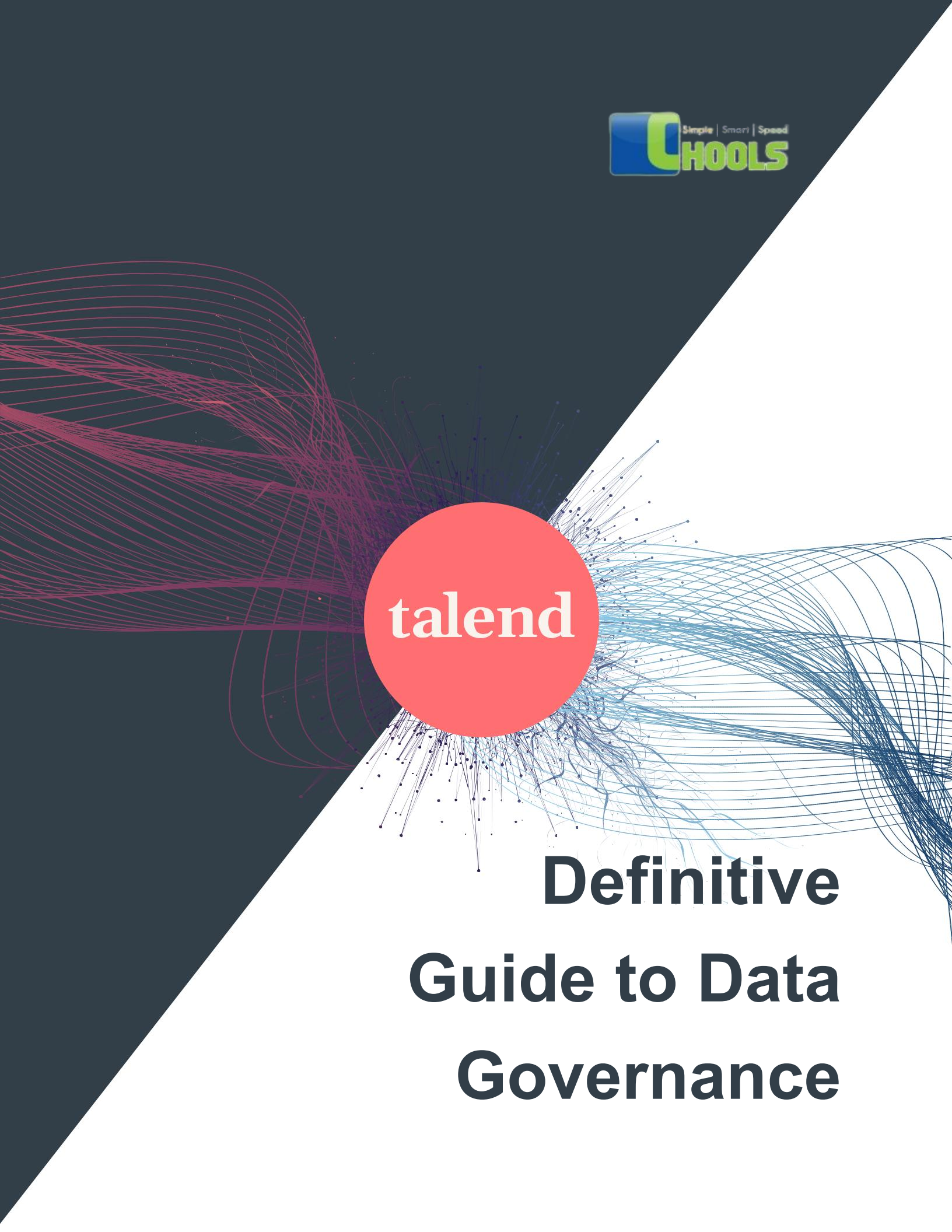


Definitive Guide to Data Governance

talend

Contents

Introduction

Why trusted data is the key to digital transformation

We've entered the era of the information economy, where data has become the most critical asset of every organization. Data-driven strategies are now a competitive imperative to succeed in every industry. To support business objectives such as revenue growth, profitability, and customer satisfaction, organizations are increasingly relying on data to make decisions. Data-driven decision-making is at the heart of your digital transformation initiatives.

But in order to provide the business with the data it needs to fuel digital transformation, organizations must solve two problems at the same time.

The data must be timely, because digital transformation is all about speed and accelerating time to market — whether that's providing real-time answers for business teams or delivering personalized customer experiences. However, most companies are behind the curve when it comes to delivering technology initiatives quickly.

But while speed is critical, it's not enough. For data to enable effective decision-making and deliver remarkable customer experiences, organizations need data they can trust. This is also a major challenge for organizations. Being able to trust your data is about remaining on the right side of regulation and customer confidence, and it's about having the right people using the right data to make the right decisions. And this too is a major challenge for organizations. According to the Harvard Business Review, on average, 47% of data records are created with critical errors that impact work.

Speed and trust are often at odds so it's common for organizations to focus on one or the other. Many organizations default to speed to meet the data users' expectations for ease and convenience and their own constraints. They allow developers to hand-code integrations or do one-off projects with niche integration tools in order to get results fast. While these tactics may solve for speed in the short term, they are not scalable as the company grows, and create quality and compliance risk due to the lack of oversight. On the other hand, organizations that try to solve the data trust problem often create a culture of "no" with the establishment of strict controls and an authoritative approach to governance. However, it's resource-intensive, cumbersome, restrictive and slow. This can hinder the innovation and agility so necessary to compete in today's business environment; businesses that operate too slowly risk being left behind.

"With our new data analytics platform, we now can better understand where the market is going, which helps us optimize energy trading while managing risk and complying with regulations."

René Greiner, Vice President for data integration, Uniper SE

According to Forrester, only 40% of CIOs are delivering results against the speed required.

Chapter 1:

What is data governance and why do you need it?

Why you should modernize your approach to data

Imagine that you are desperately looking for a rare book. The only way to get it is to visit a library, so you enter the single library found in your hometown. Entrance to the library is strictly controlled, so you have to show your ID to be granted an access card. Once you've entered, you have to weave through rows of books that are packed tightly together. You realize it will be painful to search for your book as nothing is tidy in this disordered environment. None of the books are classified by title nor author. However, you keep on searching. Since nothing is labeled, you have to look into each individual book to see if that's the right one. You could ask librarians to help, but they might be too busy to assist, because they're dealing with other incoming books in the library or other visitors waiting for their books.

After a while, you've finally found the precious book. However, when you open the book, you discover that some pages have been torn out, leaving the book hard to understand and with no value to you.

Please don't blame the librarians; they also need to deal with CDs and DVDs, new digital formats to classify, and a growing queue of visitors to manage (as well as online visitors clamoring for additional references).

You might think about ways to make things better organized so that people can find their books quicker. But nobody asked you for help — you were just here as a reader. Besides, the overall integrity of this library does not encourage you to trust it. The poor conditions, low-quality books, and your precious time wasted leave you with a negative perception of the library; it's certainly not a trustworthy institution you would recommend to others.

Does this sound like a discouraging and frustrating experience? Your data community may share the same feeling when looking for the right data sets in your organization.

Just like a library, we need to manage a growing volume of data assets, and not only the traditional data sources that we used to work with in the past but also the new ones that the digital era is creating, such as social media and sensor data.

This creates a data sprawl that almost impossible to scale. The more data you collect, the less you can meet the promise of self-service. Your data library becomes valuable for a happy few who have the broad skills required to explore the hidden value on their own. The others are left behind.

Also, with huge volumes of data coming from everywhere, you are losing control. You might not even know when some inappropriate or inaccurate content comes in, making untrustworthy data accessible to anyone. This is the "data swamp" situation that we see in many companies that can't keep up with the speed and volume of data entering their systems.

What if we could make all this data trustworthy, organize it at scale, and deliver it to everybody who needs it? What if we could give people the right tools to organize themselves and work as a team to cleanse, extract hidden value, and then assemble and deliver data everyone can trust? The ability to do this is the essence of data governance.

"Consolidating our data in a single system made us better placed to have clean data and also helped with data governance."

Senior It Manager, Enterprise Telecommunications Services Company

What is data governance?

Data governance is not only about control and data protection; it is also about enablement and crowdsourcing insights. Data governance is a requirement in today's fast-moving and highly competitive enterprise environment. Now that organizations have the opportunity to capture massive amounts of diverse internal and external data, they need the discipline to maximize that data's value, manage its risks, and reduce the cost of its management.

Data governance is a collection of processes, roles, policies, standards, and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals. It establishes the processes and responsibilities that provide the quality and security of the data used across a business. Data governance defines who can take what action upon what data, in which situations, and using what methods.

A well-crafted data governance strategy is fundamental for any organization.

A well-crafted data governance strategy is fundamental for any organization that works with data. It underpins how your business benefits from consistent, standard processes and responsibilities. Business drivers highlight what data needs to be carefully controlled in your data governance strategy and the benefits expected from this effort. This strategy becomes the basis of your data governance framework.

For example, if a business driver for your data governance strategy is to ensure the privacy of health care-related data, patient data will need to be securely managed as it flows through your business. Retention requirements (e.g., history of who changed what information and when) will need to be defined to ensure compliance with relevant government requirements, such as the GDPR and the CCPA.

Data governance ensures that roles related to data are clearly defined and that responsibility and accountability are agreed upon across the enterprise. A well-planned data governance framework covers strategic, tactical, and operational roles and responsibilities.

Data governance is not optional

An effective data governance strategy provides many crucial benefits to your organization that would be hard to live without.

An effective data governance strategy provides so many crucial benefits to your organization that it's hard to live without one.

These benefits include:

- **A common understanding of data:** Data governance offers a consistent view of, and common terminology for, data, while individual business units retain appropriate flexibility.
- **Improved quality of data:** Data governance creates a plan that ensures data accuracy, completeness, and consistency.
- **A data map:** Data governance provides an advanced ability to understand the location of all data related to critical entities, which is necessary for data integration. Like a GPS that can represent a physical landscape and help people find their way in unknown territory, data governance makes data assets usable and easier to connect with business outcomes.
- **A 360-degree view of each customer and other business entities:** Data governance establishes a framework so an organization can agree on "a single version of the truth" for critical business entities. The organization can then create an appropriate level of consistency across entities and business activities.
- **Consistent compliance:** Data governance provides a platform for meeting the demands of government regulations, such as the EU General Data Protection Regulation (GDPR), the CCPA (the California Consumer Protection Act), the US HIPAA (Health Insurance Portability and Accountability Act), and industry requirements, such as PCI DSS (Payment Card Industry Data Security Standards).
- **Improved data management:** Data governance brings a human dimension into a highly automated, data-driven world. It establishes codes of conduct and best practices in data management, making sure that the concerns and needs beyond traditional data and technology areas — including areas such as legal, security, and compliance — are addressed consistently.
- **Easy access:** A data governance framework ensures data is trusted, well-doc-umented, and easy to find within your organization, and that it is kept secure, compliant, and confidential.

GDPR, CCPA, and beyond: applying data governance for data privacy compliance

The European Union (EU) published the General Data Protection Regulation (GDPR) in May 2016. After a two-year transition period, the GDPR went into into effect on May 25, 2018. The GDPR applies to the processing of personal data of all data subjects, including customers, employees, and prospects. The definition of personal data includes "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person." Typical examples include customer names and contacts within CRM as well as employee salaries, benefits, and performance data,
but the regulation applies as well to newer data types such as sensor data that may point to vehicle location and driver behaviors.

The regulation applies to data subjects in the European Union, even when data is processed by organizations operating outside of the EU within jurisdictions like the United States, Asia Pacific, Middle East, and Africa. Noncompliance with the GDPR may result in huge fines, which can be the higher of €20M or 4% of the organization's worldwide revenues.

The GDPR isn't the only major data protection legislation on the books. The state of California has followed suit with the CCPA (California Consumer Privacy Act). It gives consumers the right to know what information companies are collecting about them, why they are collecting that data, and who they are sharing it with.

A robust data governance program is a pivotal part of the landscape for compliance with any data protection legislation. The traditional data governance disciplines of data ownership, metadata management, data quality management, and model governance also apply. Besides, GDPR compliance needs to incorporate self-service controls relating to data preparation and data stewardship to foster accountability for data protection across the stakeholders in a way that should be verifiable in practice, not just defined by legal guidelines written on paper. Although regulatory compliance is often what triggers a data privacy compliance project, it shouldn't be the only driver. Rather, the goal is to establish a system of trust with your customers regarding their personal data.

Use a modern data platform to make modern data governance a real success

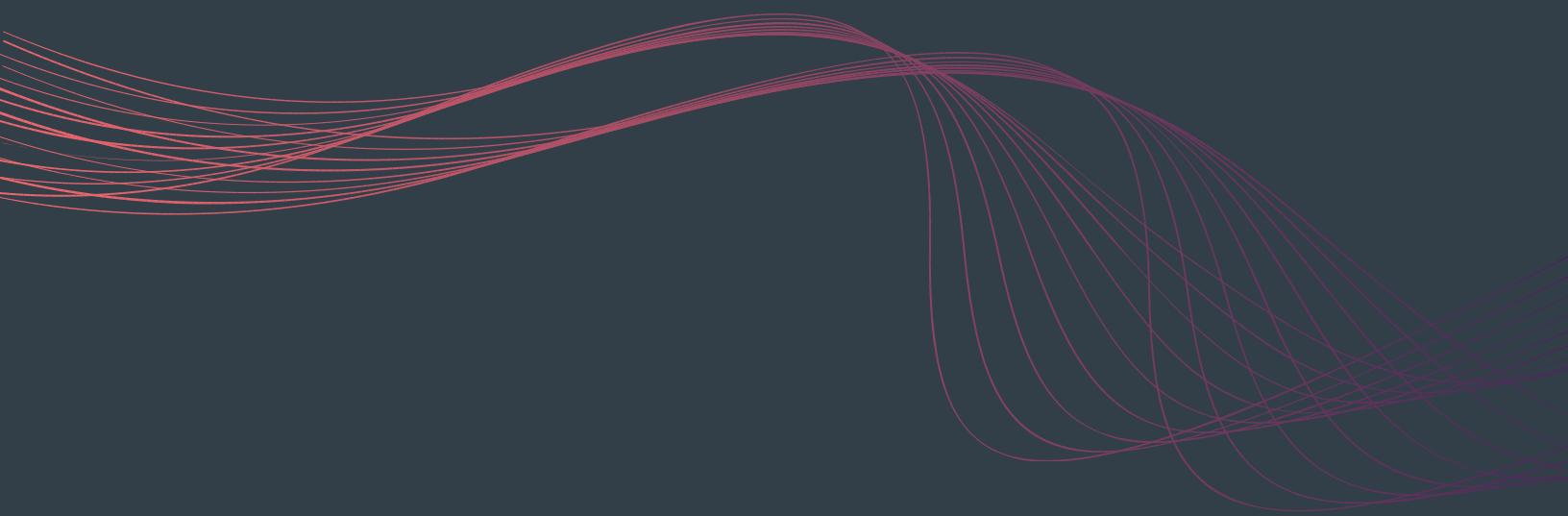
To find the right data governance approach for your organization, look for open source, scalable tools that are easy to integrate with the organization's existing environment. Additionally, a cloud-based platform allows you to quickly plug into robust capabilities that are cost-efficient and easy to use. Cloud-based solutions also avoid the overhead required for on-premises servers. As you start comparing and selecting data governance tools, focus on choosing ones that help you realize the business benefits laid out in your data governance strategy.

These tools should help you:

- Capture and understand your data through discovery, profiling, and benchmarking tools and capabilities. For example, the right tools can automatically detect a piece of personal data, like a Social Security number, in a new data set and trigger an alert.
- Improve the quality of your data with validation, data cleansing, and data enrichment.
- Manage your data with metadata-driven ETL and ELT and data integration applications so that data pipelines can be tracked and traced with end-to-end data lineage.
- Control your data with tools that actively review and monitor. Document your data so that it can be augmented by metadata to increase its relevance, searchability, accessibility, linkability, and compliance.
- Empower the people who know the data best to contribute to the data stewardship tasks with self-service tools.

