DEFINITIVE GUIDE TO

DATA INTEGRATION

talend

DEFINITIVE GUIDE TO DATA INTEGRATION: INTRODUCTION

We're at the beginning of the data age. The world has realized that data drives competitive advantage in every business. And the collective innovation of the entire global technology ecosystem has been focused on reimagining what we do with data and how we do it.

Two and a half quintillion bytes of data are created every day, and the volume of data is doubling each year. More data is being produced today than at any other time in human history. In addition, we are experiencing a time of accelerating change as everything in the data world gets continually reinvented from the bottom up. This has been going on for the last five years and will continue for at least the next decade if not longer.

In today's fast-paced, global economy, it is understood that companies must become data-driven in order to remain competitive. In fact, a report from McKinsey Global Institute indicates companies that are truly data-driven—meaning those that can gather, process and analyze data in real time as it flows through the enterprise—make better decisions. According to the report, being data-driven results in a:

- 23x greater likelihood of customer acquisition
- 6 greater likelihood of customer retention
- 19x greater likelihood of profitability

Deriving accurate insights — and acting on them — become a competitive advantage in every market. Having inaccurate analytics or even delayed insights can put you in a vulnerable position—one in which your competitors can overtake your market share by acting on industry and customer needs that you haven't yet identified.

of data are created every day, and the volume of data is doubling each year.

Organizations today are exposed to tremendous opportunities with the growth of technologies such as cloud, machine learning, IoT, and big data. But most businesses would acknowledge that the promise of these technologies has not matched the reality.

Here is that reality:

- 55% of a company's data is not accessible for making decisions.
- Only 45% of organization's structured data is actively used for business intelligence, and less than 1% of unstructured data is analyzed or used at all.
- More than 70% of employees have access to data they should not.
- 80% of analysts' time is spent simply discovering and preparing data.
- 47% of newly-created data records have at least 1 critical error.
- The estimated financial impact of poor data quality is $15m a year on average.
- Knowledge workers waste 50% of their time hunting for data, finding and correcting errors, and searching for confirmatory sources for data they don't trust.
- Data scientists spend 60% their time cleaning and organizing data.

The data ecosystem grows bigger and more complex every day because of the multiplication of data sources.

The growing demand for more data from the enterprise puts a strain on IT's resources.

DATA INTEGRATION AS A BUSINESS STRATEGY

Data integration is a cornerstone of business strategy. It is commonly thought of as a mere technical process; integration gets data where it needs to go and makes data accessible to those who need it. But, as many data professionals have discovered, integration can easily become the main bottleneck to get to the insights.

Gartner estimates that through 2020, integration will consume 50% of the time and cost of building a digital platform.

But this bottleneck doesn't have to exist — if we start thinking about data integration strategically from a business lens rather than a collection of technical processes. Businesses that win have gotten good at making their data infrastructure available, easy to use and maintain, and agile. In this definitive guide, we'll take an in-depth look at the changes and challenges associated with data integration and provide practical steps to developing your data integration strategy. We'll also discuss selecting data integration tools, and how to organize your data team to get the most out of your data investments. We'll take a look at how to get real value from your data quickly, no matter what technology, platforms, or systems you use.

It's time to help data realize its promise. Through a strategic and holistic approach to data integration, it finally can.

CONTENTS

talend

CHAPTER 1

THE DATA REVOLUTION IS HERE: BUT WHAT'S HOLDING YOU BACK?

talend

THE DATA REVOLUTION IS HERE — BUT WHAT'S HOLDING US BACK?

The day an IBM scientist invented the relational database in 1970 completely changed the nature of how we use data. For the first time, data became readily accessible to business users. Businesses began to unlock the power of data to make decisions and increase growth. Fast forward to today, and all the leading companies have one thing in common: they are intensely data-driven.

Data has the power to transform everything we do in every industry from finance to retail to healthcare — if we use it the right way. And businesses that win are maximizing their data to create better customer experiences, improve logistics, and derive valuable business intelligence for future decision making. But right now we are at a critical inflection point.

Challenges with the data revolution

We are currently experiencing a "perfect storm" of data. Data volumes are increasing due to the incredibly low cost of sensors, ubiquitous networking, cheap processing in the cloud, and dynamic computing resources. All this data could be overwhelming, but machine learning and cognitive computing allows us to deal with data at an unprecedented scale and find correlations that no amount of human brain power could accomplish. Knowing we can use data in a completely transformative way makes the possibilities seem limitless, so, business leaders feel there is a powerful enterprise imperative to put their data to work.

Theoretically, given the technology, every company should be a data-driven company. Realistically, though, there are some roadblocks to taking advantage of the power of data:

Trapped in the legacy cycle with a flat budget

The perfect storm of data is driving a set of requirements that is dramatically outstripping what most IT shops can do. Budgets are flat—increasing only 4.5% annually—often leaving companies locked into a set of technology choices and vendors. Many IT teams are still spending most of their budget just trying to keep the lights on. The remaining budget is spent trying to modernize and innovate, and then a few years later, all that new modern stuff that you bought is legacy all over again, and the cycle repeats. That's the cycle of pain that everyone has lived through for the last 20 years.

Only 3% of the data surveyed was found to be of "acceptable" quality.

Lack of data quality and accessibility

Most enterprise data is bad; it's incorrect, inconsistent, and inaccessible, which holds enterprises back from extracting the value from their data. In a Harvard Business Review study, only 3% of the data surveyed was found to be of "acceptable" quality. That is why data analysts are spending 80% of their time preparing data as opposed to doing the analytics that they're being paid for. If we can't ensure data quality, let alone access the data we need, how will we ever realize its value?

Increasing threats to data

The immense power of data also increases the threat of its exploitation. Hacking and security breaches are on the rise; the global cost of cybercrime fallout is expected to reach $6 trillion by 2021, double the $3 trillion cost in 2015. In light of the growing threat, the number of security and privacy regulations are multiplying. Given the issues with data integrity, executives want to know: Is my data both correct and secure? How can data security be ensured in the middle of this data revolution?

Keeping up with innovation

The world of data management and processing, along with analytics needs, is being reinvented from the ground up. Important movements like the rise of cloud are changing your key systems, your integration processes, and your management strategies. And change isn't going to stop; new technologies, some of which could be critical for your business, emerge every year. Your enterprise should be prepared to take full advantage of innovations that enable you to be data-driven and choose the technology stack most prepared to liberate your data, not just today, but tomorrow, and the years to come.

The global cost of cybercrime fallout is expected to reach $6 trillion by 2021 — making it more profitable than the global trade of all major illegal drugs combined.

CHAPTER 2

THE INCREASING DEPENDENCE ON DATA IS FORCING INFORMATION TECHNOLOGY TO CHANGE

talend

THE INCREASING DEPENDENCE ON DATA IS FORCING IT TO CHANGE

Data keeps growing at a dizzying rate, increasing in volume, variety, and velocity. But how, exactly, are companies going to manage it all? Often, this task falls to the IT team, which must create solutions to make data accessible to the rest of the business.

IT is undergoing transformative changes to deal with the data explosion. Five years ago, the most common Big Data use case was "ETL offload" or "data warehouse optimization," but now more projects have requirements to:

- Focus on the velocity, variety, and veracity of data: e.g., move to real-time, ingest streaming data, support more data sources in the cloud, and ensure sure data can be trusted
- Discover new insights through advanced usage of big data such as machine learning and artificial intelligence, and find out how to automate that process
- Run on different cloud platforms leveraging containers and serverless computing

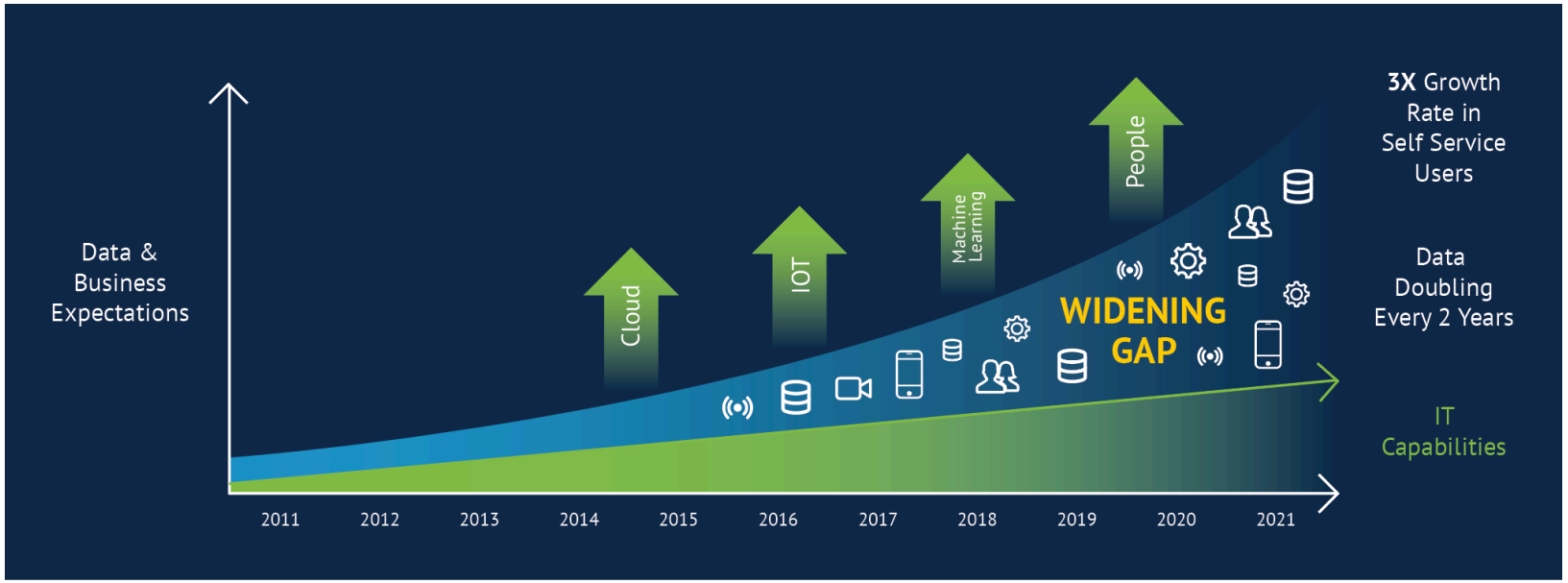
"In this world of accelerating change, how do I make sure I don't get locked into any one technology, because [technical innovations] are happening all across the world?

