk tunnel vision

Take time to take a step back and keep a company-wide vision.

Leave the clock running

Set and meet deadlines as often as possible. It will bolster your credibility. As time is running fast and your organization may shift to short-term business priorities, track your route and stay focused on your end goals.

Forget to be company-wide

Your projects will need to empower everyone to curate data. Use cloud-enabled applications that can scale across the company with a single access and intuitive interface.

CHAPTER 7: **GUIDELINES FOR DATA QUALITY SUCCESS**

WHAT'S THE TAKEAWAY?

The cost of bad data quality can be counted in lost opportunities, bad decisions, and the time it takes to hunt down, cleanse, and correct bad errors. Collaborative data management, and the tools to correct errors at the point of origin, are the clear ways to ensure data quality for everyone who needs it.

Talend offers numerous tools to achieve both those goals. Contact us today to find out more about Talend's solutions, and try Talend Cloud FREE for 30 days to get access to all the tools described in this guide.

Contact us
https://www.talend.com/contact/

Customer Support
Visit the Talend Community

Learn More
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.

talend

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

WP242-EN