Chapter 2:

Choosing the best governance model for you



Finding the right balance between top-down and bottom-up governance

Being data-driven is a business imperative, but there are tough challenges to overcome to get the most value out of data. The volume of data that companies need to manage doubles every two years, creating data sprawl. And at the same time, there is a wider variety of data to process and analyze, such as new streaming data coming from the IoT, sensors, weblogs, click streams, crowdsourced data from digital applications and social networks, and so on.

Moreover, there is a multiplication of new data-driven roles within your organization. Back in the early 2000s, we had IT developers, business analysts, and business users. Then came new data professionals, sometimes working in a central organization such as IT or an analytics Center of Excellence, and sometimes reporting to the lines of business. Some examples of those new roles are data stewards, data scientists, data curators, data protection officers, and data engineers. Today, even people in nontechnical roles have become data-savvy, aspiring to become more than passive data consumers, and desiring to autonomously turn data into insights autonomously.

Here is the quandary. Data is coming from everywhere — from traditional sources that are already under the control of central IT to new data sources that comes from everywhere: shadow IT, third-party data, internet of things, applications, etc. Plus, that data is needed faster than ever, now that companies need to ingest and analyze real-time data, instead of reacting to day- or week-old data. In fact, we are now seeing a 35% CAGR for streaming analytics. And with so many people in so many parts of the business with a variety of data analysis skills wanting access to data for business intelligence, IT is expected to provision access to it all. However, IT's budget and resources are relatively flat. There is a growing gap between business expectations and what IT can deliver.

As a result, the economics of data integration is broken. Traditionally, central organizations have established what IDC calls "governance with the no," which means that business users had to come to them with request that they had the power to fulfill or reject. This has created a gap between business and IT when it comes to data ownership, a gap that is only widening with the realities of data sprawl.

The economics of data integration are broken. There is a gap between business expectations and what IT can deliver.

The traditional approach to data governance can’t scale to the digital era: too few people access too little data

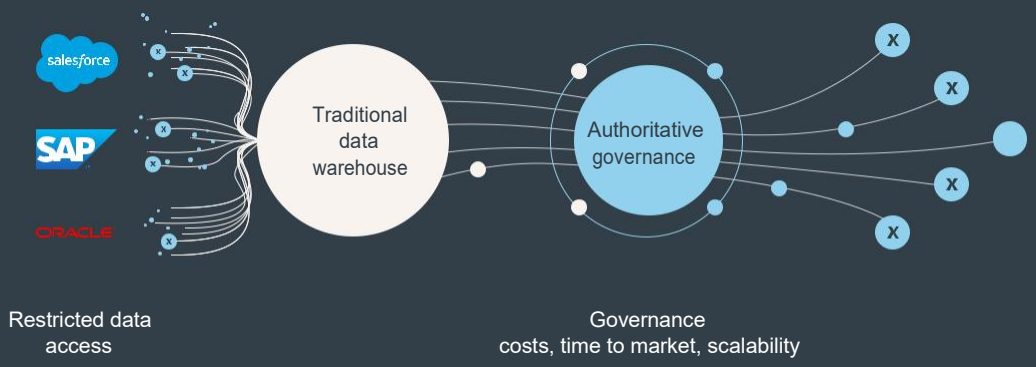
In the past, we established highly centralized approaches for creating data hubs. Some examples are Master Data Management, CRM, creating a customer 360° view, or enterprise data warehouses.

A highly centralized approach relies on a small team of highly experienced data professionals armed with well-defined methodologies and well-known best practices. When applied to an enterprise data warehouse, for example, you could start by defining a central data model where you can collect and reconcile data that has been marked as relevant for insights. Then the data is remodeled into data marts so that it can fit a business domain or problem. Finally, this data is remodeled again using a business intelligence tool that provides a semantic layer such as a “data catalog,” destined to be packaged into predefined reports. Only then can the data be consumed for analytics.

To understand the scalability problem that this model faces in the digital era, let’s compare it to a domain that has already faced a similar data sprawl with the digital era — web content. Before we entered the 21st century, we sought knowledge in an encyclopedia such as the Encyclopedia Britannica or Microsoft Encarta. The model created a clear distinction between the data providers and data consumers; only a handful of experts could author the encyclopedia. Everyone else was a data consumer.

Of course, the quality can be excellent with this model. The Encyclopedia Britannica is written by about 100 full-time editors, together with around 4,000 highly skilled contributors, including Nobel Prize winners and former heads of state.

But the issue that this model faced when the web became mainstream is that these encyclopedias couldn’t cope with the demand from the data consumer. Now people want comprehensive, up-to-date articles on each and every topic imaginable, in a single click, in their native languages.



Your organization is facing the same issue with your data. You might have the best experts in your central organization, but you do not have enough resources to bring all this data accurately to everyone who needs it as quickly as they want it, nor can you address the growing needs of the business users for new and different types of data.

Ultimately, people will find other ways, such as shadow IT or creating other bodies, to meet their data needs. The IT teams who can't evolve from this centralized model will rapidly lose control, jeopardizing speed, accuracy, and security.

Data access is tightly controlled. The encyclopedia model fails to scale in this Big Data era, when multiple people demand immediate access to the right data from every source.

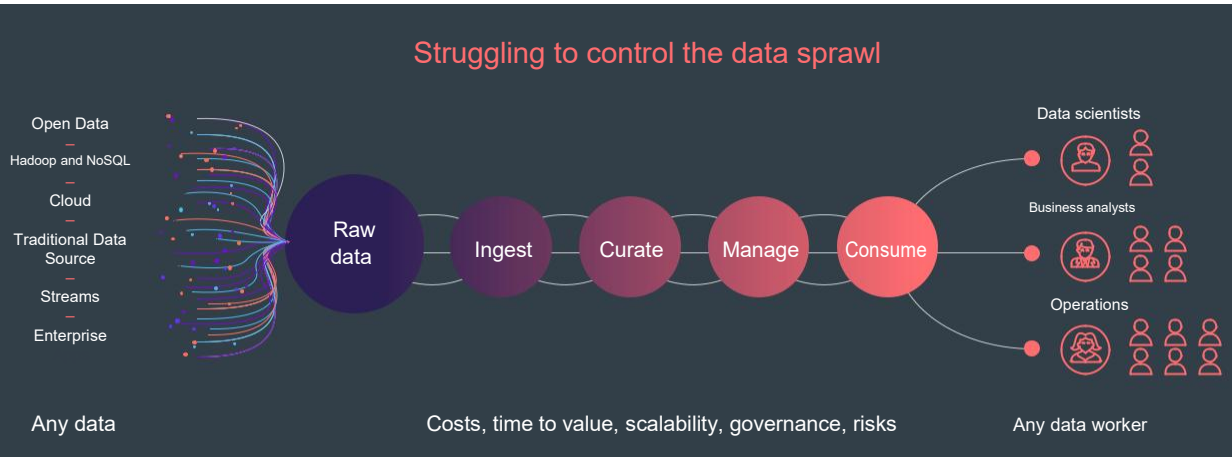
The data Wild West: struggling to control the data sprawl

With the advent of Big Data, we saw the rise of a much more agile approach to data management — the data lake. While the previous approach was to start with data modeling and data governance first and then dig in to the actual data through a top-down approach, data lakes take the exact opposite strategy. It all starts with raw data. Raw data can be ingested with minimal upfront implementation costs, generally on basic and low-cost file systems. You don't have to bother with the file structure when you bring data in. In fact you might not know what exactly is in it. Later in the process, you might create a structure on top of this data (specialists call this step "schema on read"), but also data quality controls, security rules, policies, controls, etc.

This more agile model has multiple advantages over the centralized model. It scales across data sources and use cases. And it scales across audiences, although only the most data-savvy people can access raw data, while others require structured data that is connected to their business context before they can take advantage of it.

That's why data lakes 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 turn data into smart data.

But this doesn't stop there. The next step is to share this data with a broader audience, and this audience need more structure. Many organizations 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 and data quality control.



When this is successful, then, the next step is to deliver information and insights into the whole organization. For example, you might want to provide product recommendations directly to your customers by embedding a machine learning algorithm into your front office applications, or you might want to monetize some of your data to third parties. Again, the prerequisite is to develop 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.

However, through this approach, we created a big challenge: We didn't consider data governance alongside this way of doing things, but rather as an afterthought when we had to expand to new audiences and use cases.

We didn't consider data governance alongside this more agile model, but rather as an afterthought.

Facebook faces this problem. Initially, it just provided a platform without acting as a content provider. This allowed the company to create self-governed communities on a platform with no limits on the content it could ingest and the number of users and communities it could serve.

But then came fake news, violent content, and malicious use of the platform. Because anyone can enter any data without control, control has now become very difficult to establish. New regulations have emerged for data privacy, and Facebook is under scrutiny not only by governments and regulators, but also by its own users. Some of its biggest advertisers, such as Unilever, have already publicly threatened to boycott Facebook and Google if the tech giants fail to efficiently police extremist and illegal content. Due to those market pressures, not only have trust and safety has become the most strategic talking point around Facebook over the last months, but the related costs are rising. Facebook is hiring 10,000 more employees in its trust and safety unit and plans to double those headcounts in the future.

The lesson learned is that there's no chance to succeed in the digital era without a framework; data governance can make or break your digital transformation, and when you realize it, it might be too late. Forget the governance dimension in your data lake, and it will soon become a data swamp. And once that point is reached, the effort needed to turn your lake into something that your business can safely leverage might be huge.

Delivering data at the speed of the business is the Holy Grail, but there is no compromise with data governance.

Establishing collaborative governance in the digital era

Balancing top-down and bottom-up governance

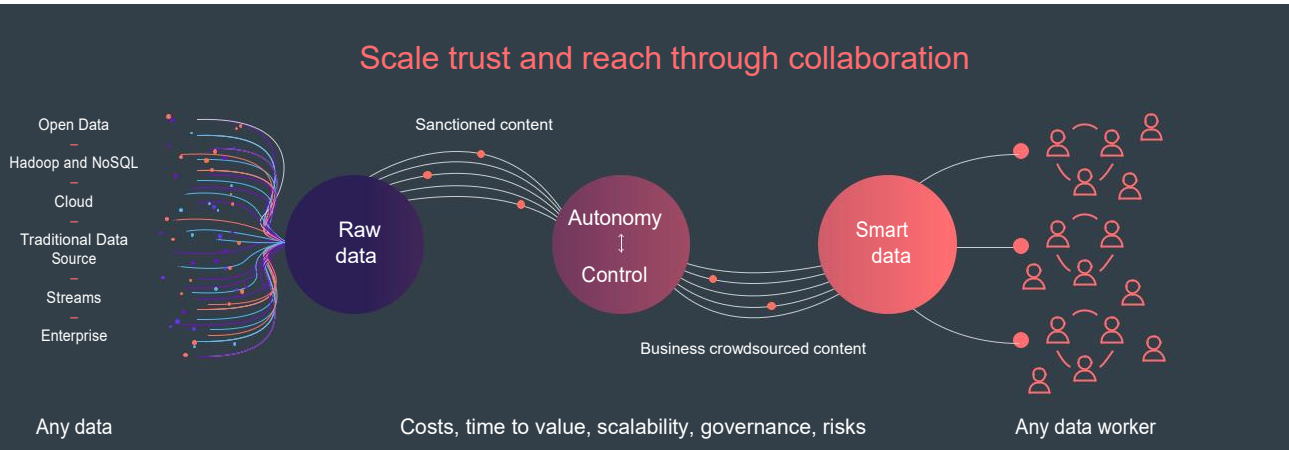
What is missed in the second model is the ability to take control of the data as it enters your systems, rather than after the fact. But at the same time, we need to recognize that there are more and more incoming data sources, introduced by more and more people from different parts of the organization. It's helpful to establish a more collaborative approach to governance up front so that the most knowledgeable among your business users can become content providers and curators. Working with data as a team from the start is essential to this approach. Otherwise, you may become overwhelmed by the amount of work needed to validate that your data is trustworthy.

Again, let's turn to the lessons learned from the web era. Wikipedia has established itself among the top 5 most visited sites globally. Wikipedia maintains more than 5 million articles, but there are only 1,194 administrators to maintain the pages. However, anyone can contribute, and regular authors number 130,000. To accommodate this, Wikipedia has well-defined principles for collaborative content curation that enhance their ability to scale and deliver a fair level of trust in the content.

By introducing a Wikipedia-like approach where anyone can potentially collaborate in data curation as long as standards are followed, organizations can engage the entire business in contributing to the process of turning raw data into something that is trusted, documented, and ready to be shared. Businesses can implement a system of trust that scales by leveraging smart and workflow-driven self-service tools with embedded data quality controls.

Note that this model can complement, rather than replace, the top-down approach; some heavily regulated processes, such as risk data aggregation in financial services, and some specific data, like consumer credit card information, require very specific attention, and in that case the bottom-up approach does not apply. Defining which data governance model applies is a typical responsibility for the data governance team.

In the next chapter, we'll see how to implement the collaborative and governed model using a three-step approach.



Chapter 3:

Three steps to deliver data you can trust

Implementing collaborative governance

Although technology can help fix the issue, enterprises need to set up the discipline to organize their data at scale. As we saw in the previous chapter, traditional data governance must be reinvented with this data sprawl: According to Gartner, “through 2022, only 20% of organizations investing in information will succeed in scaling governance for digital business.” Given the sheer number of companies that are awash in data, that percentage is just too small.

Modern data governance is not only concerned with minimizing data risks and creating a “data police,” but it is also about maximizing data usage, which is why traditional authoritative data governance approaches are not enough.

There is a need — and an opportunity — for a more agile, bottom-up approach. That strategy starts with the raw data, links it to its business context so that it becomes meaningful, takes control of its data quality and security, and thoroughly organizes it for massive consumption. Now is the time to seize this opportunity, as more and more people clearly realize the value of data and therefore understand the benefits of managing it properly. In addition, due to data scandals and data leaks that make the headlines and with the proliferation of new regulations that put higher stakes on data protection, people now also understand the challenges as well.

New data platforms empower this new discipline, and leverage smart technologies like pattern recognition, data cataloging, data lineage, and machine learning to organize data at scale and turn data governance into a team sport by enabling organization-wide collaboration on data ownership, curation, remediation, and reuse.

In this chapter, we present a three-step methodology to better deliver trusted data within your organization. We also illustrate with examples how data integation platforms can help with appropriate tools and capabilities to succeed in implementing this approach.

Top takeaway:

Make sure you can Discover and Cleanse your data as soon as it enters your data landscape. Then Organize and Empower, thereby creating a single point of trust where you can define the data accountability and empower as a team the people needed to document them, protect them, and share them widely. Once you are in control, you can Automate and Enable, so that you can deliver insights at scale to a wide range of data and application consumers.

Step 1: Discover and cleanse your data

With increased affordability and accessibility of data storage over recent years, large data repositories such as data lakes have developed, leaving teams with a growing number of various known and unknown datasets. Hopefully these data sets enrich your data lake, but unfortunately sometimes they pollute it. As a consequence, everyone in a given company may face a data backlog. Although it might take seconds to ingest the data with modern software, it might take weeks for IT teams to publish new data sources in data warehouses or data lakes.

At the same time, data consumers might not be aware that the data they need is available for use. It takes hours for business analysts or data scientists to find, understand, and place all data into context. IDC found that only 19% of data professionals' time is spent analyzing information and delivering valuable business outcomes. Instead, they spend 37% of their time preparing data, and 24% of their time protecting data. That translates into a lot of wasted time, which IDC estimates at 24 hours per week. And that also brings governance headaches, because when people in the business don't find the data they're looking for, they recreate it. Then they add their own rules on top of their newly created data sources, ultimately propagating multiple versions of "the truth".