How do I harness all of that innovation enterprise-wide going forward?"

- Mike Tuchen, CEO, Talend

Almost concurrently, data consumers across the enterprise have multiplied, due to companies recognizing the value of their data. Everyone wants access to data so they can provide better insights, and as a result, new data roles such as data scientists, data engineers, and data stewards have emerged to analyze and manage data. A recent report by Gartner states that by 2020, the number of data and analytics experts in business units will grow at three times the rate of experts in IT departments. Plus, once one department discovers the power of using analytics to make informed decisions, other teams want more and deeper insights for every aspect of the enterprise. That means that IT has to provision all these experts throughout the business with access to data from disparate systems, and bring data sources together in accessible ways.

This is the challenge. IT departments can no longer keep pace with the volume and types of data now available nor the number of users requesting access to it. To try to harness all the data, companies have made ever bigger annual investments in software solutions, in their infrastructures, and in their IT teams. However, simply increasing the IT budget and resources is not a sustainable strategy, especially as data volumes keep on growing and more users pop up eager to get their hands on it.



It's clear that current data economics are broken, and the looming challenge for companies looks particularly daunting with the rise of hybrid cloud and the influx of new applications.

If companies are to realize more value from more of their data, they need to change their thinking in four key areas:

Embracing continuous innovation: enabling the entire organization to embrace all this innovation and get the benefits from cloud computing, containers, machine learning, and whatever comes next.

Disrupting data economics: lowering TCO and improving productivity by taking advantage of native cloud performance, elasticity and security, and automating the data lifecycle to save time and resources.

Delivering data you can trust: architecting data quality and governance capabilities for continuous governance, and building workflows to allow data stewardship by those who know the data best.

Making data a team sport: enabling users throughout the enterprise to self-serve data access rather than controlling access to a few groups.

The emerging technologies most critical to your data integration projects

The organizations that can quickly put the right data to work have a competitive advantage. Modern technologies make it possible to liberate your data and thrive in today's hybrid, multi cloud, real-time, machine learning world. Here are three technology innovations having an impact on data management:

- **Cloud Computing:** The cloud has created new efficiencies and cost savings that organizations never dreamed would be possible. Cloud storage is remote and scales to deliver only the capacity that is needed. It eliminates the time and expense of maintaining on-premises servers, and gives users real-time self-service to data, anytime, anywhere. Cloud service providers such as Amazon Web Services and Microsoft Azure have played a part in encouraging company-wide data-driven practices. They enable businesses of every size to store, process, and explore more data without the monetary and resource investment that was previously required with on-premises technologies. Plus, cloud technology offers cost flexibility, capability and productivity gains that on-premises infrastructure can't compete with.
- **Containers:** Containers are quickly overtaking virtual machines. According to a recent study, the adoption of application containers will grow by 40% annually through 2020. Virtual machines require costly overhead and time-consuming maintenance, with full hardware and an operating system (OS) that needs to be managed. Containers are portable, with few moving parts and minimal required maintenance. A company using stacked container layers pays only for a small slice of the OS and hardware on which the containers are stacked, giving data disruptors unlimited operating potential at huge cost savings.
- **Serverless Computing:** Deploying and managing Big Data technologies can be complicated and costly, and requires expertise that is hard to find. Research by Gartner states, "Serverless platform-as-a-service (PaaS) can improve the efficiency and agility of cloud services, reduce the cost of entry to the cloud for beginners, and accelerate the pace of organizations' modernization of IT." Serverless computing allows users to run code without provisioning or managing any underlying system or application infrastructure. Instead, the systems automatically scale to support increasing or decreasing workloads on demand as data becomes available. Companies are only charged for what they are running at any given time, eliminating the waste associated with on-premises servers. Serverless computing scales up as much as it needs to solve that problem, runs the data, and scales it back down, and then turns off. The future is serverless, and its potential to liberate your data is limitless.

CHAPTER 3

WHAT IS DATA INTEGRATION AND WHY IS IT IMPORTANT?

talend

WHAT IS DATA INTEGRATION AND WHY IS IT IMPORTANT?

Data integration is the process of combining data from several different sources in a unified view, making it more actionable and valuable to those accessing it.

"Without an underlying data integration technology, we'd struggle to operate and meet our business commitments and we'd risk financial penalties from regulators."

- David Clifton, enterprise solutions architect, Affinity Water

There is no universal approach to data integration. However, integration solutions generally involve a few common elements, including a network of data sources, a master server, and clients accessing data from the master server.

In a typical data integration process, the client sends a request to the master server for data. The master server then intakes the needed data from internal and external sources. The data is extracted from the sources, then combined in a cohesive, unified form. This is served back to the client.

Data often resides in a number of separate data sources. Information from all of those sources often needs to be pulled together for analysis.

Without unified data, a single report typically involves logging into multiple accounts, on multiple sites, accessing data within native apps, copying data, reformatting, and cleansing, all before analysis can happen. This takes a lot of time and effort, which is why data integration is so important.

Data integration processes benefit companies in a number of ways:

- **Data integration improves collaboration and unification of systems**

  Employees in every department—and often in disparate physical locations—increasingly need access to the company's data for shared and individual projects. IT also needs a secure solution for delivering data via self-service access across all lines of business in every geography. Additionally, employees in almost every department are generating and improving data that the rest of the business needs.

- **Data integration saves time**

  When a company takes measures to integrate its data, it significantly cuts down the time it takes to analyze that data. Employees no longer need to build connections from scratch whenever they need to run a report. Additionally, using the right tools, rather than hand coding the integration, returns even more time (and resources overall) to the dev team.

- **Data integration reduces errors (and rework)**

  To manually gather data, employees must know every location and account that they might need to explore—and have all necessary software installed before they begin—to ensure their data sets will be complete and accurate. Additionally, without a data integration solution that automatically synchronizes data, reporting must be periodically redone to account for any changes. With automated updates, however, reports can be run easily in real-time, whenever they're needed.

- **Data integration delivers more valuable data**

  Data integration efforts actually improve the value of a business' data over time. As data is integrated into a centralized system, quality issues are identified and necessary improvements are implemented, which ultimately results in more accurate data—the foundation for developing correct business intelligence.

86% of companies saw a 2x improvement in productivity by using a data integration tool rather than hand coding integrations.

Types of Data Integration Initiatives

There are numerous business initiatives that require data integration projects. The most common are:

Managing ETL or ELT

Extract, Transform, Load, commonly known as ETL, is a process where data is taken from the source system, transformed, and delivered into a target destination. This traditional process is changing as technology changes; for example, ELT (Extract, Load, Transform) processes are becoming prevalent.

Creating data warehouses

Data integration initiatives—particularly among large businesses—are often used to create data warehouses on-premises or in the cloud, combining multiple data sources into a relational database. Data warehouses allow users to run queries, compile reports, generate analysis, and retrieve data in a consistent format.

Simplifying business intelligence (BI)

By delivering a unified view of data from numerous sources, data integration simplifies business intelligence (BI) processes. Organizations can easily view, and quickly comprehend, available data sets in order to derive actionable information on the current state of the business. With data integration, analysts can compile more information for more accurate insights.

Leveraging data lakes

Data lakes, central stores for both structured and unstructured data, can be highly complex and massive in volume. As more data sources crop up, more data becomes available for businesses to leverage. That means the need for sophisticated data integration efforts becomes increasingly critical.

60% of customers surveyed by Talend use data integration tools to improve the efficiency of their business.

Business data integration approaches

There are several ways to integrate data. The appropriate approach will depend on the size of the business, the need being fulfilled, and the resources available.

- **Manual data integration** is the process by which an individual user manually collects necessary data from various sources by accessing interfaces directly, cleans it up as needed, and combines it into one warehouse. This is highly inefficient and can create inconsistencies.
- **Middleware data integration** is an integration approach in which a middleware application acts as a mediator, helping to normalize data and bring it into the master data pool. Middleware is useful when a data integration system is unable to access data from one of these applications on its own.
- **Application-based integration** is an approach to integration where software applications locate, retrieve, and integrate data. During integration, the software must make data from different systems compatible with one another so that they can be transmitted from one source to another.
- **Uniform access integration** is a type of data integration that focuses on creating a front end that makes data appear consistent when accessed from different sources. The data, however, is left in the original source. Using this method, object-oriented database management systems can be used to create the appearance of uniformity.
- **Common storage integration** is the most frequently used approach to storage in data integration. A copy of data from the original source is kept in the integrated system and processed for a unified view. The common storage approach is the underlying principle behind the traditional data warehousing solution.

