StreamSets
A SOFTWARE AG COMPANY

RESEARCH REPORT

Lifting the Lid on the Hidden Data Integration Problem

Rethinking Data Integration for the Modern Enterprise

streamsets.com

Introduction

The Data Dependency Pressure Cooker

Data is at the very core of digital transformation. Businesses are dependent on insights from data to meet strategic and operational goals. Without data, enterprises cannot make smart real-time decisions, stay competitive, or accelerate innovation. Data leaders and practitioners know that to meet business demand for digitalization, information assets must move seamlessly and at speed throughout an organization.

But this is easier said than done. The modern data ecosystem is enormous, complex, and dynamic. It's also constantly evolving as data architectures become increasingly fluid. Building pipelines that connect data from source to destination requires rules to integrate, transform, and process data across multiple environments. The data supply chain is not fixed from cloud applications and services to on-premises mainframe and legacy systems. All of which has made the job of building resilient data pipelines considerably harder.

Under-resourced technical teams are struggling to keep up with the volume of requests for data from the business without ceding control. And business teams simply want data on demand to inform their operations and advance their digitalization initiatives. The end result is frustration on both sides of the aisle.

StreamSets wanted to lift the lid on the hidden problem of data integration friction and find out what it means for today's modern enterprises. And who better to ask than those on the data front lines? We surveyed 653 data decision makers and practitioners from large enterprises in the US, UK, Germany, France, Spain, Italy and Australia to understand the challenges of delivering data to the business. In this report, we explore the results and shine a light on the burden data leaders and practitioners face.

What's inside:

- Demand for data is outstripping supply
- Data chaos is holding businesses back
- Beyond friction: cracks in the pipelines
- The true cost of data integration friction
- Unleash the power of data across the enterprise
- Methodology and demographics

Demand for Data is Outstripping Supply

Access to data is critical to every aspect of an organization’s digital and strategic objectives. Whether navigating turbulent economic headwinds and volatile supply chains, launching new products and services, or simply staying competitive, organizations require real-time data analytics fueled by large volumes of accurate and timely data.

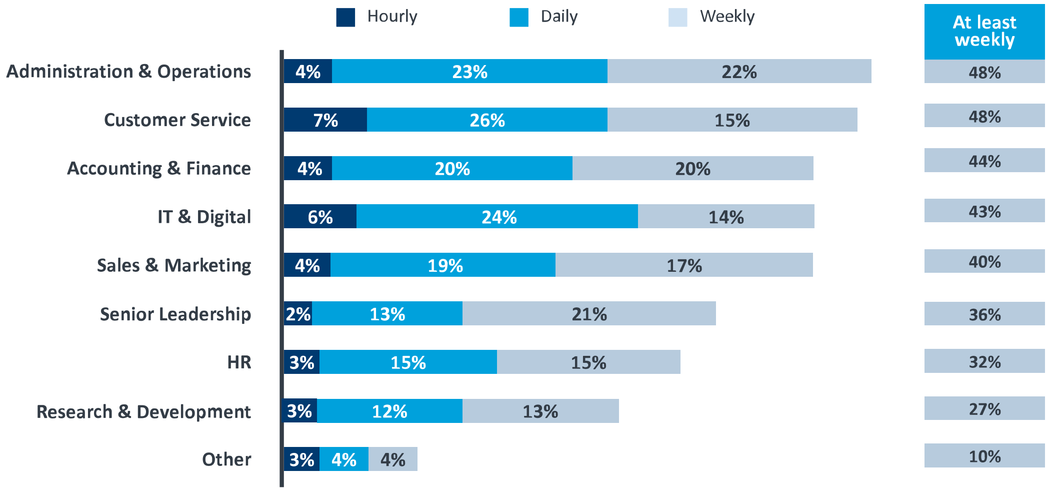
While traditionally technical teams would expect departments like finance and sales to request data frequently, this research shows that all lines of business are consuming more data as digital transformation continues at pace. Almost half (48%) of admin and operations, and customer service departments request data at least weekly. They are followed by accounting and financing (44%), other IT and digital teams (43%), and sales and marketing (40%).

The increase in requests means a classic supply and demand problem exists. The demand for data is higher than the ability of most technical teams to provide it. More than half (59%) of respondents say the acceleration of digital transformation priorities has created major data supply chain challenges.

The problem of meeting demand for data is compounded by the complexity of enterprises’ ecosystems. Data engineers must take many steps to connect, transform and process data to build pipelines that meet the individual needs of different departments. But when data is siloed in multiple systems with inconsistent formats, creating bespoke data pipelines at scale is a huge challenge. **Almost two thirds of respondents (65%) say this data complexity and friction can have a crippling impact on digital transformation.**

As a result, there is often a disconnect between the expectations of line of business teams and what can actually be delivered. Data leaders and practitioners are frustrated that non-experts expect data on demand and have little understanding of the scale of the data integration challenge.