Your challenge is to overcome these obstacles by bringing clarity, transparency, and accessibility to your data assets. Wherever this data resides — within enterprise apps like Salesforce.com, Microsoft Dynamics, or SAP; a traditional data warehouse; or in a cloud data lake — you need to establish proper data screening so you can make sure you have the entire view of data sources and data streams coming into and out of your organization.

Top takeaway:

Don't go blind with your data. As you're making your very first steps into your data strategy, you need to know what's inside your data. To do so, you need tools and methodology to step up your data-driven strategy.

Know and understand your data in the first place

When working on your data, it's critical that you start exploring the different data sources you wish to manage. You should identify patterns, explore data sets, and look at data structure from the different sources at your disposal.

In the past, this activity was processed manually by data experts using traditional data profiling tools. But this approach doesn't work anymore, since it requires working with each dataset individually. The digital era's data sprawl requires a more automatic and systematic approach. This is what modern data cataloging tools such as Talend Data Catalog can do. It helps you to schedule the data discovery processes that than crawl your data lake or other data landscapes and intelligently examine the underlying data, so than you can understand, document, and take actions based on the content of your datasets.

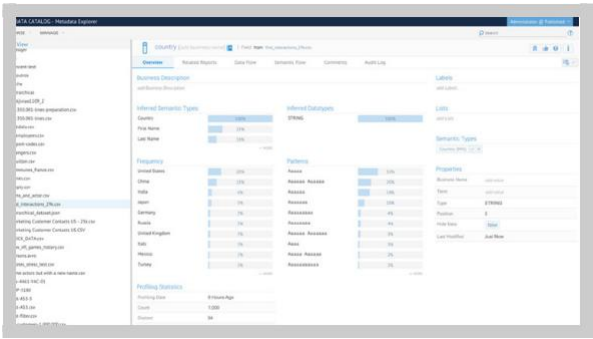
Autoprofiling for all with Data Catalog

The autoprofiling capabilities of Talend Data Catalog facilitate data screening for nontechnical people within your organization. Data Catalog provides you with automated discovery and intelligent documentation of the datasets in your data lake. It comes with easy-to-use profiling and sampling capabilities that help you to assess data at a glance. With trusted and autoprofiled datasets, you can have powerful and visual profiling indicators, so users can easily find and the right data in a few clicks.

Talend Data Catalog can automatically draw the links between datasets and connect them to a business glossary. In a nutshell, this allows an organization to automate the data inventory and leverage smart semantics for autoprofiling and relationships discovery and classification thanks to an integrated semantic flow.

The benefits are twofold. The data providers and data owners get an overview of their data and can take actions. For example, the critical data elements that need specific attention, such as personal data that requires specific consideration for privacy compliance, are automatically categorized and highlighted. The data catalog also exposes the potential data quality issues that require remediation.

The data consumers can see what's in the data before they consume it, by seeing data samples or getting the indications that the data within a column might be for example a phone number, an account number, or an email.



» *Figure 1: Data profiling with Talend Data Catalog*

Go further with data profiling

Data profiling enables you to explore your datasets and accurately assess your multiple data sources based on six dimensions of data quality. It helps you to identify if and how data is inaccurate, inconsistent, and incomplete.

Think about a doctor's exam to assess patient's health. Nobody wants to have surgery without a precise and close examination beforehand. The same applies to data profiling. You need to understand your data before you can fix it. Since data often comes into in your hands in hidden formats, inoperable, or unstructured, an accurate diagnosis helps you to have a detailed overview of the problem before fixing it. It saves time for you, your team, and your entire organization.

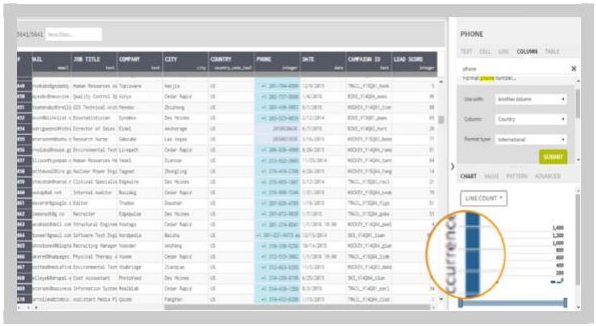
To the same extent that a family physician and specialist physician have different yet crucial roles for a health diagnostic, and perform it with slightly different approaches and tools, data profiling techniques apply for different roles and require different tools.

80% of surveyed organizations say they have been affected by GDPR or other data protection legislation.

Empowering power users with self-service profiling

In many cases, the people who know the data best are not the data experts. Think about customer contact data, for example: sales admins, sales representatives, and field marketing managers probably know the data quality issues better than the central IT team. And they most keenly feel the pain of data quality issues, as it impacts how efficiently they can go about their day-to-day jobs. For use cases such as Salesforce data cleansing, you may wish to gauge your data quality by delegating some primary profiling activities to those business users.

Of course, you won't ask them to become data quality experts. This requires new kind of smart tools that can hide the technical complexity and provide a simple, fast, and visual user experience to guide them to do rapid profiling on their favorite datasets. With tools like Talend Data Preparation, you can have powerful yet simple built-in profiling capabilities to explore data sets and assess their quality with the help of indicators, trends, and patterns.



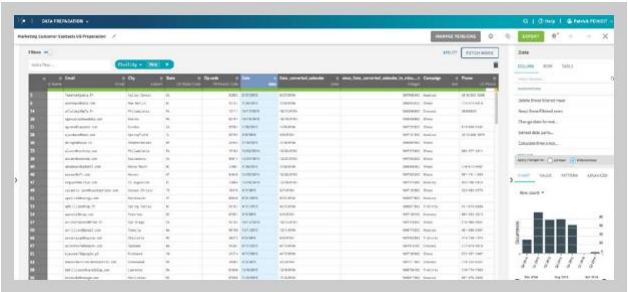
» *Figure 2: Data profiling for power users with Talend Data Preparation*

Establish trust in your data with advanced profiling

While automatic data profiling through both a data catalog and self-service profiling addresses the case for bottom-up data governance, a top-down approach might require a deeper look into the data. Think, for example, of risk data aggregation and risk data reporting. These are defined by formal principles and related regulations. Among the principles are that "Supervisors expect banks to measure and monitor the accuracy of data and to develop appropriate escalation channels and action plans to be in place to rectify poor data quality."

This kind of formal approach for measuring and monitoring accuracy of data and taking actions when it doesn't meet the standards requires the involvement of data engineers, data quality specialists, or IT developers. Using tools such as Talend Data Quality in Talend Studio, they would start by connecting to data sources to analyze their structure (catalogs, schemas, and tables), and store the description of their metadata in its metadata repository. Then they would define available data quality analyses including database, content analysis, column analysis, table analysis, redundancy analysis, and correlation analysis. These analyses carry out data profiling processes that define the content, structure, and quality of highly complex data structures.

All this deep discovery draws a "trust index" that's be calculated, reported, and tracked on a regular and automated basis. It will also trigger alerts when the trust index gets below a certain threshold.



» *Figure 3: Advanced data profiling for data engineers*

Top takeaway:

Keep in mind that your data strategy should first and foremost start with data discovery. Not profiling your data puts your entire data strategy at risk. You need to analyze the ground to make sure you build your "data house" on solid foundations.

Think, for example, of a data privacy initiative that cannot capture the fact that a new dataset with personal data has entered in your data landscape. Your whole compliance program might be compromised by new data elements that you were not able to identify.

Data integrity from the get-go

Incorporating appropriate data quality controls in your data chain is vital for the success of your data governance initiative.

Suppose, for example, that you want to start a campaign to contact customers for billing and payment and your primary source to contact appropriate people is email and postal addresses. Having consistent and correct address data is vital to be able to reach everyone. Otherwise, you may lose lots of revenue or miss out on opportunities due to missing or inconsistent data.

Data integrity issues have exploded over the last several years. The sources and volume of data is growing, and so are the number of data professionals who want to work with it. The impact of this proliferation of data across a growing number of clouds and digital channels and an increasing number and variety of people put the enterprise at risk for data leaks, data breaches, and misguided insights based on rogue and inconsistent data. As an example, 62% of end users admit they have access to data they should not. Dealing with integrity is crucial as new data governance regulations that may have a tangible impact on business are implemented, which may have a tangible 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 revenue.

There is a way to solve this problem. Organizations need to ensure data accuracy and availability 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. Everyone from IT to data scientists to application integrators to business analysts should be able to participate and extract valuable insights out of always available, good quality data.

62% of data end users admit they have access to data that they should not.

Orchestrating data integrity across pipelines

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.

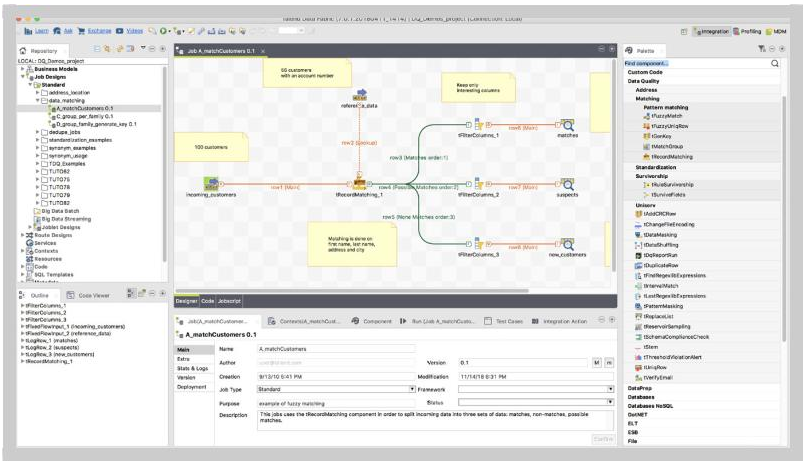
But data quality is not a stand-alone operation. To make it successful and deliver trusted data, you need to operate data quality operations upfront and natively from the data sources, along with the data lifecycle to ensure that any data operator or user or app could consume trusted data at the end.

Data quality is key to us — as well as understanding provenance for data governance purposes.

Executive, health insurance company

So before establishing your single point of trust and putting data at your disposal, you must be able to apply data quality controls and remediations to the ingested data sources. To do so, Talend Data Quality generates native code to run data quality controls at the right place, on premises, inside a Big Data cluster, or in the cloud, and at the right time, on data at rest or on streaming data.

You need to profile, cleanse, and standardize your data while monitoring data quality over time, in any format or size. This is why you need more than isolated, point solutions for data quality but rather a pervasive platform that provides a wide array of data quality controls not only for cleansing and standardizing data sources, but also for delegating some tasks to business experts using integrated self-service tools.

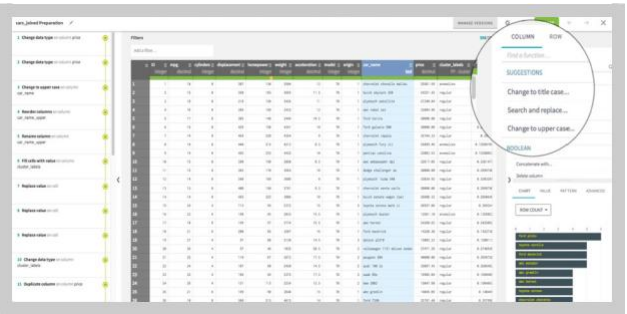


» *Figure 4: Talend Data Preparation orchestrates data with integrity across pipelines*

Delegate data cleansing with self-service

As we have learned, data isn't the responsibility of a single central organization anymore. Successful data governance frameworks require setting accountabilities and delegating authority appropriately. For example, a data protection officer in a central organization might want to delegate tasks to data stewards or business users in the operations; a sales engineer might be best positioned to ensure that contact data for their accounts are accurate and kept up-to-date. A campaign manager should ensure that a consent mechanism has been put in place and captured within a marketing database.

To support this kind of delegation, organizations need to provide workflow-based, self-served apps, such as Talend Data Preparation, to different departments. This provides them with enhanced autonomy without putting the data at risk.



» *Figure 5: Self-service data cleansing with Talend Data Preparation*

Top takeaway:

Data leaders cannot master the opportunities and challenges of digital transformation with yesterday's centralized roles and organizations, as this creates data bottlenecks and misalignment. Data is a team sport that spans across functions and lines of business.

Extend your data quality with the cloud

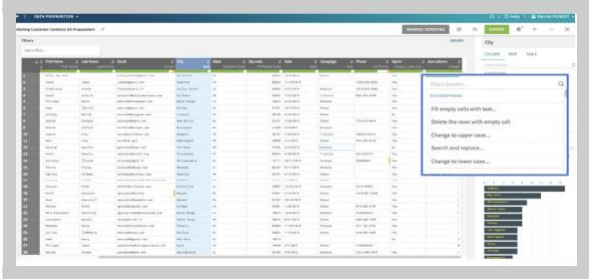
The cloud drastically extends the boundaries of data. Lines of business use their own applications, and products, people, and assets create their own data pipelines through the web and the Internet of Things. Data can also be exchanged seamlessly between business partners and data providers.

Self-service is the way to get data quality standards to scale. When trusted data is not provided in a self-service way, multiple surveys have shown that business analysts and data scientists spend 80% of their time cleaning data and getting it ready to use. Reduced time and effort mean reduced costs; as a result, more value and more insight can be extracted from data.

Self-service apps such as Talend Data Preparation deal with this problem, allowing potentially anyone to access a data set and then cleanse, standardize, transform, or enrich the data. Because it is easy to use, Talend Data Prep 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.

Talend Data Preparation unlock people's productivity when working with their data, but it captures the action they take on the data. When those actions help to bring trust in the data, they can be operationalized and embedded with data pipelines so that anyone can benefit. In addition to improving personal productivity, the true value of those self-service applications is to drive collaboration between business and IT.

At the end of the first of the three-step approach to deliver data you can trust, data sources have been identified and documented. Actions have been taken for the data sources that required attention with respect to their data quality.



» *Figure 6: Self-service access with Talend Data Preparation*

Top takeaway:

Your data governance platform's choice should take into consideration its ability to delegate data quality operations to business users in a self-service mode while keeping control. It is critical if you want to scale rapidly and mutualize your data cleansing efforts at the business speed. It would be risky not to do anything and let people prepare and cleanse data on their own, spending a considerable amount of time in repetitive tasks on uncontrolled data sources.

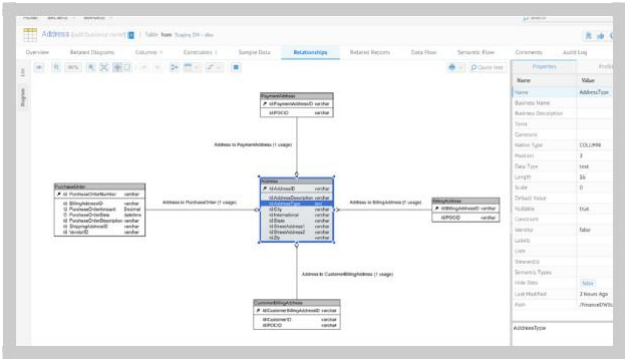
Step 2: Organize data you can trust and empower people

Organize data with integrity by establishing a single point of trust

While step 1 helped to ensure that the incoming data assets are identified, documented, and trusted, now it is time to organize them for massive consumption by an extended network of data users who would use it within the organization.

This starts by establishing a single point of trust; that is to say, collecting all the data sets together in a single control point that will be the cornerstone of your data governance framework. Then, you need to select the identified datasets and assign roles and responsibilities directly into your single point of control to directly operationalize your governance from the get-go.