Common data integration challenges

Taking several data sources and turning them into a unified whole within a single structure is a technical challenge unto itself. Common challenges organizations face in building integration systems include:

How to get to the finish line: Companies typically know what they want from data integration—the solution to a specific challenge. Anyone implementing data integration must understand which types of data need to be analyzed, where that data comes from, which systems that will use the data, which types of analysis will be conducted, and how frequently data and reports will need to be updated.

Dealing with data from legacy systems: Integration efforts may need to include data stored in legacy systems. That data is often missing markers such as times and dates for activities.

Dealing with external data: Data taken in from external sources may not be provided in the same level of detail as internal sources, making it difficult to examine with the same rigor. Also, contracts in place with external vendors may make it difficult to share data across the organization.

Keeping up with new technologies: Once an integration system is up and running, the task isn't done. It becomes incumbent upon the data team to keep data integration efforts on par with best practices, as well as the organization's needs and the latest regulatory requirements.

CHAPTER 4
DATA INTEGRATION AND THE CLOUD

talend

DATA INTEGRATION AND THE CLOUD

While classic data integration processes like ETL and ELT have been around for years, the general enterprise move to cloud computing has implications for how data is integrated, where, and why. Today's data integration tools are going to have to not only accommodate new cloud infrastructures,but will have to use all the cloud's power and promise of innovation.

Why data integration is shifting to the cloud

By Philip Russom, senior director of TDWI research for data management

A number of newly mature trends are making cloud-based data integration platforms, technologies, and user best practices more relevant than ever:

The cloud is a well-established platform.

The cloud has become a preferred computing and data platform for modern applications, data, and data-driven business practices such as analytics. The cloud is also becoming prevalent for data management disciplines, including those for data integration and related fields such as data quality, master data management, data warehousing, and reporting.

Many organizations use multiple clouds.

Due to cloud's successful proliferation, many organizations are now "multicloud" because they use multiple brands of software-as-a-service (SaaS) applications (e.g., Salesforce and Marketo) in addition to traditional enterprise applications. This means they manage data on both cloud and on-premises systems. This complex system environment presents unique cloud-to-cloud and hybrid challenges and opportunities that data integration in the cloud can handle and leverage.

Data from the cloud and the internet now coexists with enterprise data.

Many organizations have data sources and targets both on premises and in the cloud because they have embraced big data, social media, the Internet of Things (IoT), SaaS applications, and cloud storage. To accommodate the resulting hybrid data environment, future-facing organizations need to modernize and extend their integration infrastructures to support the Internet and the cloud fully.

Cloud development tends to be agile
The speed of business continues to accelerate, demanding that data, sources, and data-driven products (e.g., reports and analyses) be delivered for business use in minimal time. Cloud-based integration platforms with agile interfaces can compress development cycles by incorporating new data sources and users quickly.

Self-service is a data requirement
Business people and other users are demanding self-service access to data lakes and other cloud-based data sets for analytics so they can perform self-guided data exploration, data prep, and visualization. In a related trend, the only way to scale to increasing numbers of data consumers is to enable them with self-service data access. A cloud-based data integration solution can provide a metadata-driven central point for sharing data and collaborating.

What is cloud data integration?

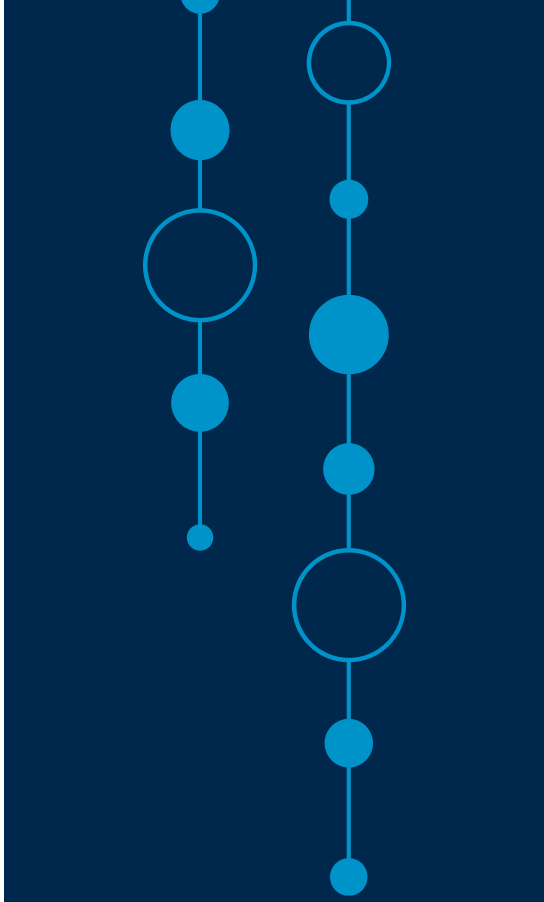
Cloud data integration is a secure, cloud-oriented integration platform that combines the power of traditional data integration capabilities with a focus on extending data integration infrastructure to natively support the cloud. It also supports related functions for data quality, master data, metadata, and event processing plus handling Big Data, IoT data, and other new data sources or targets from the cloud or the Internet.

What does cloud data integration do?

Data integration in the cloud enables developers to design unified solutions that can run natively for local performance or functionality advantages. Because user organizations increasingly have multiple cloud-based applications and data sets, data integration in the cloud should likewise support the native interfaces, calls, and data models of SaaS applications, cloud storage, and cloud data warehouse platforms.

What is the point of data integration in the cloud?

It addresses a real-world need for data integration infrastructure and solutions that reach all applications, data, and people regardless of their types or locations. Comprehensive data integration of this scope is mission-critical as enterprises of all sizes and industries deal with increasingly hybrid and distributed data environments. Obviously, data integration in the cloud helps user organizations establish enterprise-grade integration solutions for cloud environments. It also helps users migrate to the cloud, including the migration of multiple applications, data sets, and user constituencies.



Cloud-based integration platforms can compress development cycles by incorporating new data sources and users quickly.

How can the cloud enhance data integration (and vice versa?)

Several cloud capabilities can enhance data integration solutions in important ways:

Cloud's elastic scalability.

Many data integration workloads ramp up quickly, demand considerable server resources, and then subside just as quickly. Common examples include data ingestion, data transformations, and preprocessing data prior to loading targets. When these occur, an elastic cloud can automatically marshal needed resources, then reallocate resources after intense data integration workloads complete.

Cloud centralization of semantics and collaborative capabilities.

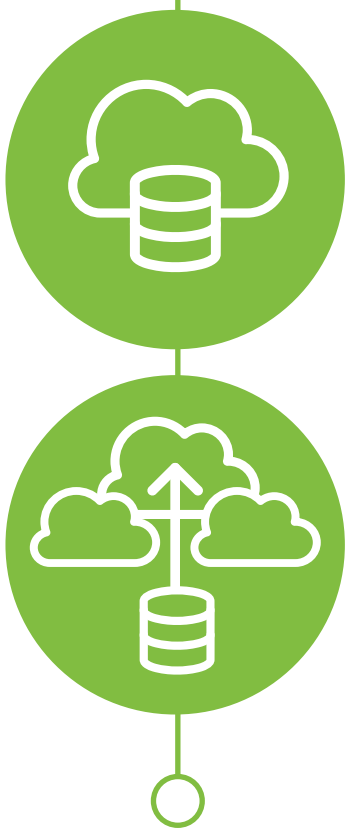
Centralizing shared resources and services makes data management consistent and governable while increasing developer productivity and collaboration. These practices originated on premises but are now available on the cloud, too, so that resources and services can be shared even more broadly among geographically dispersed people and departments, as well as applied in production among the multiple platforms of hybrid data integration workflows.

Cloud's favorable economics.

Server and storage resources tend to cost less on cloud platforms compared to traditional on-premises resources. Furthermore, the cloud provider handles server capacity planning, optimization, upgrades, and maintenance, taking those time-consuming distractions off the plates of data management professionals. Finally, by using cloud-based servers and storage, data management staff need not devote time to system integration or burn up budget on capital expenditures.

"Cloud enables us to develop scalable, robust, timely apps. We selected Talend Cloud as a key part of our cloud strategy, because it allows us to easily integrate our cloud-native applications and our back-end, legacy on-premise customer relationship management environment."

—Tom Murphy, University CIO and vice president for Information Technology, University of Pennsylvania



Data integration's support for cloud falls into two categories:

Data integration running natively in the cloud

Data integration processes executed in the cloud benefit from cloud's scalability, neutrality, and affordability. Furthermore, this architecture places data integration upstream near clouds and other internet-based sources and targets, thus enabling new practices in ingestion, triage, real-time, and streaming.

Data integration interoperating with multiple clouds

To succeed with the extreme complexity of today's hybrid data environments, you need deep support for new technologies (such as Spark) as well as open source tools, Big Data sources, and cloud storage. For the greatest speed and scale (plus richest functionality), cloud data integration must also support interfaces to popular SaaS apps. For use cases in analytics and data warehousing, cloud data integration must also support interfaces to cloud-based data warehouses and other databases.