Figure 1. Administration & Operations and Customer Service request data most frequently, with almost half (48%) requesting it at least weekly



DATA DEMAND:
The IT and Business Disconnect:

69% **are frustrated** non-data experts think you can click a button and data magically appears

59% say line of business owners are **“blissfully unaware”** of how hard it is for IT to deliver data

50% **are tired of people thinking cloud makes data access easier** when the opposite is true

The bottom line is that it is difficult to quickly fulfill diverse requests with scarce resources. Technical teams are under increased pressure to deliver data and support digital transformation. Meanwhile, the skilled employees needed to build smart data pipelines are in short supply. Data will continue to grow in volume, complexity, and urgency. If the data integration friction problem is left unresolved, the inability to empower teams with data will impact businesses' ability to thrive.

Data Chaos is Holding Businesses Back

In a bid to mitigate data integration friction, many businesses have invested heavily in "enabling" technologies to both increase agility and drive digital transformation. These include everything from moving to the cloud and implementing AI, to adopting elastic and hyper-scalable data platforms. However, keeping up with constant change introduced by technology is hard and can in fact add to the difficulties of data integration friction.

As data sources and technology platforms proliferate, businesses end up with a patchwork of systems where data becomes increasingly siloed. Whether legacy systems, point solutions, custom-built tools or solutions from a cloud service provider, the result is a fragmented and chaotic data environment.

As a result, what should be a simple pipeline-building task becomes a complex job requiring expensive expert skills. Inevitably, this hampers technical teams and slows them down. **The research finds that over two thirds (68%) of data leaders say data friction is preventing them from delivering data at the speed the business requests it.** And more than four-in-ten (43%) say data friction is a "chronic problem" in their organization.

There are several factors contributing to this friction. The most cited issue by respondents was the variety of data formats, both structured and unstructured (38%). This was followed by the speed that data is created at (36%) and the presence of legacy technologies (30%).

> For many, getting the data **"out" of legacy systems is the biggest obstacle,** but it's **also the biggest gain** when it comes to business insight.

Legacy Technologies

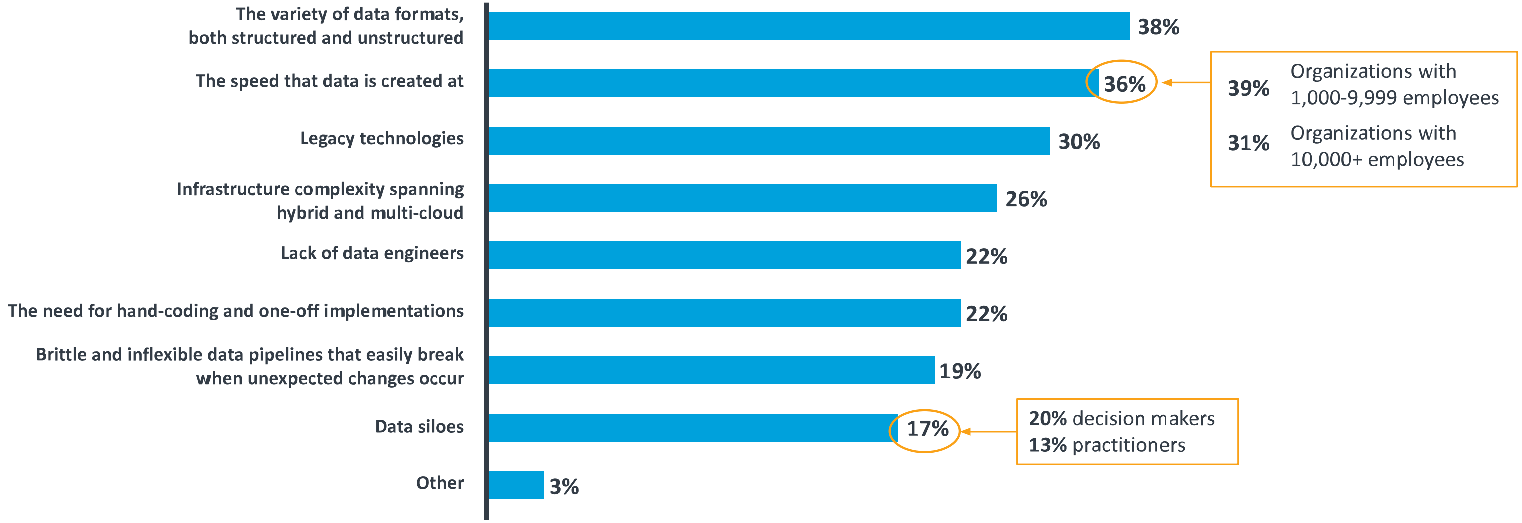
To further underline the point around legacy technologies, 51% of respondents say data in legacy systems, such as mainframes or on-premises databases, are hard to access for cloud analytics, so they often "don't bother" to include it when creating data pipelines — a considerable risk.