It is one of the advantages of data cataloging: regrouping all the trusted data in one place and giving access to members so that everybody can immediately use it, protect it, curate it, and allow a wide range of people and apps to take advantage of it. The benefit of centralizing trusted data into a shareable environment is that it will save time and resources of your organization once operationalized.



» *Figure 7: Establishing a single point of trust with Talend Data Catalog*

Top takeaway:

According to the Gartner Magic Quadrant for Business Intelligence and Analytics Platforms, 2017: "By 2020, organizations that offer users access to a curated catalog of internal and external data will realize twice the business value from analytics investments than those that do not."

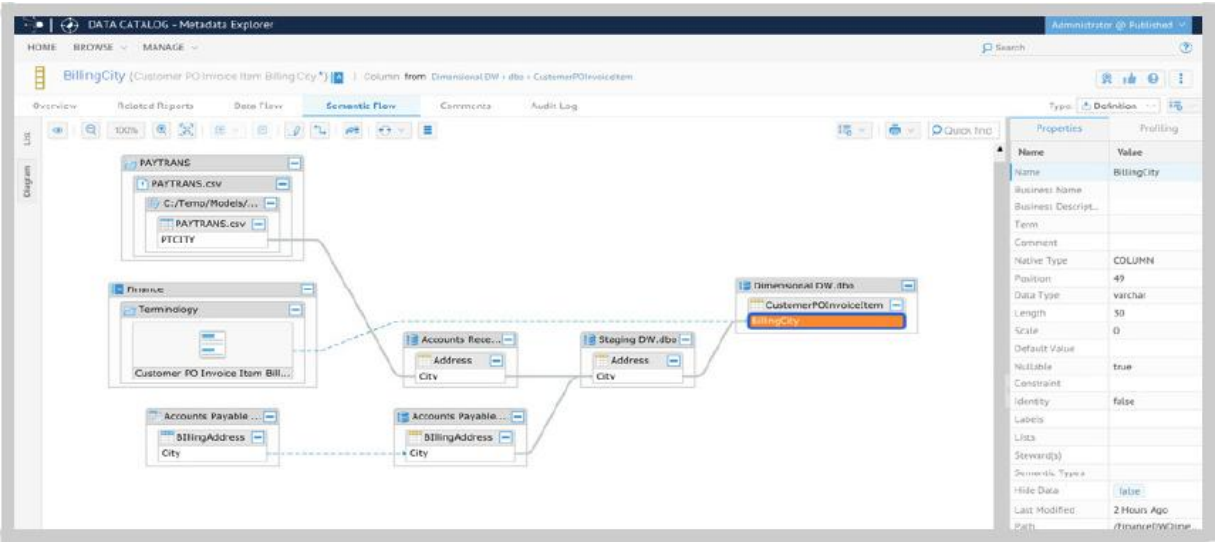
Define your data in a business glossary and make it accessible through data lineage

Within a data catalog, a business glossary is used to define collections of terms and to link them to categories and subcategories. Building a business glossary can be as simple as dragging in an existing well-documented data model, importing the terms and definitions from other sources (e.g., CSV, Microsoft Excel), or interactive authoring via the user interface during the process of classifying objects. Once published, the glossary can be accessed by potentially anyone who has proper authorizations through a search-based interface (see figure below).

This dataset becomes a living resource, as you enable authorized people to edit, validate, and enrich the data. Doing it automatically through a data catalog saves lots of time and resources.

Even more importantly, data lineage gives you the big picture to track and trace your data flows from sources to final destination.

Imagine that you find some inconsistent data in your data systems that have been created and perpetuated in one of your datasets and you're asked to explain it, identify it, and correct it. Having data lineage accelerates your speed to resolution by helping you to spot the right problem at the right place and ensure that your data is always accurate. Moreover, if new datasets come to your data lake, establish a data lineage helps you to identify these new sources very quickly.



» *Figure 8: Build a business glossary with Talend Data Catalog*

Data lineage helps you quickly answer to audit trails as requested by the competent authorities: concrete examples are privacy regulations such as GDPR or CCPA, or data risk regulations for financial services such as BCBS 239, that encourage the creation of an always accurate data inventory that can track the data provenance and related data processing activities that are applied to specific data elements.

The typical questions Typical questions that might arise include: When an error is identified in a management report, where did it come from? When did the error occur? Who is accountable for that? How can you solve it? All of these questions could find an answer into a metadata management solution that integrates the data lineage. Lineage gives you a picture overview of the data views so you can easily spot the problem.

Lineage is also important for change management. For example, suppose that a data element in a data chain has to be updated. What would be the impact upstream in the data chain? What other data elements are impacted? Answering those questions can take weeks, while proper data cataloging and data lineage principles and tools provide answers in a matter of clicks.

"Data intelligence software is critical in the support of data enablement and governance programs because it provides organizations with the knowledge needed to deliver the right data to the right resource at the best time. Data lineage is a core element in those data intelligence solutions, bringing more insight about the data itself and delivering even more impact and value for data-driven organizations."

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

Identify roles and responsibilities and establish ownership

Once you have defined your data categories, sometimes also referred to as “critical data elements,” you have a more accurate picture of the sources in your data environment.

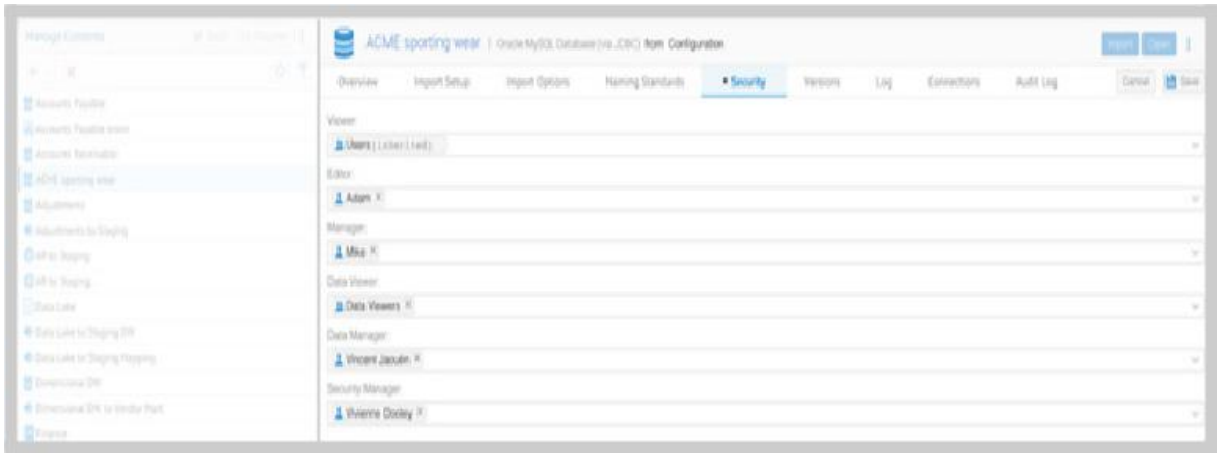
You can also to define better data owners: Who is responsible for this particular data domain? Who is responsible for viewing, accessing, editing, and curating the data sets?

At this step, using a RACI model will help you save time defining and assigning roles and responsibilities between your stakeholders.

RACI is an acronym derived from the four key responsibilities most typically used: Responsible, Accountable, Consulted, and Informed. The RACI model is an assignment matrix that’s easy to understand and use. It’s easy to understand and easy to use. It’s particularly useful if your data governance involves different departments and divisions in your organization.

The next step is to define data owners who are ultimately accountable for one or more data categories and subcategories. These data owners areresponsible for day-to-day operations regarding the data or appoint data stewards for those tasks. They identify critical datasets and critical data elements (CDE) as well as establish standards for data collection, data use, and data masking. Talend Data Catalog may also support catalog owners and stewards for data categories and sub-categories and assign their related roles and workflows.

For example, Talend Data Catalog may catalog the data owners for “customer” as well as “customer identity,” “customer billing,” “customer contact,” and “customer ship-to information.”



» *Figure 9: Identifying roles and responsibilities with Talend Data Catalog*

Empower people for data curation and remediation

Your data strategy success can be measured by the number of people who connect to your sources, but also by their level of trust with the data. You need to offer role-based easy-to-use applications that help you to engage people and make them accountable for data accuracy and value.

"Cleansing and consolidating consumer data enables us to deliver the kind of personalized experience today's consumers deserve and expect."

Steve Brennan, Vice President, Strategy and Analytics, Carhartt, Inc.

Empower people to curate data

According to Wikipedia, data curation "is the organization and integration of data collected from various sources. It includes annotation, publication, and presentation of data to make sure it's valid over time." You can enable data curation by putting in place an explicit RACI model where you define who can define, edit, validate, and enrich data in your systems.

Now it's time to encourage people to actually do it. Communication is vital; you may have decided with your executive sponsor to communicate about your project launch officially. When doing so, involve your internal communication department so that you explain the broader purpose of your project.

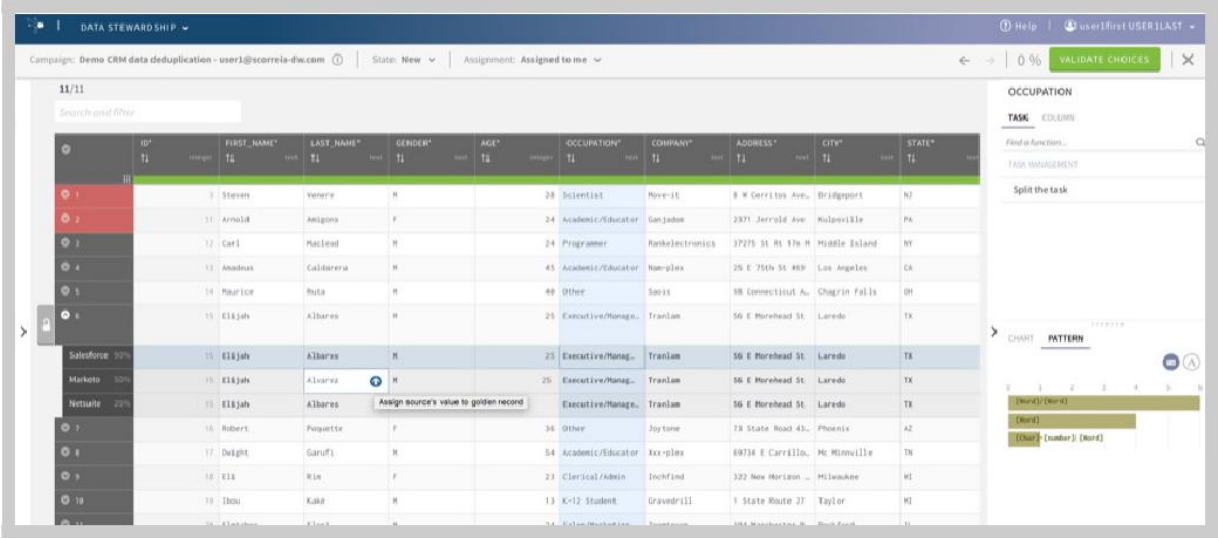
Why is empowering people so important? A data governance project is not just intended to let trusted data be accessible to all. It's also about promoting data custodians' acountability to the rest of the organization so that they can enrich and curate trusted data and produce valuable, accurate insights out of the data pipelines.

Use email and internal collaboration systems. Choose an impactful name or visual that can depict the purpose of your data program. Along with your communication plan, make sure you touch different audiences and different departments that use data in their daily operations so that your communication program sensitizes the ones who are closer to the customers or the services or the products: These are the kind of people you want to embark in your data journey.

Empower people to remediate the data

In many cases, data owners realize that they should not manage everything in their data domains, and thus need to act as orchestrators rather than doers. The collaborative part of data management here makes a great deal of sense. You need to engage with — occasionally or regularly — the ones who know the data best for data certification, arbitration, resolution, or reconciliation. Using applications such as Talend Data Stewardship, data stewards can design, orchestrate, and launch “stewardship campaigns” that ask for identified contributors’ key inputs to enrich your data dynamically.

Through this process, anyone can be promoted at any time to be a data steward who participates in the data value chain. These data stewards can resolve and validate inconsistent data in a user-friendly application, which is fully operationalized by the steward campaign manager.



» *Figure 10: Perform data remediation tasks with Talend Data Stewardship*

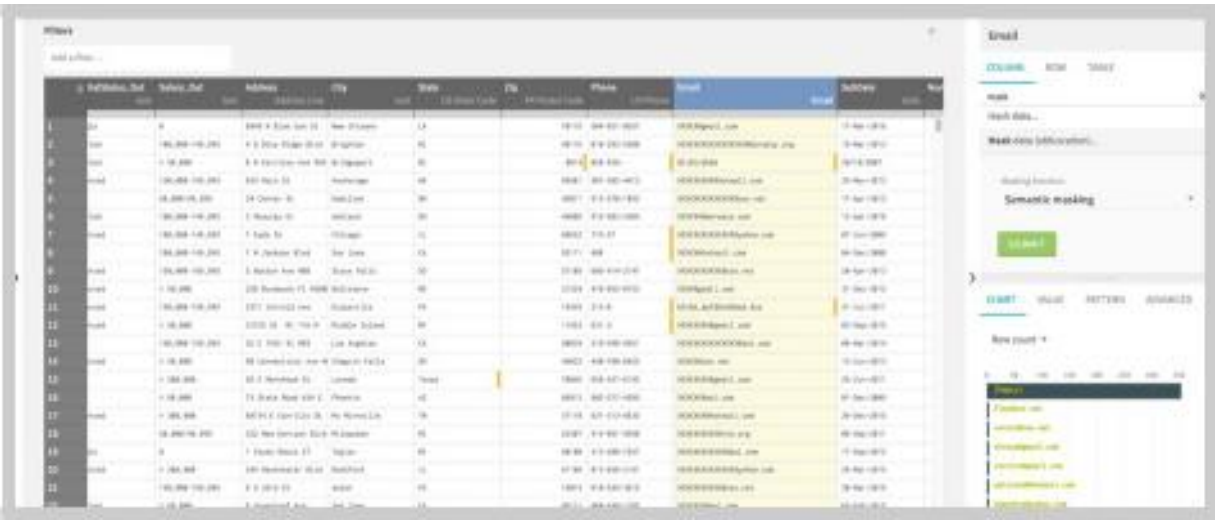
Empower people to protect and mask data

The data governance team may also delegate responsibilities for data protection. A typical example is data masking. In a data lake, for example, IT specialists might not be responsible for data masking and might even not have the authorization privileges to process the data before it has been masked. You need to be able to delegate the data protection task to people who might not be technical experts with deep expertise in the data masking discipline.

This is why it is important to empower a large audience to mask the data on their own so that once they identify specific scenarios where sensitive data may be exposed, they can proactively act on it automatically with a user-friendly tool. That's why Talend delivers data masking across its apps, from Talend Data Catalog to Talend Data Preparation and Talend Studio.

Let's consider this use case. A campaign manager prepares an event with a partner, but lacks explicit consent from customers to share personal data with third parties. Thankfully, Data Preparation offers a collection of drag-and-drop data masking actions that the campaign manager can apply directly to this data to ensure that the data can be easily shared without violating data privacy rules.

We've now reached the end of step 2 of the three-step approach to deliver data you can trust. Data is now accessible in a single point of access and reconciled so that you can understand the relationships and lineage between the datasets. You can then define responsibilities for those datasets and allow the data owners to curate, remediate, or protect data, or delegate those tasks to other data professionals or business users. Now you have created an always accurate, single point of trust for your data assets.