What are some use cases for data integration in the cloud?

Advanced analytics

One of the strongest drivers for change in data management today is the demand for a wider range of approaches to analytics. Many organizations are choosing to extend their portfolio of advanced analytics by using cloud-based systems. This begins with cloud data integration to pull together large volumes of data from diverse sources as required for the cross-source correlations that most advanced analytics tools operate on.

Cloud data warehousing

The modern data warehouse is, by definition, a compilation of numerous data collections. As users make decisions about how to modernize their warehouse, they migrate some data collections from legacy on-premises databases to cloud-based data platforms. A growing number of users choose to migrate the whole warehouse to the cloud. Whether wholly or partially on a cloud platform, the modern data warehouse needs cloud data integration to migrate warehouse data to the cloud initially as well as to feed the warehouse from hybrid sources during daily production. With so much valuable operational data originating from multiple clouds and the internet, operational reports would not be complete and up to date without cloud data integration.

Multicloud data sync

Some of the most popular SaaS apps today automate sales and marketing business processes—plus other customer-facing functions such as customer service, billing, and shipping. Achieving complete customer views in a multicloud environment—and synchronizing customer data across related customer-oriented applications (whether on premises or in the cloud)—demands sophisticated and high-performing cloud data integration.

“Modern, comprehensive data integration can run anywhere—on premises or in the cloud—to liberate siloed systems to provide the business with the greatest data value.

Data integration in the cloud enables developers to design unified solutions that can run natively for local performance or functionality advantages.”

— Philip Russom, senior director of TDWI Research for data management

talend

Collaboration and the Cloud

New cloud infrastructures bring opportunities and challenges for companies to adopt agile software development models.

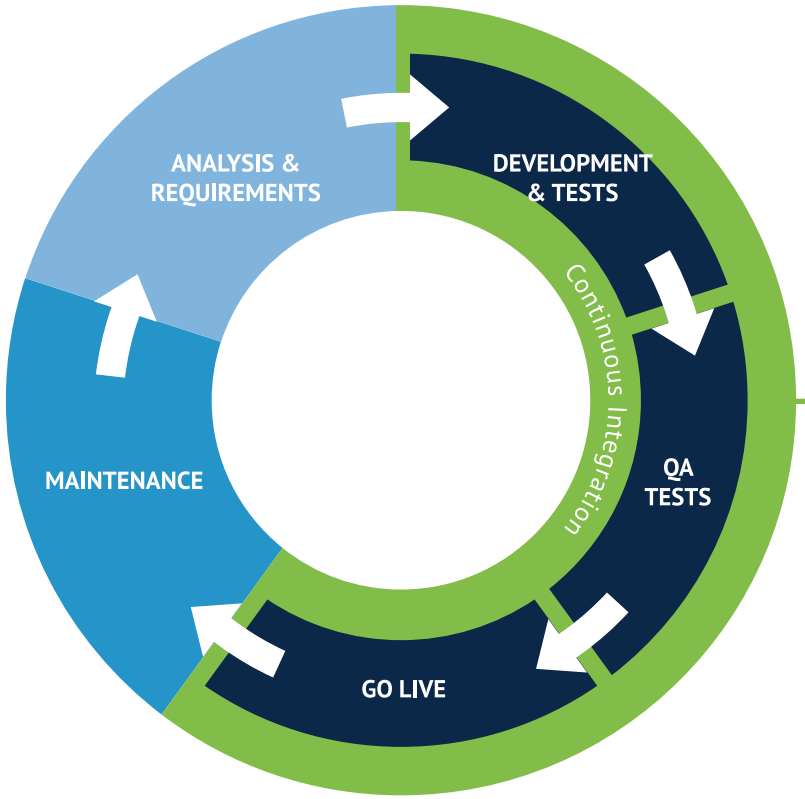
These models allow companies to develop faster and more innovative applications, changing how teams collaborate.

When software is being developed, several people are often working on a single project, and they need to share code with their colleagues so that their work isn't duplicated.

When the software is launched and its maintenance cycle comes around, new and old developers will need to know where the code is and how to work on it. That's why moving to a cloud infrastructure demands well-built tools for facilitating teamwork and collaboration.

For example, continuous delivery (CD) automates delivering software from development to production, ensuring fewer software defects so that the code quality is higher and integration builds can be faster.

In addition, if implemented correctly and practiced regularly, continuous integration (CI) helps in reducing integration problems, thereby allowing you to deliver jobs, code, or software more rapidly.



Also, by integrating regularly, you can detect errors quickly, and locate them more easily. By using the right tools, you could have fewer conflicts and easily resolve them while integrating the code. The most important point is you end up with less of a chance of breaking what already exists; even if it breaks, you can easily solve any problems and recover the data.

All this is made possible with a shared software repository. This repository is designed to consolidate all project information and enterprise metadata and share it with all stakeholders in the integration process, to better facilitate collaboration and communication throughout the Software Development Life Cycle.

There are other key collaboration capabilities for cloud software development tools to consider:

- Versioning, which facilitates item reusability and you can revert to a previous development stage if you need to.
- Git, a widely used source code management system. Git enables better auditing and smaller repositories, such as automated unit testing and revision control.
- Enforcing security and reuse of code through shared projects, rules and user profiles

"Collaboration breaks down the technological and psychological barriers between enterprise data keepers and information consumers. This concept has the power to transform entire industries."

— Jean-michel Franco, senior product marketing director, Talend

CHAPTER 5

HOW TO CHOOSE THE BEST DATA INTEGRATION STRATEGY FOR YOUR ORGANIZATION

talend

HOW TO CHOOSE THE BEST DATA INTEGRATION STRATEGY FOR YOUR ORGANIZATION

Selecting the best data integration method for your organization's needs is only one part of meeting the challenge of how to turn your organization into a data-driven company. There are numerous other factors to consider when thinking about how to use data to better serve your customers and improve business operations: where will your infrastructure be housed? What kinds of technologies will your customers use to interact with you tomorrow, next year, or five years from now? Who in your organization will want access to data? And if the answer is "everyone," how will you make sure they're getting correct data to analyze? How will you comply with not only current data protection regulations, but ones coming online in the future?

This is why data integration isn't a matter of selecting the right software or even the right project; it's a holistic business strategy that impacts your company's capacity to innovate and grow. In order to become a data-driven company, it's essential to understand your organization's business goals, needs, available resources, as well as the overall direction of the data management market to ensure that you create a future-proof strategy that sets you up for success.

"Data agility means procuring, integrating, and enriching data and turning that into business value, on any platform, anywhere, anytime."

— Tim Derrico, director of global analytics, Johnson Controls

The four pillars of a data integration strategy

A common mistake when embarking on a data integration project is to not think about how their integration plans may evolve beyond their initial integration projects.

Your organization's data integration strategy is critical to your ability to grow as a data-driven organization, which often determines your ability to act on data-driven insights before your competitors. As your company's business needs grow, your data needs will grow as well. You need to make sure that your strategy takes that growth into account.

When developing your data integration strategy, consider these four key factors:

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

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

2. **How will you embrace emerging technologies (even the ones you haven't heard of yet)?**

   Most organizations would like to think that they will take advantage of any new technology that will bring their business some tangible improvement. However, some data integration strategies— like manually building your integration stack— will be less capable of enabling you to use new technologies without an enormous amount of development time or procuring a new tool. If you know that you will want to have flexibility in your data environment and the technologies you leverage — for example, accommodating cloud applications and infrastructures — you will need to have an integration strategy that is flexible enough to handle those changes easily. One advantage of using

open source integration technologies, is that they tend to be able to work with new data technologies very easily—especially since many of the new innovative data technologies are based on open source projects themselves.

3. What quantifiable business value is this data integration strategy supposed to bring your company?

While this guide contains many reasons why data integration strategy is important to most organizations profiled here, it is essential to map out how an integration strategy will impact your particular organization. Understanding the clear business value of an integration platform will help you prioritize which features should be most important to you when evaluating different integration methods and vendors. Also, communicating your business goals to your technical teams will help them more efficiently create the data infrastructure that your entire business needs.

4. Do you have the people and technical resources to affect the change you seek in your organization?

Obviously, you will need technical resources to develop, maintain, and scale your integration initiatives. However, the strategies that enable a data-driven company go beyond a single technical project. Instead, they require cultural and organizational change that encompasses all company functions, not just the technical ones. If your integration project is meant to change the data landscape for your company, be sure to work with other stakeholders to envision what those changes mean to IT and the business units who are working with the data.

CHAPTER 6

THE BIG QUESTION: HAND CODING OR A DATA INTEGRATION TOOL?

talend

THE BIG QUESTION: HAND CODING OR A DATA INTEGRATION TOOL?

Many data professionals ask themselves why they need a data integration tool when hand coding can often get the job done quickly and at lower upfront cost. Every IT manager must consider numerous factors when evaluating the trade-offs between hand coding and a tool-based approach to data integration.

Hand coding and data integration tools often go together; understanding when to use each of these methods can be difficult to determine. Most companies use a combination when tackling their technical challenges.