Legacy systems typically have many decades of valuable business insights held within them. This statistic also highlights that cloud analytics are not a panacea. Many analytics products focused on sourcing data from SaaS applications are unable to extract data completely from complex multi and hybrid cloud environments, let alone data trapped in legacy systems.

No modern enterprise today can afford to simply ignore legacy data, especially because legacy systems typically hold years of proprietary data and much of a company's IP. This data is the "secret sauce" that gives a business the edge. This specific, transactional and granular data ensures the insights from machine learning and AI models drive optimal business decisions. For many, getting the data "out" of legacy systems is the biggest obstacle, but it's also the biggest gain when it comes to business insight.

Without the assurance that data from all sources is collated, businesses cannot fully trust their data. And in today's world, data must come without caveats. However chaotic the data ecosystem, technical teams need the capacity to run dynamic data pipelines in any cloud or on-premises environment to unlock insights that drive innovation.

Figure 2. The biggest causes of data friction are the variety of data formats (38%) and the speed that data is created at (36%)



Beyond Friction: Cracks in the Pipelines

As we've discussed, the chaos of modern data ecosystems means building smart and resilient data pipelines is hugely difficult. For many businesses, establishing a pipeline is labor intensive and requires expert data engineers to hand-code one-off solutions that can't be templatized or re-used. These pipelines are not automatically insulated from unexpected shifts in the environment, resulting in brittle pipelines that are vulnerable to breakage.

The research finds that 39% of data leaders and practitioners admit their pipelines are too brittle and crack at the first bump in the road. A noteworthy 87% of respondents have experienced data pipeline breaks at least once a year, with more than a third (36%) saying their pipelines break every week, and worryingly, 14% say they break at least once a day.

Considering that large enterprises can have thousands of critical data pipelines in place, this represents a huge amount of disruption. It leaves line of business teams working with outdated information and technical teams dealing with a mountain of repair work. The business impact of broken data pipelines can be significant. For example, a supply chain director working with old data may over or under order goods. Customer facing teams in a haulage and logistics company cannot accurately inform customers of when to expect their orders. And a trader in a financial services firm may be left to make stock picks on out-of-date intel.

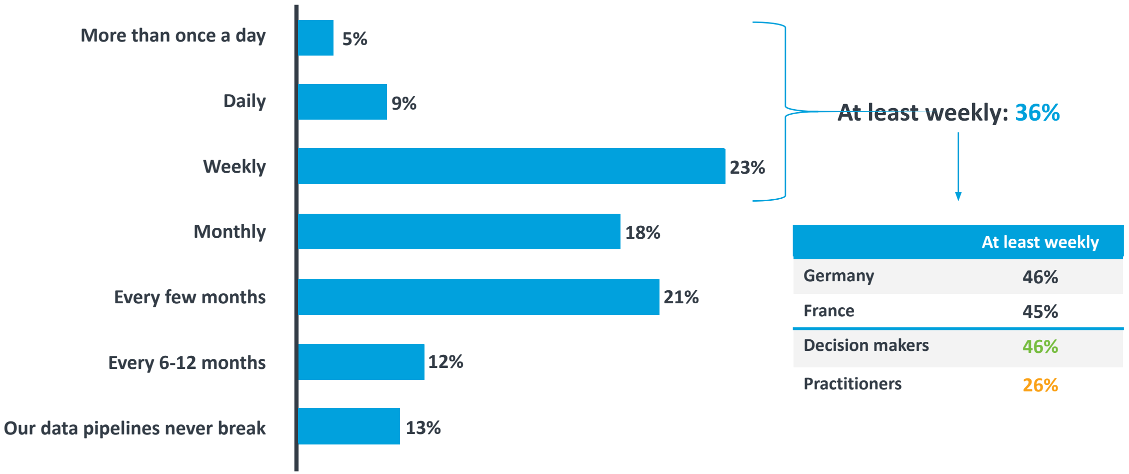
Pipelines break when they are not resilient to changes in the environment. The most cited reasons for breakage by data leaders and practitioners in our research include bugs and errors being introduced during a change (44%), infrastructure changes such as moving to a new cloud (33%), and credentials changing or expiring (31%).