» *Figure 11: Masking data with Talend Data Preparation*

Step 3: Automate your data pipelines and enable data access

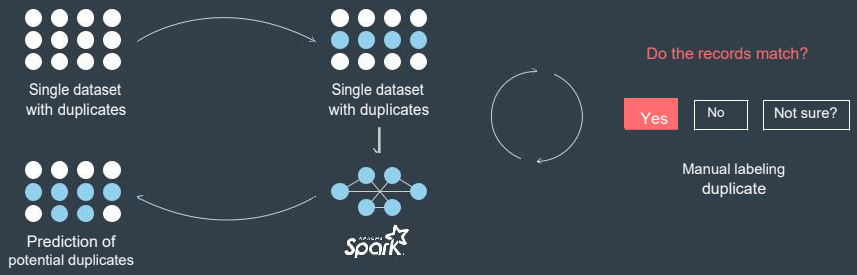
Now that your data is fully under control, it is time to extract all its value by delivering at scale to a wide audience of authorized humans and machines.

In the digital era, scaling is a lot about automation. In the second step of this approach, we've seen how important it is to have people engaged in the data governance process to make it happen, but the risk is that they become the bottleneck (remember the Encyclopedia Britannica metaphor?). That's why you need to augment their skills, free them from repetitive tasks, and make sure that the policies that they defined can be applied on a systematic basis across data flows. Technologies such as data integration and machine learning can help.

Leverage the power of automation to streamline your **dataflows. Use machine learning** *to learn from remediation and scale faster.*

Advanced analytics and machine learning help democratize data governance and data management because they make things much simpler. They improve developers' productivity and empower non-data experts to work with data as well by suggesting next best actions and guiding users through their data journey. Let's take an example from Talend Data Preparation; when the software identifies, by introspecting the data, that it might be worthwhile to standardize the value of data in a textual column, it suggests that the user try the function "find and group similar text." This function uses an advanced data quality function called text clustering. Through the use of smart functions such as pattern recognition or recommendations, the software can make this sophisticated function useable and relevant to a non-data-savvy audience.

Streamlining dataflows with automation

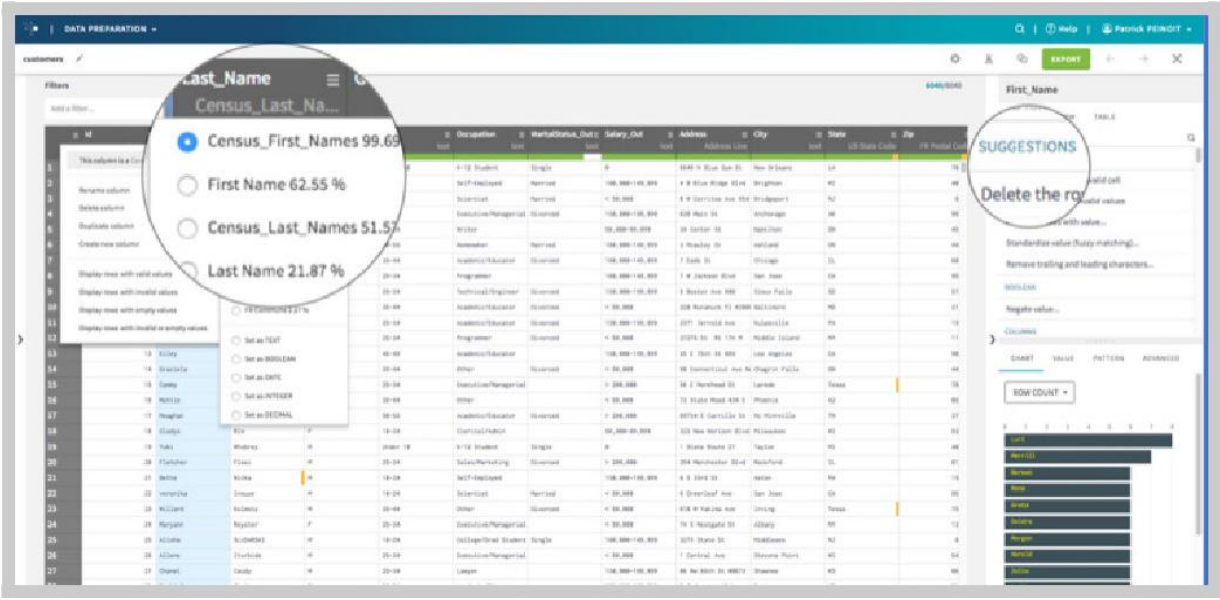


Machine learning also allows the capture of knowledge from business users and data professionals. One typical use case is data error resolution and matching. By using self-service tools such as Talend Data Stewardship for deduplicating records on a data sample and then applying machine learning to the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

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.

Our vision is to put data at the center of everything we do. All routine actions performed, customer engagements made, and decisions taken are powered by big data and analytics.

Adrian Vella, head of data and business intelligence , Tipico



» *Figure 13: Smart assistance with machine learning in Talend Data Preparation*

Automate protection with always-on data masking

Data masking gives you the ability to selectively share production-quality data across your organization for development and analysis without disclosing personally identifiable information (PII) to people who aren't authorized to see it.

If you fail to establish data privacy, you will leave your company exposed to risk, negative reputation, and data privacy regulatory penalties. To deal with that, you need to find a way to automatically spot sensitive datasets, which is one if the core capabilities brought by data cataloging technologies.

A data catalog is the typical starting point for automating the personal data identification process. Once you have defined your data elements with PII, you can automatically spot the data-sets that relate to them. You might want to apply data masking on those elements to lessen the burden of compliance. The more you protect, the better your data management will be secured. If personal data is not necessary for testing or analytics, why would you take the risk of exposing it? Minimize risks while protecting the data meaning thru masking. By doing so, you can anonymize data sets so that individuals can no longer be identifiable.

In the past, usage of disciplines like data masking has been limited to a happy few. With the explosion of data privacy scandals and the proliferation of regulations, you need a much more pervasive approach to data masking that you can embed in your data flow. Only then can you share production-quality data across your organization for business intelligence without exposing personally identifiable information.

» *Figure 14: Add always-on data protection*

GDPR focus:

Article 25 in the GDPR establishes data protection by design and by default, while recital 26 states that the principles of data protection should apply to any information concerning an identified or identifiable natural person. The laws of data protection should therefore not apply to anonymous information, namely, information that does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not, or is no longer, identifiable.

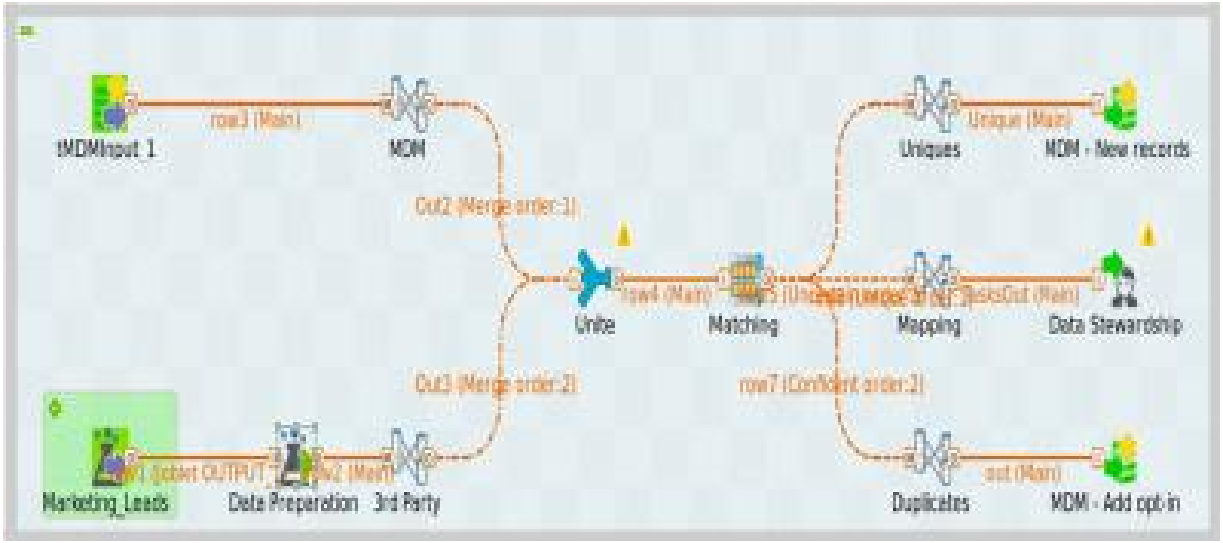
A data governance office must establish controls to mask or encrypt sensitive personal data appropriately. The data masking standards need to ensure that data cannot be reconstructed when multiple fields are combined. For example, data scientists may request that the employee name field should be masked before any analytics. However, a smart data scientist may be able to discern the identity of an employee by looking at title, compensation, and gender (e.g., “Director of HR who is a female with a base salary of $200,000”). In this situation, it may be more appropriate to mask job title and to provide a salary band, such as “above $100,000.”

Automate your data pipelines

Many data governance approaches fail because they can-not be applied in a systematic way. Take the example of a data inventory. Surveys show that in most cases, data inventories are created with a declarative approach, based on interviews of data and process owners and documen-tation using form-based tools or Excel. Creating this data inventory is a short, human resource-intensive process, and the inventory becomes outdated as soon as soon as the data landscape changes.

This is why modern data governance controls need to be embedded into the data chain, so that it can be operationalized and cannot be bypassed. It helps data engineers to orchestrate and automate all your data pipelines. Talend can act as an orchestrator to operationalize and automate any jobs or flows so that you keep on structuring and cleaning your data along the data lifecycle, all the while putting stewards at work for validation, users for curation, or business users for data preparation.

Within Talend, Talend Studio is at the cornerstone of all the data flows: It offers a rich set of technical functionalities that span from integration to data profiling and numerous data quality controls. It allows businesses to operationalize tasks that have been captured in self-service tools such as Talend Data Preparation.

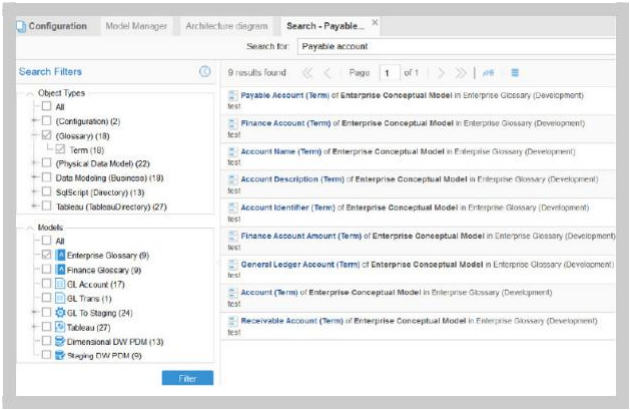


» *Figure 15: Automate data pipelines with Talend Studio*

Enable data you can trust with easy search-based access to good quality data

Delivering data you can trust is a discipline that allows businesses to centrally collect data, maintain its accuracy, and publish it under specific rules and policies. The beauty of the approach is that not only does it control data but liberates it for consumption as well. It allows data professionals and to find, understand, and share data 10 times faster. Data engineers, scientists, analysts, and even developers can spend their time extracting value from those data sets rather than searching for them or recreating them — removing the risk of a data lake turning into a data swamp.

A data catalog is not only a place for data owners to curate and govern the data. It makes also data more meaningful for data consumers, because of its ability to profile, sample, and categorize the data, document the data relationships, and crowdsource comments, tags, likes, and annotations. All this metadata is then easy to consume through full text or faceted search, or through visualization of data flows. Talend Data Catalog makes it possible to locate, use, and access trusted data faster by searching and verifying data's validity before sharing it with peers. Through its collaborative user experience, it turns data consumers into data curators, and enables anyone to contribute metadata or business glossary information.



» *Figure 16: Providing search-based access to data*

Top takeaway:

In its report "Data Intelligence Software for Data Governance," IDC advocates the benefits of modern data governance and positions the data catalog as the cornerstone of what they define as data intelligence software. In the report, IDC says a "technology that supports enablement through governance is called data intelligence software and is delivered in metadata management, data lineage, data catalog, business glossary, data profiling, mastering, and stewardship software. "

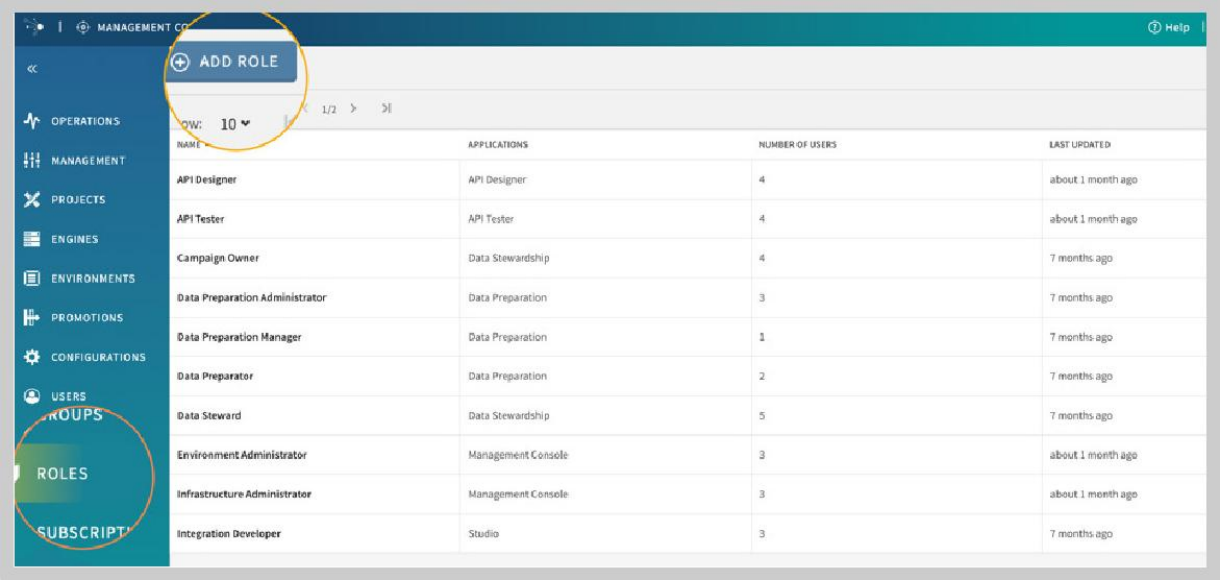
Enable everyone: Establish one platform for all and leverage user-friendly apps for your stakeholders' community

Although a data catalog helps data consumers find their data, their data experience doesn't stop there. Now that they have found the data, they need to put it to work. That's where they need simple, user-friendly apps to fits to their roles. No one-size-fits-all applications can meet the needs of a business analyst, a data engineer, and an IT developer. The cloud is a perfect deployment model to deliver those kinds of ready-to-use applications that can point to any datasets wherever they might be located. Data architectures such as Talend Data Fabric are designed to provide a whole set of collaborative apps that help to consume the trusted data that the three steps presented here helped to create.

Data services publish trusted data into apps

Alongside your data governance project, you have now created data assets that are ready for reuse. These shouldn't be limited to the business users in your organization; business applications should also be able to benefit. These apps may include financial payment services, marketing campaigns orchestrated by Marketo, or learning apps that crawl your profile to offer you the best programs. Identify apps that require trusted data and connect them to the extremely valuable datasets that your data governance program has established. All the created and automated data pipelines can find a destination not only into business intelligence dashboards but also into apps that will get the most of it.

That's one of the key benefits of APIs; they allow a wide range of applications to consume data assets in an easy way. By leveraging environments such as Talend API Services as part of your data governance platform, you can make sure your investment reaches far beyond analytics to feed any apps your organization uses.



» *Figure 17: Enabling everyone in the business to use trusted data with Talend Management Console*

Chapter 4:

Dos & don’ts: the 12 labors of the data governance hero

Dos

Set clear expectations from the start

One big mistake would be to forget or ignore the rationale behind data. So don't just govern to govern. Whether you need to minimize risks or maximize your benefits, link your data governance project to clear and measurable outcomes. As data governance is a nondepartmental but company-wide initiative, you need to prove its value from the start to convince leaders to prioritize and allocate some resources.