"The allure of hand-coding SQL scripts for data integration can be overpowering. But hand-coding can create unforeseen long-term consequences. What started out as a simple hand-coded SQL script turns into dozens of pages of undocumented, non-compliant scripts that are difficult to repeat, audit, verify, and validate."

— Nick Piette,
product marketing director, Talend

When considering initiating custom-coded projects, here's what IT decision makers must ask themselves:

Be sure to look at both short and long-term costs: While your deployment costs might be reduced by 20% with a custom-coded approach, the maintenance costs will increase by 200%. If you are looking to build a repeatable process that your business is can depend on, a data integration tool may be a more sustainable choice.

There is a time and place for hand coding, but only in very specific situations: Custom coding can make sense for very targeted, simple, one-off projects that do not require a lot of maintenance. It could also be necessary for situations in which there are no tools capable of doing the work required.

Understand that data integration projects requiring multiple developers will benefit from the visual design environments provided by tools: If you have several developers working on your integration initiative, it is important to think about how your developers' work will fit together. With custom coding, there is no guaranteed consistency from one developer to the next, which can make development and maintenance complex and costly. A tool with the option to reuse prior development elements will keep your integration team from duplicating effort and result in more efficient data integration flows.

It's important to understand that maintenance and support costs that will go along with any projects. If different people maintain and support the code once it's in production, their learning curve with a hand-coded approach will be high. Even worse, if the code is in production for years and the person who developed the code leaves the company, understanding how the integration job works (and how to fix it when it breaks) becomes exponentially harder and more expensive.

The hand-coding checklist:

Given the pros and cons of hand coding, we've developed a checklist of questions to determine whether hand coding or a tool is the right choice for your situation:

1. **Does my development team have the expertise to do this using hand coding?**

   If you're using a new technology such as Hadoop or cloud platform, who is going to do the work and how much ramp time will they need?

2. **Is this what I want my hand-coding experts spending time on?**

   Hand-coding experts are typically about a quarter of the full development team, making them a scarce resource. If a non-expert could do the same work using a tool—and save hours of time doing so—wouldn't you rather have the experts doing something for which their unique skills are required?

3. **Can I do this same work with a tool cheaper and faster than my team can hand code it?**

   Most IT teams are constantly being asked to do more with less. A tool-based approach often allows a lower cost per developer to do the work and accomplish it quickly.

4. **Is this a one-off, stand-alone project, or is this an area where I plan to continue doing more and more development over time?**

   If you are embarking on an initiative using a Big Data or cloud platform, chances are you are going to want to do more and more on that platform over time. If so, relying on expert hand coders will be a very hard approach to scale given the scarcity of these resources.

5. **How portable will this code be if I want to repurpose it on a new technology platform like Spark or Flink?**

   When budgeting your costs and time for an integration project, think about the additional efforts needed to re-develop all of your previous work in addition to the efforts needed for new development. Leading Data Integration tools allow you to simply move from one data processing framework to another, eliminating legacy code situations.

6. Will multiple developers be collaborating on this project?

There are multiple benefits with a tools-based approach when you have multiple developers working together including: easy reuse and code sharing, visual design environments, automated documentation, even wizards and experts to advise the developer.

7. How long will this code be in production?

When embarking on a new project, it's tempting to focus on the time needed to develop and forget how long something will be in production. Often something that takes six months to develop will be in production for five years or longer. If that is the case, the support and maintenance costs of that code will continue for 10 times longer than the initial development work making it critical that you understand your support and maintenance costs.

8. Who will own the maintenance of this code?

If you have only a handful of developers, then they will be the ones forced to maintain and support their code. Eventually, support and maintenance will consume all of their capacity, making it impossible for them to take on new projects that could potentially be tasks that help your organization gain a competitive edge.

9. How often will the code need to be updated to accommodate new business needs or changes in the data sources or targets?

Data sources, targets, and business needs are constantly evolving. If it's reasonable to expect this constant stream of changes, then the cost of maintenance and support will be significantly higher.

How to select the right data integration tool

When you have decided to purchase a data integration tool and you have determined whether it makes sense to move your data integration to the cloud, there are a number of features that most organizations will need to have to meet their integration challenges.

What you should be looking for in your data integration tool:

- It should be able to both read and write from the entire breadth of the data sources you need, whether located in the cloud or on-premises.
- It should be able to do data transformation processes like sorting, filtering, and aggregating.
- There should be data quality and data governance capabilities built in, like deduplication, matching, and data profiling.
- Collaboration tools should be included.
- With the shift to cloud systems, the ability to accommodate Continuous Integration/Continuous Development (CI/CD) processes is a necessity.
- Your data integration tool should be able to work in any environment, across on-premises, cloud, or hybrid infrastructures.
- There are some tasks that may be so custom that an data integration tool wouldn't come with all capabilities out-of-the-box, and that's fine. Just make sure that your tool is flexible enough to allow you to extend its capabilities via custom-built components.

- It should be able to accommodate changing providers easily. It's important to have an data integration tool that works in a multi-cloud environment. It should also be able to accommodate switching providers and deployment environments by simply swapping out a few components while keeping the business logic and transformation logic the same.
- A data integration tool should work well with the latest innovations and should accommodate new technologies easily. Good tools will be able to integrate with serverless technologies, Spark, Snowflake, machine learning, and more — and can quickly adapt to new technology that has yet to emerge.
- Scalability is very important when choosing tools. IT means the tool can easily handle simple processes as well as increasingly complex integrations. That will help you keep pace with the growth of your analytics operation and reuse elements from one project to another, saving time and resources.
- Portability is an important but sometimes overlooked capability for data integration tools. For example, the Apache Hadoop ecosystem is moving incredibly quickly. In 2014 and 2015, MapReduce was the standard, but by the end of 2016 Spark emerged as a new standard. If you hand coded, it was impossible to port that code from MapReduce to Spark. Leading tools allow you to do this seamlessly.
- Choose an ETL tool that fits the future needs of your company. If you know that a major data quality or master data management project may be in the works, make sure you choose a tool that can enable you to expand your data environment.

CHAPTER 7

HOW BECOMING A DATA-DRIVEN ENTERPRISE CHANGES YOUR DATA TEAM'S ORGANIZATION

talend

HOW BECOMING A DATA-DRIVEN ENTERPRISE CHANGES YOUR DATA TEAM'S ORGANIZATION

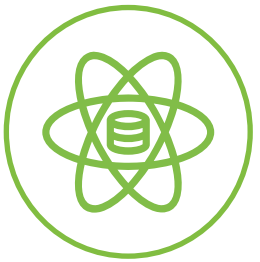
After choosing your data integration tools, you may think your job is done and that your organization can now organically reap the benefits of being a data-driven company. But selecting the right tools is only the beginning of the journey. Making your organization data-driven isn't just about the software you choose — it's about organizing your teams from individual practitioners to company leadership to place data at the heart of the organization. This is what we mean by making data integration a strategy, not just a reactive decision. By developing a company-wide data integration strategy — which includes making data a team sport and enabling collaborative data management — you will be better able to succeed in today's hyper-competitive business environment.

There are four things you must do on a people and process level to make your organization truly data-driven:

Close the business/ it divide: Create built-in collaboration for developers, data scientists, analysts and operations. Work together across the entire data lifecycle.

Enable a data-driven culture: Put the right data in the right hands at the right time. Provide self-service apps tailor-made for every role.

Scale up dramatically: Allow IT to give huge numbers of people anywhere in the organization access to trusted data in a managed and governed way.



Leave no data behind: Use one platform for multiple use cases: Big Data, Cloud, App Integration, MDM, Prep, Stewardship.

The new enterprise paradigm: collaborative data management

IT is facing difficulties meeting the demands for data integration throughout the enterprise. Users want access to the growing pools of data that companies have amassed and the insights they likely contain. If IT doesn't have the resources to provide other employees access to the information residing in corporate data lakes, employees will — just as they do in this BYOD era—find a workaround that could likely put enterprise information at risk.

The best option for IT is to deliver data through self-service access across all lines of business. However, IT must find the proper way to do so in order to prevent exposing company assets to unnecessary risk. They must adopt a model of collaborative data management.

How to implement collaborative data management

The transition from authoritative to collaborative management and governance of company data might be hard, but there's an opportunity for corporate IT departments to create a system of trust around enterprise data stores, in which employees collaborate with IT to maintain or increase the quality, governance and security of data. The good news is that IT professionals have a blueprint from the companies that pioneered the use of the World Wide Web. Just as Web 2.0 evolved around trends that focused on the idea of user collaboration, sharing of user-generated content, and social networking, so too does the concept of collaborative data management.

Collaborative data management breaks down the technological and psychological barriers between enterprise data keepers and information consumers, allowing everyone within an organization to share the responsibility of securing enterprise data.

The greatest challenge – and enabler - for this model has always been trust. Information used to be designed and published by a very small number of data professionals targeting their efforts to consumers who were ingesting the information.

Web 2.0 evolved around trends that focused on the idea of user collaboration, sharing of user-generated content, and social networking — and so has the concept of collaborative data management.

Today, the proliferation of information within companies is uncontrollable, just like it is on the Web. We're all experiencing the rise of a growing number of cloud applications coming through sales, marketing, HR, operations or finance to complement centrally designed, legacy IT apps, such as ERP, data warehousing or CRM. Digital and mobile applications connect IT systems to the external world. To manage these new data streams, we are watching new data-focused roles emerging within corporations, such as data analysts, data scientists or data stewards, which are blurring the lines between enterprise data consumers and providers.