What is your "Emerald City"? Define your meaning of success

In "The Wonderful Wizard of Oz," the Emerald City is Dorothy's destination at the end of the yellow brick road.

Success can take different forms: reinforcing data control, mitigating risks or data breaches, reducing time spent by business teams, monetizing your data, or producing new value from your data pipelines. Meeting compliance standards to avoid penalties is crucial.

Secure your funding

As you're building the fundamentals of your projects and you're defining your criteria for success, explain the why, the what, and the how. Then make sure you don't forget "how much." Identify associated costs, involved resources. If you're a newly assigned data protection officer and make sure you have a minimum secured operating fund. If you're a chief data officer, ally with the chief technology officer to secure your fundings together. Then defend your proposal to your finance team so that they understand how the company's risks are linked to failed compliance, and explain the value of your data strategy and all the hidden potential behind data. Make sure you give them the perspective of data as a financial asset.

Don't go it alone

As we know, and it cannot be said often enough, a data journey is not a single project to be tackled by IT.

Even if you can go fast with tools and take advantage of powerful apps, delivering trusted data is a team sport. Gather your colleagues from various departments and start a discussion group around the data challenges they're facing. Try to identify what kind of issues they have. Frequent complaints are:

- "I cannot access datasets easily."
- "I don't find the right data I am looking for."
- "Salesforce data is polluted."
- "How can I make sure it's trusted?"
- "We spent too much time removing duplicates manually."
- "I cannot access datasets easily."

You will soon discover that one of the biggest challenges is to build a data value chain that various profiles can leverage to get more trustworthy data into the data pipelines. Work with peers to clarify, document, and see together how to remove these pains. Embark people on your data journey and give them some responsibilities so your project won't be your project but a team project. Show that the entire success will not be for you but for all team members.

Bring people on your data journey and give them some responsibilities so your solo project becomes a team project.

Apply governance with a yes

Avoid too much control and an overly authoritative top-down approach whenever possible. On the contrary, apply the collaborative and controlled model of data governance to enable controlled role-based applications that allow your data stakeholders and the entire stakeholder community to harness the power of data with governance put in place from the get-go.

Make sure that the business understands the benefits, but also that stakeholders are ready to participate in the effort of delivering trusted data at the speed of the business.

Start with your data

Traditional governance projects often plan to apply a non-negotiable top-down approach to assign accountabilities to data. While you should spend time in getting directions on your data governance, the truth is it won't be super productive as you'll often confront high levels of resistance. Instead, start with your data and the people using it. Modern data experts should listen to business experts and collaborators, get into data sets to detect business value and potential business risks, then identify who is using the dataset the most. Power users are often most inclined to protect, remediate, and maintain dataset integrity."

Data governance is not a project; rather, it's an ongoing process.

Nitin Kudikala Customer Success Architect, Talend

Consider the cloud on your route to trust

Gartner predicts that "by 2023, 75% of all databases will be on a cloud platform, increasing complexity for data governance and integration." The move to cloud is accelerating as organizations need to collect more data, including new datasets that are created beyond their firewalls, deliver that data in real time to a wider audience, and seek more agility and on-demand processing capabilities.

Because your data can be off premises, running on top of third-party infrastructures, using the cloud might require stronger data governance principles. Take the example of data privacy, where regulations mandate that:

- you establish controls for cross border exchange of data;
- you make policies for notification of data breaches, that you establish key privacy principles such as data portability, retention policies, and the right to be forgotten;
- you establish rigorous practices for managing the relationships with vendors who process your personal data.

The cloud brings new challenges for your data governance practices, but it brings many opportunities. At Talend, we see that a majority of our customer are now selecting the cloud to establish their single source of trusted data. Depending on your context, there's a good chance that the cloud is the perfect place to capture the footprints of all the data in your data landscape, and then empower all the stakeholders in your data-driven process with ready-to-use applications to take control of and consume the data.

Be prepared to explain “data”; don’t expect people to have your expertise

Employees often lack digital literacy, but as data becomes predominant in organizations, data literacy becomes a requirement. That’s one part of the problem. As data is becoming more predominant in organizations, everyone will require data literacy. You can use a data catalog to make your data more meaningful, connected to business context, and easy to find. Leverage cloud-based applications such as Talend Data Preparation and Data Stewardship so that they can access data in a few clicks without specific training.

Prove the data value: “start small to deliver big”

Skeptics will challenge you on your ability to control and solve their problems. Don’t take for granted people that will understand your data has value. You will need to prove to them that they will save resources and money by working with trusted data. Take a data sample like a Salesforce dataset, for instance, or a Marketo data source. Use data preparation tools to explain how easy it is to remove duplicates and identify data quality issues. Show the recipe function that allows users to effortless reproduce their prep work to other data sets. Make sure that everyone understands the benefits of data quality, and that they can, for example, use proofed customer contact data to improve the ROI of their sales and marketing activities.

Another quick win is to show them how easy it is to mask data with Talend Data Preparation.

Talend Data Preparation has a very clean, easy-to-use interface, which allows us to get value out of our enterprise data much faster.

Jermaine Ransom, Vice President of Data Services, DMD Marketing Corp

Don’ts

Don’t expect executive sponsorship to be secured

Once you prove business value with small proofs of concept, and you gain some support from the business, ask for a meeting with your executives. Then, present your plan to make data better for the entire organization. Be clear and concise so that anybody can understand your project value. Explain they will gain visibility by endorsing you and improving the entire organization’s efficiency.

You will gain the confidence you need to have your project supported, and your work will get easier.

Be hands on, not hands off, with data. Lead your trusted data project

By starting to meet with different people to listen to their challenges and offer your assistance, you will be seen as the project leader. Make sure all your actions are efficient. As the old saying goes, you have to plan the work and work the plan. Follow up and outline the next milestones of the project. You will confront some obstacles and have to realign priorities as your organization adapts to changing business conditions. Don’t give up. Keep convincing people and explaining how your project helps the company overcome its challenges.

Make sure your data governance is really connected with your data. Too many data governance program have established policies, workflows, and procedures, but are failing to connect with the actual data. For example, a survey showed that among the 98% of the companies surveyed that are claiming GDPR compliance, only 30% could deliver on their promises to fulfill data access requests. When their customers are asking to respect their rights for data accessibility. This means that most companies have established strong governance principles, but are failing to operationalize them.

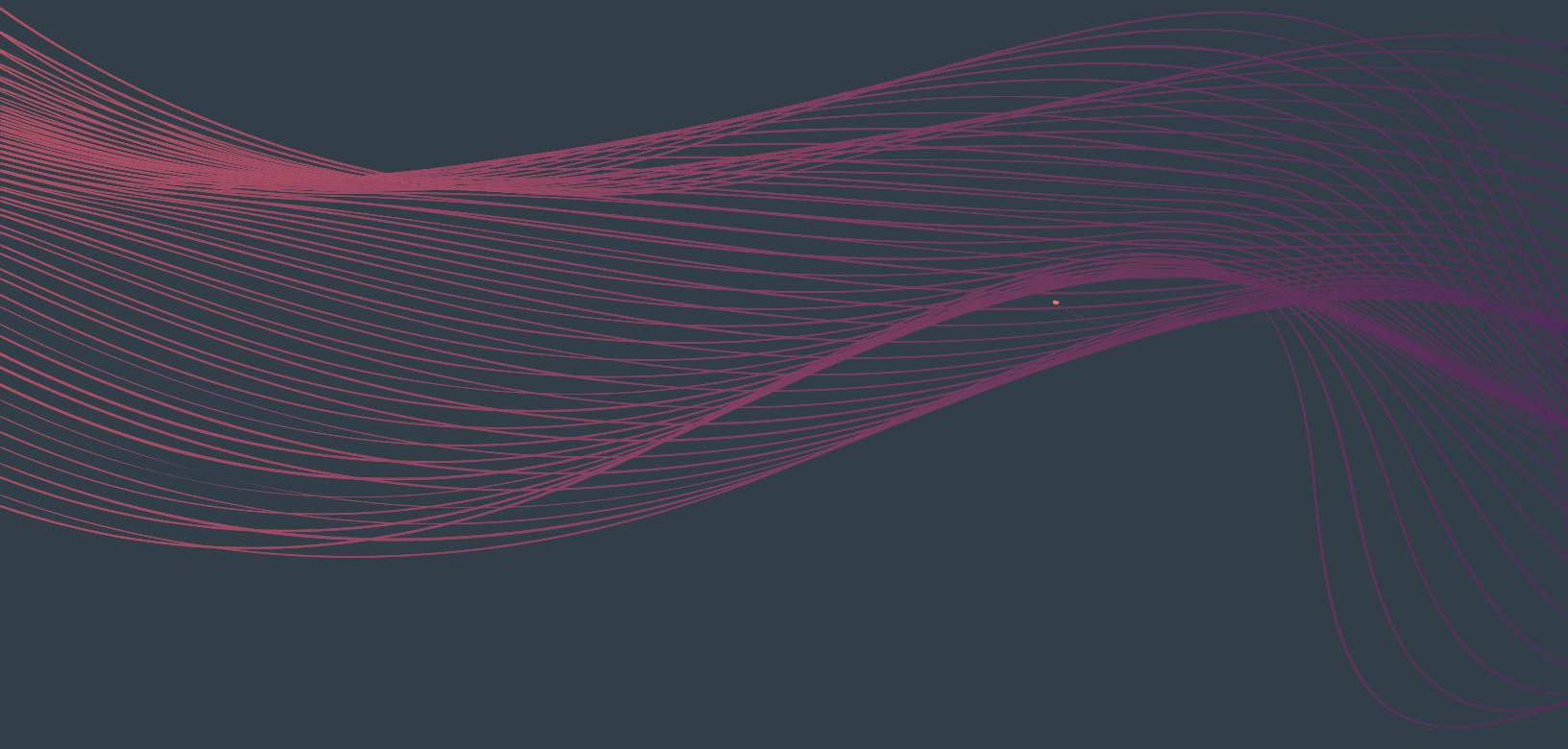
Live your data challenges

Here’s an experiment regarding a crisis in a real-life situation. Imagine you’ve experienced a data breach or a data leak and see if your data governance framework is working in a worst-case scenario. Practice an audit trail. Is all your sensitive data masked? Are you able to track and trace all of your data? Do the data owners feel accountable about the data they’re responsible for? Get in your customers’ shoes and consider their rights for data access or their right to be forgotten.

Consider running a team drill. Make up a breaking news scenario and see how well your plan works, then use those lessons learned to improve it. It’s always better to be proactive rather than just experiencing a privacy incident for real with all the consequences that this entails. A drill can help make data governance more concrete, turning it in operational challenges rather than high-level principles.

Chapter 5:

The new roles of data governance



Let's talk about the roles you should have on your data management team and how they function in the collective data management approach.

Keep in mind there isn't a one-size-fits-all perfect model of the data governance team. Depending on your company's culture and its perception of risk, the team can be more risk-driven rather than value-driven. This said, your team should include specific skills and expertise to understand both compliance regulations and data management so that you cover the full data spectrum to implement the data strategy.

The truth is that data roles in most organizations have been radically evolving over the last few years. They are clearly shifting from centralized to decentralized positions in line of business departments as organizations embrace collaboration to better govern data. Gartner notes, "Key roles such as the data steward are shifting from the IT group to placement either purely in business units or in an IT-business hybrid combination."

In general, there are six critical roles in a collaborative data management framework:

Chief data officers are the conductors of the data strategy. They are responsible for defining, deploying, and tracking the data strategy with the help of the data governance team. The CDO's mission is to make sure data is considered a valuable business asset at the executive level so that the executive committee will invest in resources to be data compliant, to minimize risks, and extract value out of data flows to maximize revenue. Some CDOs become chief trust officers once they make data a trusted asset that can generate positive outcomes for the company.

Data protection officers must data compliance standards as defined by the relevant authorities. DPOs make sure personal data processed by the organization is fully compliant. Examples of personal data can be customers, suppliers, or other individuals' data processed by the organization for their daily operations. Their role is to make sure data is protected according to specific laws and regulations of the related industry or geography. DPOs should have direct access to the upper level of your organization as specified by regulations such as GDPR or CCPA.

Data roles in most organizations have been radically evolving over the last few years.

Data architects are in charge of making sure the data house remains solid and robust yet flexible enough to help users to become masons of their own datasets while annotating, enriching, or certifying data. They set the foundations of the data as an asset to make it meaningful and business-driven for the whole company. They may prioritize information based on measurable business outcomes.

Data stewards are responsible for maintaining the data integrity on specified datasets. For example, the CRM manager may be responsible for the stewardship of customer database. Data stewards have a great responsibility; they need to make sure their datasets meet data quality standards as defined by the data governance team. They may work with information stewards who are their partners in deploying data integrity in their specific geo/division/department.

Data engineers and developers are responsible for designing, deploying, and maintaining the architecture to process complex flows within the organization. Their background is technical, and they will use technical data environments to process data flows. Traditionally, they were in charge of the movement of data, but now they have to know what's in the pipeline, and ensure that the content has gone through quality checks. Often submerged by increasing requests of business users, there are more and more data engineers seeking to give autonomy to the whole data community whilst keeping control of access, authorization, and product administrations.

Data scientists are well-versed in the science of extracting the value out of data pipelines to deliver valuable insights for the company. Data scientists solve complicated data problems using mathematics, statistics, and computer science. They often are experts in statistics, data mining, and predictive analytics. They often also complement their expertise with programming capabilities.

Business analysts analyze the trends and opportunities, identify and calculate risks, and determine the pulse of the business. They are heavy users of business intelligence tools such as Tableau, Power BI, and Looker. Their mission is to get the trusted insights from the data pipelines and to present them in a digested way through comprehensive dashboards.

Top takeaway:

The more you target business people, the simpler and more intelligent self-service apps need to be.

Business users/data curators/data custodians are generalists who come from every department. Their skills and capability level may vary. They may be frustrated when it comes to getting access to data, but they are also eager to get value from data with modern and simple self-service tools.

It's critical that these roles work together in a team. Collaborative data governance is a team sport — like America's Cup — where everyone works on the same boat to win the race, using their unique skills and capabilities and working together to stay ahead of the competition.

Here's an example of how these roles work together in practice.

Let's imagine a company that wants to incorporate weather data to improve the precision of their sales forecasts.

It might start with a data scientist who provides weather data in a data lab to refine the forecasting model.

Once the data scientist confirms that weather data will impact the model, this new dataset might be checked for quality, compliance, and copyrights by a data curator.

Then, a data engineer will automate data integration to ingest a real-time data flow within a corporate data lake.

Finally, business analysts will get access to datasets. They will also share them with and interpret them for business users who are waiting for the data to be delivered.

One approach to executing this data project is the siloed way, where IT controls the access to and distribution of data using data integration tools. Then the data scientists in their data lab could use a data science platform, while the CIO's office would use this data governance framework to ensure compliance. However, how could they work as a team with this siloed approach? Moreover, who could control this disparate set of practices, tools, and datasets?

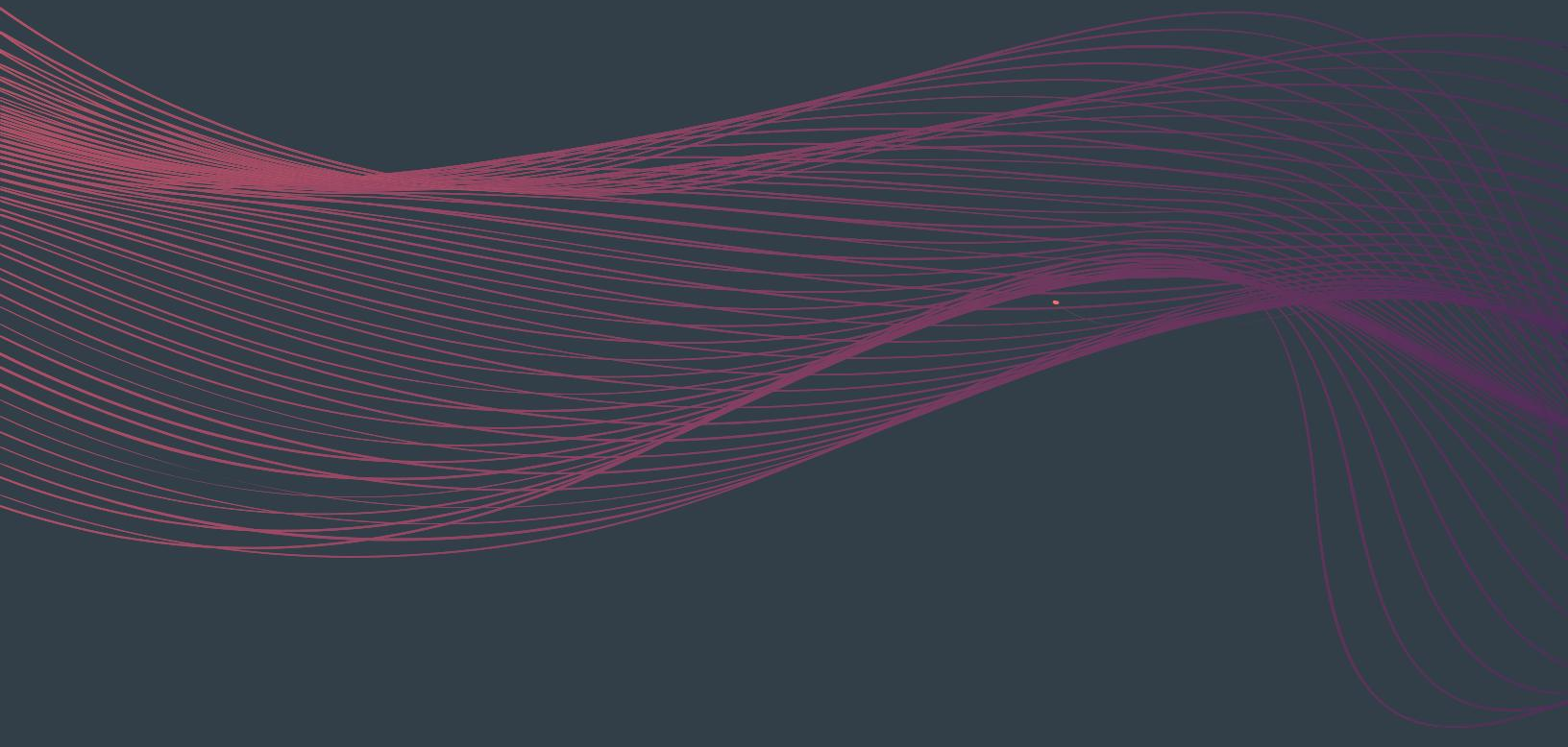
This scenario is what collaborative data management is all about — allowing people to work as a team to reap all the benefits of your data.

GDPR and data protection

Article 25 in the GDPR establishes data protection by design and by default, while article 26 states that the principles of data protection should apply to any information concerning an identified or identifiable natural person. The laws of data protection should therefore not apply to anonymous information; namely, information that does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not, or is no longer, identifiable.

Chapter 6:

Successful trusted data delivery stories



talend

EURONEXT

In the stock exchange sector, we follow three watchwords: integrity, because it is impossible to lose a single order; permanent availability; and governance in a highly-regulated market. Talend has met these expectations.

Abderrahmane Belarfaoui, Chief Data Officer, Euronext

Customer story

Euronext becomes a data trader with a trusted data environment

Following its split from the New York Stock Exchange in 2014, Euronext became the first pan-European exchange in the eurozone, fusing together the stock markets of Amsterdam, Brussels, Dublin, Lisbon, and Paris. Euronext comprises close to 1,300 issuers, reporting a total market capitalization of 3,700 billion euros at the end of March 2018.

In 2016, Euronext began the typical process of migrating its data to the cloud — except that this migration had nothing typical about it at all. First off, the Euronext database contained 100 TB of data — one of the biggest databases in Europe. Then there was the fact that this was not just a simple transfer of a database to a hosted platform. The idea was to create a governed data lake with self-service access for business units and clients in an effort to monetize new services and generate additional revenues.

Migrating to a governed cloud

Before this project was launched, Euronext relied on an on-premises data warehouse appliance with proprietary hardware from one of the big names in the industry.

"Our IT infrastructure had reached the end of its lifecycle in our European operations, where regulators were expecting that Euronext store more and more data," Abderrahmane Belarfaoui, Chief Data Officer, Euronext recalls.

"Moreover, sometimes we had to wait six hours after market close, and in some case even more, before we could send the data to business units and clients."

The situation prompted the CDO to look at moving to a hybrid cloud model, while remaining independent of the cloud provider.

In an ultraregulated world, Talend has also proven to address the challenges of data lake governance and regulatory compliance. Being able to safely open data to new usages — including monetization — and address data privacy — including GDPR — involves knowing it inside out, keeping track of changes and the history of data feeds, and knowing how to classify them in a granular structure.

"We have an Amazon S3 storage that is shared by everyone. I have to know who owns data (the data owner), who has access to what, whom to ask, who can use it, and who has priority over whom. Our data stewards protect the organization of our data," adds Belarfaoui.

This governance strategy is applied in very specific tools, such as the Talend Data Catalog. A dictionary is created together with each technical project for each individual market. These dictionaries are used to find the history of end-to-end data, from the sources to the reporting. "Now I can see when S3 is the data source, I can add value to the data, combine it with other data, and convert it into other data in Redshift," says the CDO, who is very satisfied with the new process. "I can also add tags. Typically, we add the storage duration. For example, whether data has to be kept for 10 years, or five years (per MIFID II), or if it should be archived."

At the same time, data lineage with Talend drastically reduces impact analysis costs. "One simple example comes to mind: We plan to change the value of an index on the British stock market. Once we integrate it into our systems, it propagates itself pretty much everywhere. Currently we have to figure 200 person-days just to find the index in our different systems. But with the dictionary, we are able to run this data lineage with just one click."

Monetizing stock market data

Two years after its launch, the governed lake project with Talend and AWS is a success. "The initial returns are more than positive," says Belarfaoui. "On the technical side, we can manage 10 times more iso-budget data."

Beyond the improved architecture and the business improvement for regulatory compliance, the new platform is also positioning Euronext to become a "data trader." The stock market operator wanted to be able to refine and add to its wealth of data in order to monetize it. In fact, the monetization of data already makes 20% of Euronext's revenues.

This project also involves giving data scientists and business units self-service access to this data, which they can analyze in data sandboxes for tasks such as market monitoring.

This is a real turning point for Euronext. "In 2016, we identified the need, but we didn't have the capacity to do it. At the time, we could only relay the volumes of market activity to market regulators (Mifid II). Today, we can dig deeper. Under the General Data Protection Regulations (GDPR), I have to know where personal data is stored. If I receive requests for modification or deletion, I can find the data, thanks to the dictionary," elaborates Belarfaoui. "Similarly, a user who searches a transaction can instantly see if it is confidential. Once data is identified as being critical, the data steward can deny user access."

Modern data governance allows organizations to modernize their data landscape

- Deliver data portability, accelerate change management, and migration cycles
- Engage lines of business and IT for trusted and meaningful data
- Track and trace data movements and processing activities across disparate information systems
- Protect sensitive data and operationalize data policies
- Enable a wider audience with trusted data in a controlled way

Talend is everywhere in our project. We ***benefit from the most advanced concepts****, such as data cataloging, serverless computing, and continuous integration.*

Abderrahmane Belarfaoui, Chief Data Officer, Euronext

Customer story

Uniper delivers trusted data at the speed of demand

Uniper generates, trades, and markets energy on a large scale. With about 36 gigawatts of installed generation capacity, Uniper is among the most significant global power generators. The company also procures, stores, transports, and supplies commodities such as natural gas, LNG, and coal as well as energy-related products.

Uniper customers include large industrial customers and municipalities in German and neighboring countries.

Providing self-service data and analytics in real time

The utility and power industry is in the midst of the biggest disruption in decades. Energy liberalization brings increased competition. Renewables and smart grid technologies have upended assumptions about capital planning, centralized vs. decentralized generation, and the underlying foundation of the business.

“We are in an increasingly complex world of ever-changing technologies and markets,” said René Greiner, Vice President for Data Integration, Uniper SE.

Fast facts

INDUSTRY:
Utilities

INFORMATION:
HQ: Germany
10,000+ employees

USE CASE:
Operational efficiency

CHALLENGE:
Providing self-service data and analytics in real time

TALEND PRODUCTS USED:
Talend Real-Time Big Data
Talend Data Catalog
Talend Data Preparation

RESULTS:

- Integrated more than 120 internal and external sources
- Reduced integration costs by more than 80%
- Supplying data in real-time 10 times faster and 10 times cheaper

PARTNER ECOSYSTEM:
Azure, Snowflake

"We produce energy. We buy and sell energy via marketplaces. How much coal and gas do we need to produce today and in the future? Is the market going to turn in a completely different direction? How shall we expand our market positions? How can we maximize our profit and loss? Before we embarked on our cloud journey, we didn't have our data readily available to make these decisions quickly.

"Once the idea of an organization-wide data strategy emerged, we decided to go with a public cloud solution for reasons of scalability and cost. We concluded that Talend would be the best software for such a cloud architecture," said Greiner. "Talend's ability to connect to a wide range of source systems and its modular product design were also deciding factors."

To make informed decisions, Uniper relies on marketing analytics based on real-time information. Now that the relevant information is aggregated in the data lake, market analysis teams can access data faster and provide answers more quickly to questions they get every day from traders. Questions that previously required months of research can now be answered right away, or in just a few days.

Speed is important in answering questions because the earlier trading teams can react, the earlier they can take a position and that can make a difference of millions of euros.

For gas, for example, it's essential to understand the demand for power and how it changes with temperature, which affects the volume delivered. The data lake makes data available continuously and enables trading teams to automate trading when prices reach a pre-established threshold. So now, Uniper can move immediately to take a position, when, in the past, it would take days to pull together the necessary data.

Modern data governance allows organizations to deliver data lakes and analytics on a wider scale

- Accelerate time to market from initial discovery to publishing
- Improve efficiency for finding data and turning it into insights
- Enforce control and data protection for self-service access
- Crowdsource knowledge for data curation, recommendations, and remediation
- Reach a wider audience with meaningful and trusted data

talend

AIR FRANCE KLM

Fast facts

INDUSTRY:
Air Transportation

INFORMATION:
HQ: France
84,000+ employees

USE CASE:
A 360° approach to caring for each and every customer: "Customer intimacy"

CHALLENGE:
Meeting customers' specific travel needs

TALEND PRODUCTS USED:
Talend Data Management
Talend Metadata Management

RESULTS:

- Tens of millions of unique experiences
- One million pieces of data processed each month
- Access to customer information ten times faster

"Our goal is now to become the airline that best caters to its customers."

Gauthier Le Masne, Chief Customer Data Officer, Air France-KLM

Customer story

Air France-KLM offers 'made -just-for-me' travel experiences with the Talend platform

Air France-KLM is a world leader in three main business lines: passenger transportation, cargo transportation, and aeronautics maintenance. With 90 million annual customers, 27 million FlyingBlue members, and nearly 2.5 million unique visitors on the web each month, customer data processing is a key issue for the Air France-KLM group.

Meeting customers' specific travel needs

In the field of air transport, it goes without saying that competition is intense. It has proven difficult for Air France-KLM to set itself — and its prices — apart from low-cost companies. Making its products stand out against those of its Asian and Gulf competitors has also been a challenge. And the challenge no longer really lies with customization of the customer experience, but with hyper-customization.

"We are entering an era where we must cater to each and every one of our customers," Gauthier Le Masne, Chief Customer Data Officer, explains. Nowadays, "the products themselves are no longer enough. The quality of our relationship with customers and the services they receive is what will set us apart from the competition. When it comes to what satisfies customers, the product only ranks tenth behind more service-related motivators. Customers don't expect their carrier to transport them, but rather to meet their specific travel needs."

Within a few years, the amount of data available to airlines has exploded. Sites and applications also generate numerous interactions. For example, a sale is made every five seconds on AirFrance.com. In addition, there are exchanges with the company's 16 million Facebook fans and three million Twitter followers, as well as data from media campaigns, since Air France-KLM is one of the few advertisers to carry out its own media buying process online.

A Big Data platform for centralizing customer data

While the group began collecting customer data several years ago through call centers, social networks, and its staff at airports, airport lounges, and on airplanes, the data collected to date has not been centralized. Thus, the first challenge was to combine all customer data on a common platform for all Air France-KLM businesses. “The idea was to have our customers’ data centralized on a Big Data platform so that they can be contextually redistributed in real time to all of our customer service points,” Gauthier le Masne continues. The platform was set up in the first half of 2016. “We relied on the Hadoop platform that we already had in place.”

In addition, Air France-KLM may collect and process personally identifiable information (PII) 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 carried out with Talend data masking, which makes it possible to anonymize certain sensitive data and make them unidentifiable in order to prevent unauthorized access. “Talend Data Catalog has helped us implement data governance with data stewards and data owners to document data and processes,” Damien Trinité, CRM Big Data Project Manager, Air France-KLM concludes. “Air France-KLM can locate customer data, determine its origin and destination, and share the information within the company 10 times faster than before.”

Improving the travel experience through a 360° approach to the customer

“As soon as the customer departs, they’re on the move and need support,” Gauthier le Masne explains. The company has to be able to identify its customers’ main stress factors to be able to anticipate them to the highest possible extent and be as proactive as possible.

In the field, call center agents were the first to take advantage of this data management solution. As for the flight crews, all flight pursers have an iPad. “This gives them access to all information about flights and customers. In concrete terms, if a customer usually opts for a vegetarian meal but the reservation agent forgets, the company will take the initiative to offer the option.”

On the sales and marketing side, a recommandation engine has been set up, while data-driven algorithms offer customers promotional rates for their next preferred destinations.

Modern data governance allows organizations to create a data marketplace for their highly shared assets

- Create a data inventory of all data assets that relate to a specific domain
- Reconcile disparate data into a quality proofed 360° view
- Foster data ownership for data stewardship and curation
- Protect sensitive data against inappropriate usage
- Enable people and applications with trusted data

Each and every traveler is unique. With our Big Data and **Talend platform, we offer** *‘made-just-for-me’ travel experiences, from purchase planning through* **the post-flight stage.**

Gauthier Le Masne, Chief Customer Data Officer, Air France-KLM

Customer story

Financial services firm reduces risk exposure with a cloud-based data lake

This financial institution leveraged the Talend Real-Time Big Data platform to collect and integrate hundreds of datasets from disparate sources into the cloud data lake.

Applying data governance for reconciliation, aggregation, and risk reporting

Beginning in the 2007 global financial crisis had severe impacts on the whole economy and on the financial services institutions. Not only did it highlight the need for more holistic approaches for aggregating risk exposures across financial instruments, business lines, and legal entities, but it raised also the regulatory bar and cost of noncompliance.

For this leading financial services company, this challenge drove the creation of a company-wide, centralized data governance shared service. Its goal was to address data quality, data reconciliation, and data reporting through data stewardship and data management best practices.

IT modernization was at the core of the project as well, with cloud and Big Data as the two pillars for bringing all relevant data together onto a flexible, scalable analytics and reporting platform. Through a data lake on Amazon Web Services (AWS), this company was able to collect all the raw data needed for risk aggregation.

Using a data catalog together with data quality management, the company was able to bring trust and transparency on top of this data lake and to propagate it across their risk management process.