As the Web 2.0 model evolved, trust between consumers and their service providers was established by crowdsourced mechanisms for rating, ranking and establishing a digital reputation (like Yelp). These same positive returns can be realized by enterprise IT departments that adopt selected strategies embraced by their more freewheeling consumer counterparts.

Delivering a system of trust through collaborative data management and self-service is just one of the opportunities available to evolving IT organizations. Through self-service, line of business users become more involved with the actual collection, cleansing, and qualification of data from a variety of sources, so that they can then analyze that data and use it for more informed decision-making.

Collaborative data management is a way for IT to help ensure that the quality, security, and accuracy of enterprise information is preserved in a self-service environment. It allows employees in an organization to correct, qualify and cleanse enterprise information. IT can create governed workflows to provide models for this collaborative, governed data stewardship. The master data records are therefore updated by the people most familiar with them.

Additionally, fostering the crucial shift to more business user involvement with an organization's critical data leads to numerous other benefits. For example, users save time and increase productivity when they work with trusted data. Marketing departments improve their campaigns. Call centers work with more reliable, accurate customer information, much to everyone's satisfaction. And the enterprise gets better control over its most valuable asset: data.

The 6 dos and don'ts of collaborative data management:

Dos

Set your expectations from the start

We've seen how important developing and documenting your data integration strategy is to your business. Setting expectations of how it will achieve your goals is a key part of that. What KPIs should you set? How deeply will it impact your organization's business performance? Make sure these questions are answered from both business and technical perspectives. Make sure you know your finish line, so you can set intermediate goals and milestones on a project calendar.

Build your interdisciplinary team

Of course, you'll need to have the right technical expertise as part of your data integration initiatives. But you also need to include the right people who will understand how your data integration strategy impacts the business and you need to make them your local champions in their respective departments.

Deliver quick wins

While it's key to stretch people capabilities and set ambitious objectives, it's also necessary to prove your data integration strategy and its associated projects will have positive business value very quickly. Don't spend too much time on heavy planning. You need to prove business impacts with immediate results. If you deliver better and faster time to insight, you will gain instant credibility and people will support your project. After gaining credibility and confidence, it will be easier to ask for additional resources when presenting your successes to the rest of the company. Remember that many small wins make a big win.

Don'ts

Don't isolate data integration with technical teams

We often think technical projects need technical answers. But data integration has to be treated as a business strategy as well as a technical challenge. To succeed, your goals need to be widely known within your organization. You have to take control of your own project story instead of letting bad communication spread across departments. You must therefore master the perfect mix of know-how and communication skills so that your results will be known and effectively communicated within your organization.

Don't overengineer your projects

It's important to deliver fast results so you can start small and deliver big. Sometimes those fast results can be really simple. One example is from retailer Carhartt, who struggled to integrate clean data into its systems. So a key initial project was to clean 50,000 records in a single day with Talend Data Preparation. That provided the momentum to keep going with the data integration initiatives.

Don't get distracted — stay focused on clear goals for project delivery

Lastly, it's important to set and meet deadlines as often as possible to bolster your credibility. Your organization may shift to short term business priorities; don't get distracted, but track your route and stay focused on your end goals. Make sure you deliver all your projects on time. Then, when a project milestone is finished, make sure you take time to celebrate with your team and within the organization.

The future of integration: enabling new integrators without losing oversight

Being data-driven is a mandate for modern business, and the strain cannot be placed on IT to simply keep pace with the latest innovations.

This means the business must fully equip its employees at every level to empower their decision-making with highly available and insightful data. As well as providing self-service technologies and applications, there should be training and internal communications to define a data-driven culture throughout business divisions.

There are three steps to enable citizen integrators and move data integration outside IT:

1

Organizations need to guarantee a process in which data can be worked on by non-central IT users, easily tracked with proper metadata and data lineage management, and operationalized by the data engineering team. Without such a governed platform and process, opening data access to ad hoc or citizen integrators can only lead to more data chaos.

2

When data management is considered a team sport, every person working with data will need to understand their role. They need to know how to interact and contribute to the team to get the most accurate data (and insights) possible. An agreed-upon process helps teams work together.

3

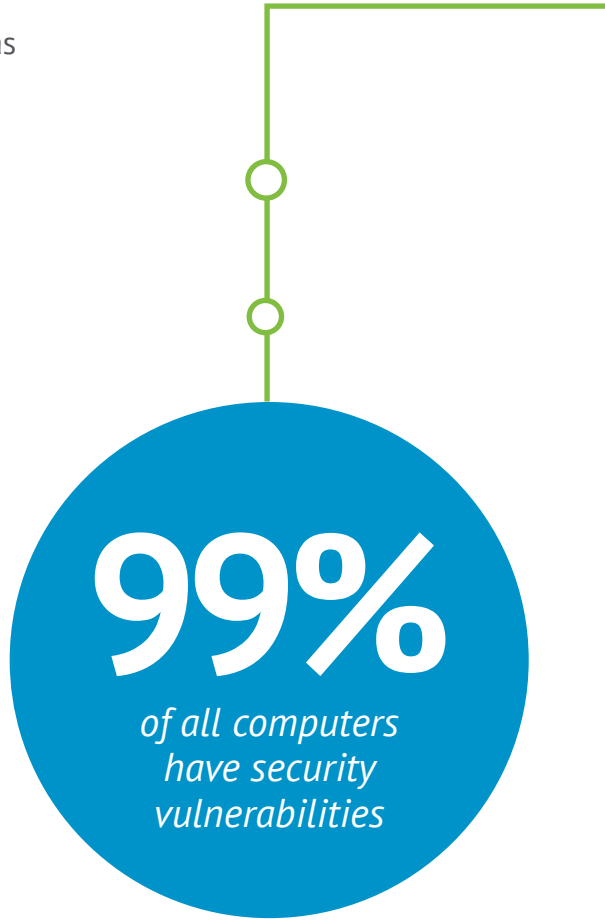
You need to find a data integration platform that will help you enable all of your data team members and operationalize the processes set between data team members all while being governed by IT.

The data security challenge

Everyone in nearly every corporate role, from the CIO to the business process analyst, need data to be right at the user's fingertips. These data professionals must have access to data to ensure they can strategize, execute and deliver for the business with the most relevant and up-to-date insights available.

The knee-jerk reaction to this might be to make as much data as possible available to as many people as possible. However, this is not viable. With regulations like the GDPR coming online, organizations have an increasing obligation to make sure only the right people have access to every piece of information; otherwise, they place their entire organization at risk.

The solution to the problem is to implement governed self-service IT solutions, which automate functions such as data access requests and data preparation. This is fundamental to allowing business employees quicker access to the right data, as well as providing clear lineage of who accessed what information, and when. At the same time, automated data preparation tools are essential to reduce the burden on the IT team, relieving them from performing manual cleansing and formatting tasks. This, in turn, will enable IT departments to focus on delivering new technologies for the organization, rather than troubleshooting small problems for the rest of the business.



CHAPTER 8

BUILD YOUR DATA INTEGRATION STRATEGY WITH UNIFIED DATA MANAGEMENT CAPABILITIES

talend

BUILD YOUR DATA INTEGRATION STRATEGY WITH UNIFIED DATA MANAGEMENT CAPABILITIES

One of the most important factors in any modern data integration solution is that they don't just do data integration. There are a host of other capabilities that need to be built in to data integration software, and ideally, they should be built with one design environment and one set of management tools. This means that developers don't have to install or learn new tools when they want to address a new type of data integration (batch, real-time, big data, cloud and on-premises). This gives you a great deal of of agility and dramatically lowers the cost of ownership - because you can move from one style of integration to another without having to install and learn new integration technologies.

Some of those key capabilities are data quality, data security and privacy, embracing innovation with open source architectures, and automation.

Data integration is useless without data quality

We've underscored the importance of data-driven business intelligence to win in today's hyper-competitive business environment. But the insights that a business can extract out of data are only as good as the data itself. Poor data can lead to difficulties in extracting accurate insights and ultimately poor decision-making. This is something that many executives are worried about. According to the Forbes Insights and KPMG "2016 Global CEO Outlook", 84% of CEOs are worried about the quality of the data they're using for business insights.

The reasons for their concerns are numerous: integrating new sources of data with their existing systems; the financial investment and competitive pressure needed to capitalize on all available enterprise data; and the difficulty of extracting data from the silos which it resides. And their concerns are not unfounded. A study conducted by MIT Sloan School of Management/Review notes that bad data can cost as much as 15-25% of a company's total revenue.

The old adage – garbage in, garbage out (GIGO) – still holds true. Poor data quality adversely affects all organizations on many levels, while good data quality is a strategic asset and a competitive advantage to the organization. If data fuels your business strategy, bad data can kill it.

A proactive approach to data quality allows you to check and measure how clean your data is before it gets into your core systems. Accessing and monitoring that data across internal, cloud, web, and mobile applications is a big task. The only way to scale that kind of monitoring across all of those systems is through data integration processes.

With the right tools, you can create whistleblowers which, as your data is being processed and integrated, can detect some of the root causes of overall data quality problems. Then you will need to track data across your landscape of applications and systems. That allows you to parse, standardize, and match the data in real time. You can set up data checking and correction wherever needed.

How to mitigate data security risks

One of the greatest risks in the modern business landscape is the catastrophic data breach. Chief security officers, CIOs, and even CEOs have lost their jobs following a data breach exposing their customers' sensitive data to external parties. But the repercussions aren't solely limited to an individual or department. A large security failure can cost a company millions and sometimes billions of dollars in fines, but also a loss of public trust, brand deterioration and significant loss of business.

Data privacy and protection are increasingly capturing the attention of business leaders, citizens, law enforcement agencies and governments. Data regulations, which used to be limited to heavily regulated industries such as banking, insurance, healthcare, or life sciences, are now burgeoning across countries and apply to any business no matter their size or industry.

These regulations impact the protection of data and are affected by governmental regulations for data privacy, data storage, data processing, and data transfers across country boundaries. These laws are emerging as an impediment to cloud-based data storage, and they need to be fully understood and considered when information is created in one country but then moved to another country for analytics or processing.

With the acceleration of the digital economy, legislation on data protection and cybersecurity is spreading globally, mandating organizations to establish policies to safeguard personal information, manage the risks related to their data, and address their legal responsibilities.

Organizations should seek to drive alignment between the legal, compliance, privacy, and enterprise data management teams to reuse existing data governance artifacts to support data compliance. In particular, organizations should define personal data elements for data sovereignty and map these attributes to applications in the metadata repository.

For more resources on how to set up data security and privacy controls at your organization, compliant with current data protection regulations, take a look at the Talend step-by-step data governance plan for GDPR compliance.

Open source innovation

Data integration platforms with their roots in open source give you access to the latest innovations faster than other competitors in the market, because so many of the systems being integrated are open source themselves.

Data infrastructures are becoming more complex, so a data integration platform has to run across all of them: on-premises, cloud-only, hybrid, or multi-cloud. Portability should be universal; you should be able to build your jobs once and run them anywhere, on any provider.

In addition, it's important to generate optimized, native code for whichever deployment you choose. There should be nothing proprietary to install or manage on the cluster - that way, you can see the code, and debug more easily. You can design your integration flows once and then quickly redeploy them on other technology platforms as your needs change. This allows you to get the most out of today's IT investments because you can continually reuse these investments on new technologies as your needs change. Your integration strategy can grow with you.

Finally, it goes without saying that your data integration platform should include hundreds of connectors out of the box. Your company needs to connect to many systems. Use pre-built connectors for added speed and productivity.

"Integrating online and offline data helps us develop more ways to communicate with our customers across channels. That kind of interaction drives loyalty."

– Matt Steell, director of information and integration architecture, Office Depot

Scaling through automation

An important capability of data integration platforms is the incorporation of automation and machine learning. To scale the data integration strategy, take the pressure off of IT, and truly make data a team sport, it becomes necessary to automate the data lifecycle: ingest, transform, blend, govern, and discover.

Eliminating hand coding and human intervention in these processes allows you to operate at a new speed and gets more data out of silos and into the hands of analysts who can use it. In addition, machine learning has improved data quality, providing enrichment and insights that get better every day.

This means that for the first time, IT can give huge numbers of people access to trusted data, improving productivity and speed as well as data quality.

A single tool with multiple capabilities

Talend Cloud builds data quality, data governance, and data privacy into the integration process so your team can make trusted data available to anyone. With easy onboarding, embedded quality controls, and rules management, data is enriched, protected, and available from a single, unified platform.

Talend provides a complete data integration, data governance, and data quality solution with built-in data connectivity, profiling, cleansing, matching and monitoring to address all your data quality and data governance needs. With Talend, you can eliminate inconsistent data, enforce business rules and create consistent information through standardization. Data governance teams become more productive by working together to access, understand and standardize data for any data domain such as name, address or product data.

Because of large unstructured data volumes, there is a need to have automated help and assistance when proceeding with data analytics. This allows users to accelerate their time-to-insights when dealing with various data sources and formats. Machine learning plays a key role in this collaborative process.

Talend combines smart semantics and machine learning to turn data into insights faster. Smart semantics automatically capture data footprints in a data pipeline to accelerate data discovery, data linking, and quality management. Machine learning helps to suggest the next best action to apply to the data pipeline or to 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.

All of these tools and capabilities are on a single platform. Everything you need is right at your fingertips for your data integration strategy can grow with your organization.

CHAPTER 9
CASE STUDIES

talend

CASESTUDIES

High tech: Lenovo

Lenovo is a $46 billion personal technology company, the #1 PC maker and #4 smartphone company in the world, serving customers in more than 160 countries.

"Customer expectations have been changing over the years.," says Marc Gallman, director, Lenovo Analytics and Data Platform. "We needed to answer those typical questions: Which options influence customer decisions for our computers the most? Which type of hard-disk would be more preferable to our customers?" To answer those questions, Lenovo is working with a variety of large data sets. But the influx of on-premise, cloud and SaaS-based technologies were creating a data connectivity problem.

"Our focus is to enhance customer satisfaction through marketing tactics. Combining all our data has helped us better know our customers and better serve them."

— Marc Gallman, director, Lenovo Analytics and Data Platform

"We decided to do a hybrid big data architecture of Amazon Web Services (AWS) and our own Lenovo servers. The idea was to maintain the privacy and security of our data and also benefit from the cloud.Talend quickly became a core component of this architecture."

With Talend, Lenovo has built an elastic hybrid-cloud platform to analyze +22 billion pieces of customer information annually. 250+ terabytes of data and 60+ different types of data sources are captured annually across Lenovo 's business units. 8,300 reports are delivered annually to over 615 users across Lenovo providing real-time dashboards, API data feeds and data analysis.

"Using Talend, we have nearly 300 data integration processes running at the same time against a multiplicity of data types and sources, and we expect these numbers to keep rising as we develop the approach," says Gallman. "We have been able to drive up a revenue per unit by 11%. The attach rate for ThinkPad laptop series has also increased by 18%."

With Talend, Lenovo has saved about $140,000 in initial migration costs alone. Lenovo operational costs (employee costs) were reduced by over $1 million within 1 year (34%) while productivity increased by 2-3 times. In addition, Talend has helped improve reporting performance and cut certain process times by a matter of hours to minutes. The ease of use of the Talend platform also allows Lenovo to deliver on requests to continually increase velocity in acquiring data. The time to market on 95% plus of requests is 14 days.

Retail: Office Depot

Office Depot Europe operates in 14 countries through its two main brands, Office Depot and Viking, and is now the leading reseller of workplace products and services. The traditional retail industry is changing with new competition coming into the market and going after corporate customers as well. This is a direct challenge to office supply stores, pulling away some business from their core customer base.

"We're banking on a mix of online/offline business and technology services to drive revenue," explains Matt Steell, director of Information and Integration Architecture, Office Depot Europe. "But there were questions we couldn't get answers to, like Why have we lost certain customers in certain segments and gained in others? What would it mean if we were to increase our spending on certain channels? We needed to be a lot more predictive and be able to test various hypotheses."

The challenge was that data at Office Depot Europe was stored in silos and managed separately by each function, which prevented senior management from getting an enterprise-wide view of customers, operations or finances. "Getting answers to high-level questions that crossed different functions like supply chain, e-commerce and marketing required pulling data together from different systems in different countries, which was too time-consuming," says Steell. "We knew we needed to centralize data and have it managed by one group, so we could get an integrated view of operations and customers and use that instead of intuition to drive our decisions."

> ***"We're very pleased with Talend. It helps us serve up information to our users faster and do it at a lower cost."***
>
> *– Matt Steell, director of Information and Integration Architecture, Office Depot Europe*

Office Depot's approach to solving this problem was to build an integrated data hub. "Our Enterprise Data Hub, powered by Talend and Hadoop, has transitioned Office Depot from being application-centric to truly data driven; senior managers through to business analysts have access to the right information at the right time in order to make better business decisions," Steell says.

Office Depot is also using Talend Data Quality to perform checks and quality control on data before ingesting it into the Hub's data lake. "We have a data quality framework for all data flowing through the data hub," says Steell. "We want to make sure we have quality data that can be trusted if it's going to be used for predictive analytics and business critical decisions."

Among the benefits Office Depot is receiving from Talend and the new data hub architecture are increased efficiency and reduced costs.

Healthcare: AstraZeneca

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

AstraZeneca had data dispersed throughout the organization in a wide range of sources and repositories. Having to draw data from CRM, HR, finance systems and several different versions of SAP ERP systems slowed down vital reporting and analysis projects.

"To be able to easily analyze that data, we knew we needed to put in place an architecture that could help with a mass consolidation and bring data together in a single source of the truth," says Simon Bradford, senior data and analytics engineer at AstraZeneca.:We wanted to consolidate everything and get a single set of global metrics so we could monitor activity across divisions and markets and do comparisons that were not previously possible."

"We started using the data lake and, each week another manager will come in with a new set of business questions. We've only scratched the surface."

— Simon Bradford, senior data and analytics engineer, AstraZeneca

AstraZeneca resolved to build a data lake on AWS to hold the data from its wide range of source systems. To capture that data, they selected Talend. Andy McPhee, science and enabling units data and analytics engineering lead, explains: "Talend is responsible for lifting, shifting, transforming and delivering our data into the cloud, extracting from multiple sources and then pushing that data into Amazon S3. The Talend jobs are built and then executed in AWS Elastic Beanstalk. After some transformation work, Talend then bulk loads that into Amazon Redshift for the analytics. Talend is also being used to connect to AWS Aurora."