Paving the route to trust from raw to compliant data

Using a top-down approach starting with raw data ingested on an Amazon S3 file system, the financial institution leveraged the Talend Real-Time Big Data platform to collect and integrate hundreds of datasets from disparate sources into the cloud data lake.

Data that contributes to risk calculation can be reconciled, documented, tracks, and traced using Talend Data Catalog so that risk professionals can see the end to end data lineage, thereby understanding where the data comes from together with the processing steps applied for risk aggregation. Additionally, Talend Data Quality delivers the reports on the quality and accuracy of risk data and drives data remediation when applicable.

Operationalizing data governance for measurable results

Armed with this solution, not only did the institution reduce the risk of noncompliance, it also reduced the compliance costs by cutting the time and resource needed for reporting — while increasing trust on risk data. Because of the extreme scalability of the cloud, the data lake technologies, and the Talend, billions of data rows from highly heterogeneous sources could be collected and processed daily.

The solution not only provides more accurate and precise reporting and transparency on how risk is calculated, but also brings more visibility into how to remediate to data quality issues. Profiling makes data actionable now that the software pinpoints potential data flaws, links it to the root cause, and empowers people with the tools for fixing it.

Using the Talend Platform on top of AWS, the institution collects data in its data lake, reconciles it against a business glossary of “critical data elements” and tracks and traces at a detailed level the whole process for risk data aggregation. Talend Data Quality then runs on top to report on quality metrics, trigger alerts, and empower data stewards to fix data issues. Now, the company relies on trusted and shared data for both its business and regulatory risk reporting. Backed up with solid and extremely scalable technology foundations, the compliance solution was implemented with a start small, grow fast approach, initially developed for one specific use case, but ready to span across all the regulated activities.

Modern data governance allows organizations to meet their regulatory compliance challenges

- Capturing massive and exponentially growing amounts of varying data across multiple sources
- Creating a data inventory to track and trace where the data came from, where it goes, and how it is processed
- Verifying the quality of that data and easily monitor quality issues
- Generating trusted reports quickly, without manual fixes
- Fostering data accountabilities, setting up policies and data protection rules

Chapter 7:

Managing the transition from data integration to data integrity

Why your employees must be data literate

Now that we've seen the big picture for delivering data you can trust, we need to be conscious that data governance is a journey. It has to take into account that people, processes, organizations, and customers are at different points in their maturity curve.

And that's where Talend can offer a "start small and grow fast" approach. Most data-driven initiatives start by creating a "data place" where companies capture all their data. You could call it a data hub, a data warehouse, a data lake, or a customer 360° — the rationale is the same.

Generally it starts with data capture and data movement and then transformation (for example, aggregation or reconciliation). This is the starting point for data governance. This is where businesses build and run their data pipelines at the speed of their business; with respect to data governance, this is the the origin of data management.

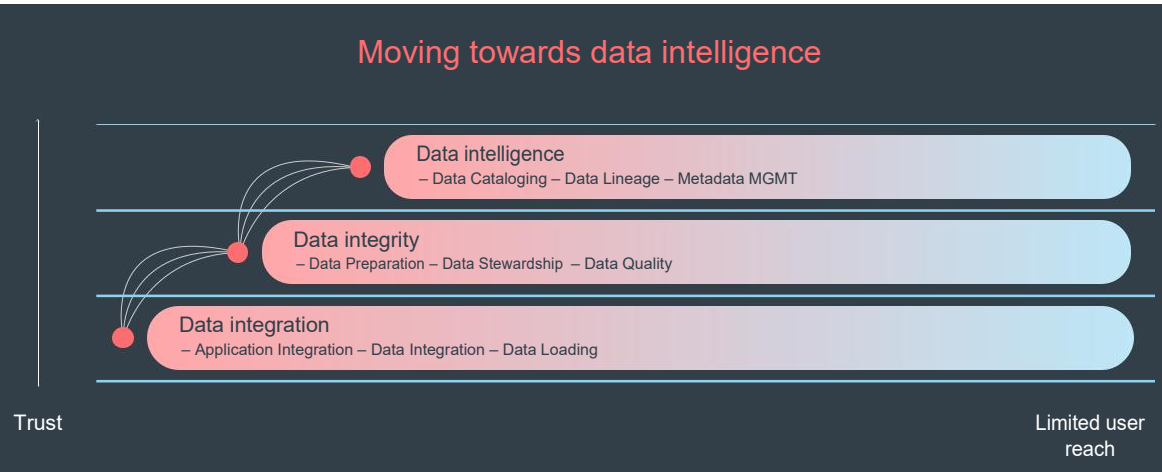
The imperative to have data you can trust requires you put data quality as a core element of your data governance strategy.

Once that's in place, you can take control of your data, with powerful data profiling, data matching with machine learning, and data masking features. If you want to explore integrity challenges in depth, we recommend you download our definitive guide to data quality.

This simple diagram shows the three maturity levels organizations go through as they become more data-driven companies.

This said, the road to data integrity is scattered with traps. One of the biggest obstacles you'll be confronted with is the ability of your communities to understand why data is an asset and how to make it better.

According to Accenture, 78% of business leaders expected their organizations to be digital, yet only 49% of them said they had a strategy for the management and development of the skills needed for the digital world.



Data literacy: still under the radar

On one hand, enterprises are urged to invest in more critical and more flexible storage to cope with a growing amount of data. On the other hand, they put BI and collaboration tools at their employees' disposal.

While focusing on digital tools, they often omit spending time and resources to promote what is needed to deliver trusted data at scale: data literacy.

The lack of data literacy jeopardizes your data integrity transformation

Although processes and responsibilities are of the utmost importance, it's key to focus on your employees' capabilities and literacies so that they become active contributors to your data programs. Otherwise, it's very likely your efforts will be useless.

As you are about to deploy your data strategy, you confront different levels of data skills. Assess your organization's capabilities, or you may deliver an an inconsistent and inadequate learning program that won't engage people and one that may hinder your data strategy distribution.

There are some essential learning steps to follow if you want to improve the integrity level

Ideally, you'll meet with your Human Resources department and Learning and Development (L&D) departments to explain your intentions and align your needs to the professional development offerings provided by your organization. This is an excellent opportunity to ask for learning analytics and assess digital competency readiness inside departments that are active members of your data strategy.

Top takeaway:

Data literacy is the ability to read, write, and communicate data in context, including an understanding of data sources and constructs, analytical methods and techniques applied, and the ability to describe the use case, the application and resulting value. More informally, "do you speak data?"

You can use the GDPR constraint as an excellent way to make this training mandatory in your company. To get an idea of the types of learning programs available, pick and choose in LinkedIn to see the content. If you have lines of business data people such as sales representatives in your radar, consider with L&D investing into a mobile learning application so you can target sales representatives on the go at the point of need on their mobile devices. Make sure L&D invests in the right tools to maximize the efficiency of your data literacy programs.

By 2020, 80% of organizations will initiate deliberate competency development in the field of data literacy.

Apply the 70-20-10 model to your data literacy strategy through digital tools

When defining your skills program, make sure you choose digital apps that get their inspiration from a 70-20-10 model that leverages social interactions, onsite experience, shared discussions, and online/offline training programs. Your key to success is to get the best engagement from your data communities:

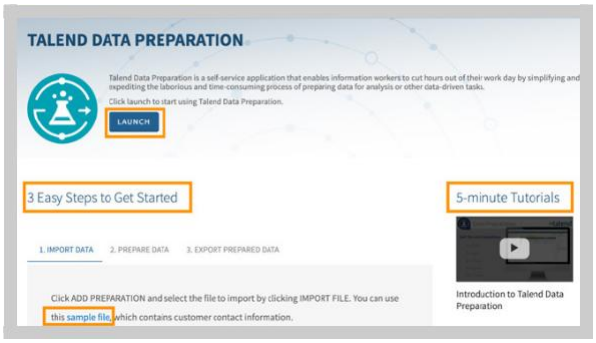
The 70-20-10 Model for Learning and Development corresponds to a proportional breakdown of how people learn effectively.

- 70% from challenging assignments
- 20% from developmental relationships
- 10% from coursework and training

Talend Data Fabric **offers** *what you need to start a data governance program with trials, hands-on, step by step integrated videos, and an integrated user guide to explore more.*

Choose “learning inside” self-service tools to reach the integrity level

Consider investing in simple — but powerful self-service tools built on a unified experience but make sure the work can be operationalized by IT. It will facilitate business and IT collaboration and ease up people work to prepare, curate, and protect data so that you can start using self-service tools without requiring extensive training skills.



Consider deploying blended learning programs

Blended learning programs are learning management system structured paths that combines a mix of online and offline learning experiences. These are the ones that will be best suited to deploy your data literacy programs. People gather online and offline in physical and virtual training rooms to share their experience and benefit from the group effect. Trainers can also follow up the program on their own with delegated autonomy given by your HR L&D administrators.

Blended learning programs are always preferred to 100% online training programs where learners' engagement is generally very low.

Choose your most active members and encourage them to foster these communities through learning apps.

Align internal enablement with your data strategy deployment

Make sure your internal enablement program is synced with your data governance program: Your training program needs to be aligned with your data strategy calendar so people can understand its value and act on its opportunity.

Set regular meetings with Learning and Development colleagues to regularly measure learning completion and decide together on the next steps for data literacy.

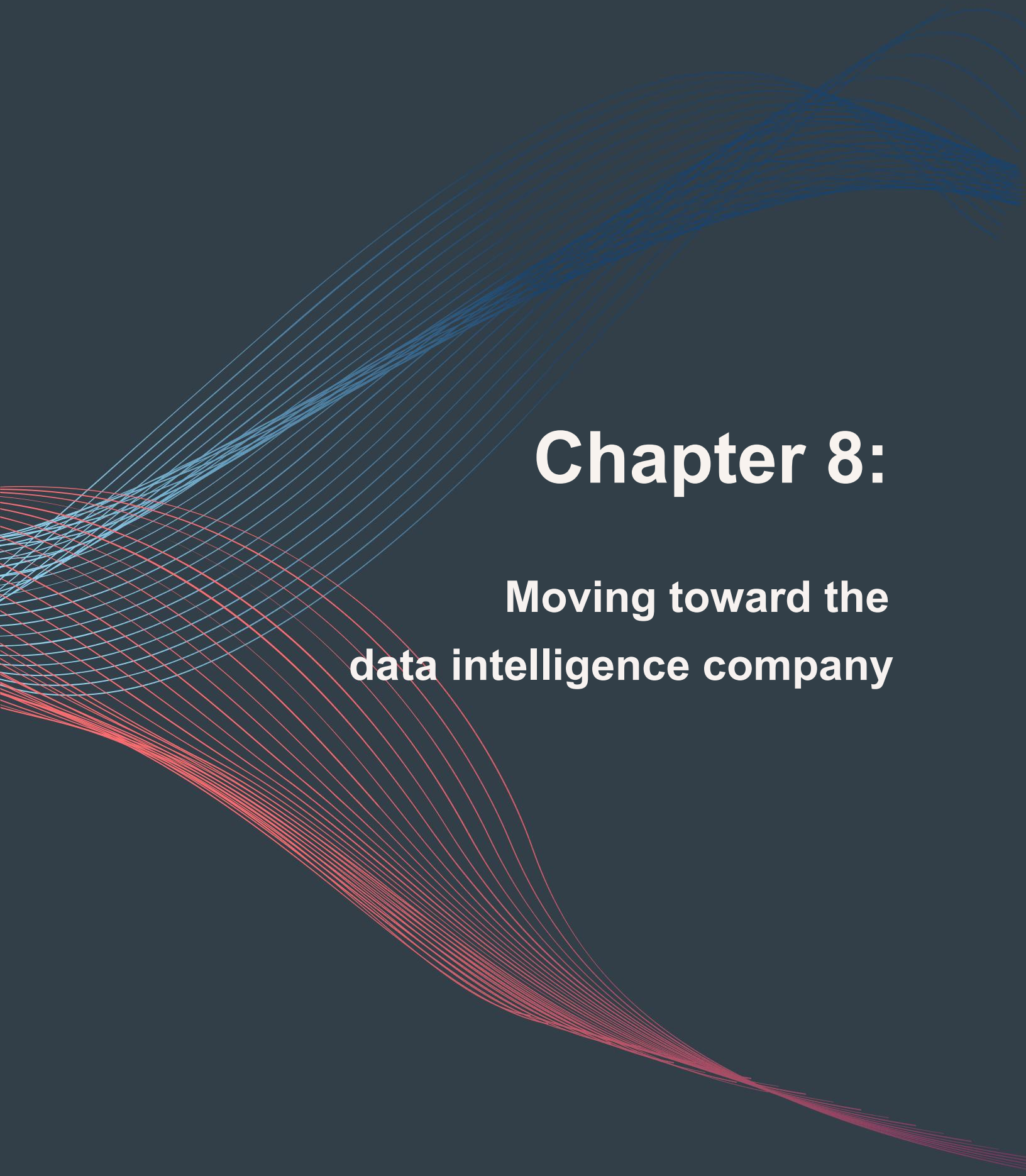
As an example, in the context of the GDPR, off-the-shelf content are provided by learning companies to explain the fundamentals of data privacy so that you can make people aware of privacy fundamentals during the data privacy rule deployment.

Build content that focuses on data as a value

Make sure that your training positions data as a precious asset to protect and a value to monetize. It's frequent that mandatory courses focus on risks to minimize but not on opportunities to maximize. It's critical you empower people to not only protect but also curate and validate the data, so they feel accountable for it. Simply offering training about data protection is not enough.

Data in the 21st century is like oil in the 18th century: "an immensely untapped valuable asset. Like oil, for those who see data's fundamental value and learn to extract and use it there will be huge rewards."

Joris Toonders, WIRED



Chapter 8:

Moving toward the data intelligence company

Sustain your data intelligence strategy with the help of experts

Data intelligence goes even a step beyond with our Data Catalog. That's where you can apply a systematic and automated approach to document your data landscape, create a single source of governance on top, and enable access to trusted data with a search-based interface. This level of maturity is your long-term strategy.

To reach this level, requesting external help and guidance is always the preferred way to ramp up quickly with a partner equipped with the right capabilities and capacities.

As an example, if your program's reason is the need for compliance such as the GDPR Data Privacy Agreements, you might want to request help and assistance to make sure you're compliant with data processing rules, people, and responsibilities.

Expertise in data lineage and data cataloguing enables you to be ahead of any compliance requirements or audit trails by understanding the provenance of your data, who uses it, and how it relates to other data.

Working with an external partner is always profitable if you want to become more intelligent about data. Consulting partners can spot difficulties and identify alternatives to get around the obstacles. You will gain speed, experience, and minimized failure in delivering the right approach.

Make sure your partner is not just there to connect the dots and do simple integration. Data governance requires your partner comes equipped not only with data cataloguing and lineage capabilities but also with some methodology and consultancy on top of data management capacities. So, your partner candidate should combine both technical implementation skills and consultancy services.

We often see failed governance projects because of lack of followup or deficiencies in project leadership over time. So, choose a partner with a consultancy approach and good knowledge of data management with top skilled resources that will stay along with your project from the beginning to the end.

Think about your data governance project like a dream house. You will need a plan and an architect to guide you along with the right resources.

Otherwise, you risk spending too much time with individual contracts and exceeding budgetary limits.

Don’t settle for imperfect data.

Talend, a leader in data integration and data integrity, enables every company to find clarity amidst chaos.

Talend Data Fabric brings together in a single platform all the necessary capabilities that ensure enterprise data is complete, clean, compliant, and readily available to everyone who needs it throughout the organization.

Over 4,250 organizations across the globe rely on Talend to deliver exceptional customer experiences, make smarter decisions in the moment, drive innovation, and improve operations. Talend has been recognized as a leader in its field by leading analyst firms and industry publications.

For more information, please visit www.talend.com.


talend