The data lake that Astrazeneca has built shows the value of a data integration strategy built on a reusable infrastructure. "The data lake enables us to pull large volumes of valuable data from disparate systems and make our data discoverable across divisions," says McPhee.

Charitable organization: Save the Children UK

Save the Children UK (SCUK) saves lives by preparing for, and responding to, humanitarian emergencies caused by natural disasters, disease outbreaks and armed conflict. The organization works in more than 120 countries to reach breakthroughs in the way the world treats children and brings about immediate and lasting change in their lives.

One key data management challenge connected to operational efficiency was trying to meet KPIs for cleansing and loading data from 50-plus streams into their CARE NG CRM. "Prior to the introduction of Talend, the import team was under increasing pressure to deal with the growing volumes of complex data imports in line with our SLAs," says Penny Kenyon, CRM import and integrity manager, Save the Children UK. "We knew we needed to find a more-efficient way to manage our data streams that would improve the quality of incoming data and better identify potential duplicate records."

"Talend rose to the top for several reasons," says Kenyon. "It was more affordable, and, as a charity, being cost-efficient is a priority with us. We also thought Talend was outstanding from an ease-of-use standpoint."

The efficient importing of higher-quality data supports the production of better, faster reporting by analysts and provides SCUK greater insight into donor behavior and motivations. Kenyon says SCUK now needs only half a person-day to load data, versus the three person-days it used to take before Talend was implemented.

> ***"Talend was more affordable, and, as a charity, being cost-efficient is a priority with us. We also thought Talend was outstanding from an ease-of-use standpoint."***
>
> *– Penny Kenyon, CRM import and integrity manager, Save the Children UK*

CHAPTER 10
DATA INTEGRATION CHECKLIST

talend

DATA INTEGRATION CHECKLIST

The need for data integration tools exists in every company, small to large. Whether it is extracting data that exists in spreadsheets, packaged applications, databases, sensor networks or social media feeds, there is a significant benefit to share and reuse information instead of having duplicate processes and silos of information. It is also important to select a solution that can address all your data integration needs, whether it be data integration, data migration, big data integration, data warehouse integration, or integration with business intelligence systems.

This checklist provides key functional requirements for implementing and deploying data integration in an enterprise environment. Use the list to validate and prioritize your needs.

	Included	Description
Connect and deliver		
Connect to Traditional Data Sources		Connect to data stored in relational databases, OLAP applications, non-relational structures like flat files, XML, common packaged applications like SAP, cloud-based applications such as salesforce.com, semi-structured (e.g Excel) data, unstructured (e.g. audio, video) data, and messaging systems. Support for industry standards like EDI.
Connect to Big Data and NoSQL		Integration with big data technologies (e.g. Hadoop, Hbase, Hive), Big Data platforms (e.g. Cloudera, Hortonworks, MapR) and NoSQL databases (e.g. MongoDB, Cassandra)
Data Movement		The ability for data consumers to receive data in many ways. Support bulk data movement, data services, data federation, change data capture (CDC) and direct data replication between data sources.
Data Synchronization		Support Extract, Transform, and Load (ETL), and Extract, Load, and Transform (ELT), real-time delivery, and event-driven delivery (trigger or changed data).

	Included	Description
Transformation		
Simple Transformations		Such as calculations, data type conversions, string manipulations, aggregations, automatic lookup and replace operations.
Advanced Transformations		Such as slowly changing dimensions, normalization of data, advanced parsing capabilities and transformation to complex standards (for example, EDIFACT, HL7, and others)
Custom Transformations		Ability to create new custom transformations, as well as extend existing transformations.
Enrichment		Solution should have the capability to use enrichment data from a wide variety of sources. Enrichment data might come in various file formats and schemas both internal and external. It may come from online sources through service APIs, commercial partners or data providers.
Development and data modeling		
Graphical Tooling		Easy-to-use, graphical, drag-and-drop tools to build processes and transformations, and design data models, metadata and data flows. Graphical representation of objects and connectors. Wizards to automate common tasks.
Business Model Tooling		A non-technical tool that enables collaboration between technical and business users to structure all relevant documentation and technical elements supporting the data integration process.
Data Model Creation and Management		Ability to create and maintain data models. Use graphical tools to define relationships.

	Included	Description
Development and data modeling		
Metadata Management		Provides automated discovery of metadata. Ability to search metadata across multiple sources and show its lineage. Use a single repository of metadata across all product features, with the ability to seamlessly share and synchronize metadata between data integration tools and other tools (e.g. data quality, data profiling and master data management).
Business Rules and Workflow		The ability to define and manage business rules and execution flows. Process execution can be scheduled immediately, at a set time, or based on an event.
Versioning		Developers can easily version metadata, routines, processes, transformations or any other object used in the integration process. Then have the ability to see changes and roll-back to a prior version if necessary.
Collaboration		A set of tools for each user, i.e. business users, developers, and IT operations staff; and a shared repository consolidating all project information and enterprise metadata shared by all stakeholders.
Testing, Debugging and Tuning		Tools to test processes with data in the graphical tool, then interactively debug and tune for optimum performance.
Impact Analysis		Uses graphical tools to compare processes, assess the impact of change and view data lineage to see where changes occurred.
Standards Support		To facilitate ramp-up time and leverage existing resources, products should embrace standards such as Eclipse, Java, JDBC, ODBC, and Web services.
Reusability		Should be able to reuse projects, metadata processes, cleansing, validation, enrichment and other highly used routines in a fast and easy manner.
Customizable		Generated artifacts can be customized for maximum flexibility. Ability to create your own custom components. Easy to customize and extend transformations.

	Included	Description
Data governance		
Integration with data quality tools		Integrated functionality with tools that profile and cleanse data, parse and standardize data, and match, merge and identify duplicate records to then be rationalized based on your requirements. The ability to define business rules to be applied to data.
Integration with data profiling tools		Integrated functionality with tools that do column-based analyses, dependency analyses, trend analyses and custom analyses.
Integration with MDM tools		Integrated functionality or out-of-the-box integration with tools to create a unified view of information and manage that master view over time.
Reports and Dashboards		Pre-built and customizable reports that show key data quality metrics over time. Provide the ability to export results in a variety of formats including XML, PDF, HTML, etc. Provide a dashboard (web-based) reporting system of data quality metrics and provide metadata to business intelligence (BI) systems.
Deploy		
Multi-Platform Runtime Support		Ability to seamlessly deploy to Unix-based, Linux-based and Windows systems. Ability to run on-premises and in the Cloud and virtualization environments. Ability to run in big data (MapReduce) distributed processing environments. Ideally generates code for portability and performance.
Load Balancing and Scalability		Clustering capabilities to spread server load over several machines. The ability to handle very large data volumes, working with big data and multi-terabyte data warehouses.
Failover		Ability to rollback a transaction and continue processing if there is a server failure without losing data.
Remote Execution		Ability to run processes remotely on various operating systems using the same configuration

	Included	Description
Deploy		
Data Integration Services		Ability to deploy all aspects of run time functionality as services within a service-oriented architecture.
Middleware Compatibility		Integrated functionality with MOM and ESB systems
Hadoop Support		Deploy native MapReduce jobs directly to a cluster with no needed appliances or additional software installed on the cluster. Have the ability to scale MapReduce processes with the cluster without code changes.
Monitor and manage		
Centralized Administration		Ability to monitor and manage all resources and deployments from one location.
Web-based Monitoring		Ability to monitor resources and deployments from any browser.
Reports and Dashboards		Pre-built and customizable reports that show key data integration metrics. The dashboard shows information and statistics over time, e.g. performance, load volumes, subtask individual metrics such as database read and rights or enrichment service response times.
Exception Reporting and Management		Ability to define, report and handle exceptions when they occur. Capability to invoke special processes when violations to data integration rules. Examples include an email alert, text message, or halting a process.
Security Controls		A mechanism to secure in-flight messages between applications as well as user/role-based security in the tool itself, LDAP.
Business User Interaction		Solution should provide an easy-to-use environment for business users to follow the key performance indicators for data integration, e.g. PDF reports and web-based portals.
Cloud Support		Ability to setup, deploy, and shutdown a cloud instance, e.g. Amazon EC2. Enables you to expand your computing capacity for your integration processes.

	Included	Description
Support		
Comprehensive Support		Provides the support you need when you need it, e.g. community forums, web knowledge-base, email support, and phone support. 24x7 mission critical support. SLAs for response time, bug fixes and maintenance updates.
Training		Classroom, online, and on-demand training for newbies, advanced developers and administrators.
Professional Services		Vendor offers a complete spectrum of consultative services: assessment, strategy and architecture, quickstart, design and development, tuning, technical audit, and custom offerings.

CHAPTER 11
WHAT'S THE TAKEAWAY?

talend

WHAT'S THE TAKEAWAY?

Now's the time for your organization to realize the value of enterprise data, and to use data integration as a way to get business-critical insights and analysis. Building your data integration strategy with the right tools, the right team, and the right capabilities will make this happen.

Talend offers a comprehensive data management platform for achieving your goals.

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

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

Customer Support
Visit the Talend Community

Learn More
www.talend.com

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

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

talend