It is not surprising to see cloud so high on this list. Migrating to the cloud can cause as many problems as it can solve if businesses don't have a clear data strategy for multi and hybrid cloud environments. When companies opt for a basic "lift-and-shift" approach to a cloud move, technical teams are forced to carry out extensive rework to orchestrate systems and connect them to data pipelines.

Without this remedial work, every change to how data is stored and consumed heightens the risk of breaking the data flow. But this data drift – the unexpected changes to data structure, semantics, and infrastructures – is a fact of modern data architectures. When businesses can't evolve their data architecture in tandem with the infrastructure and platform choices made, sub-optimal data pipelines are the result.

Businesses need to be able to introduce changes without having to worry about the stability of their data pipelines. They need the capability to ingest more data without needing to build more infrastructure. When they can do this, data engineers are freed up to complete high-value work, empowering line of business teams with data that allows them to innovate. But the research shows that today, many technical teams are not equipped with the right tools to satisfy this need. Almost half (46%) say their ability to tackle broken data pipelines lags behind other areas of data engineering, and 43% say they struggle to fix data pipelines in motion.

Figure 3. More than 1 in 3 (36%) say their data pipelines break at least weekly



	At least weekly
Germany	46%
France	45%
Decision makers	46%
Practitioners	26%

The Human Impact of Data Integration Friction:
Stop Apologizing For Your Data

Having trust in the data you are using is essential. Forecasting P&L, predicting sales, analyzing marketing campaigns, reporting financial results to the board...imagine being able to complete this work with *complete confidence* in the data.

No one wants to have to apologize for their data. Or justify why the last quarter's figures aren't included. Or caveat their datasets with explanations of why it's not quite up to date.

It makes recommendations less powerful and can even damage a department's reputation.

Yesterday's data is not the same as today's. Businesses must be powered by resilient data pipelines that automatically ingest the most up to date information and serve it on demand, wherever it is needed. This gives data users complete confidence in the trustworthiness and accuracy of data, so they can stop apologizing for it.

The True Cost of Data Integration Friction

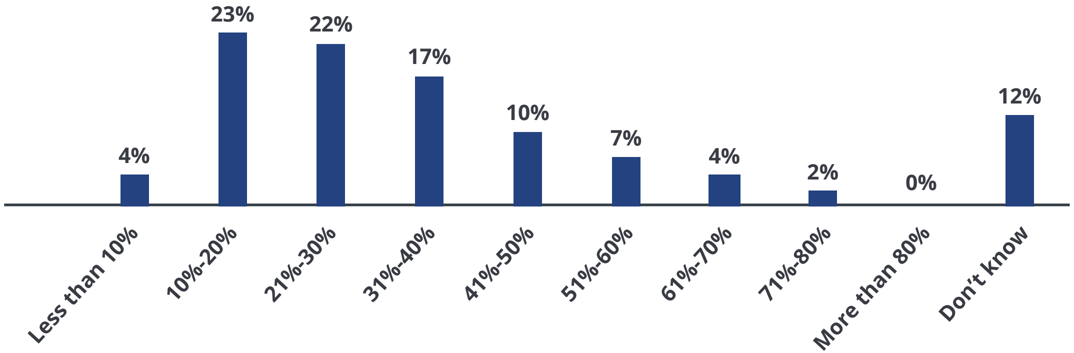
The inability to consistently build flexible and resilient data pipelines has major ramifications for enterprises. It forces already under-resourced technical teams into firefighting mode as they scramble to repair pipelines. Respondents said that, on average, data engineers spend 31% of their time troubleshooting and recoding broken data pipelines. As everyone in business knows, time is money. And this time spent firefighting swiftly adds up. When you consider that businesses spend $6.13 million annually on data experts, repairing data pipelines equates to $1.9 million of their time per year.

$6.13M **Average yearly spend on data experts**

$1.9M **Average yearly cost of fixing broken data pipelines**

Given that expert data talent is both scarce and expensive, enterprises need to find ways to automate the creation of resilient data pipelines and eliminate the need for hand-coding and one-off solutions. Because it's not just the wasted spend that businesses should be concerned with, but also the lost opportunity cost. Data engineers are too valuable to spend large amounts of time troubleshooting. These highly trained individuals should be free to innovate, create, and help drive the business to achieve its strategic goals.

Figure 4. On average, data engineers spend 31% of their time troubleshooting and recoding broken data pipelines



Unleash the Power of Data Across the Enterprise

Data is a critical success factor in the modern enterprise. It drives digital transformation, experimentation and prototyping, and real-time analytics to keep businesses competitive and thriving. But as the results of this research have demonstrated, data integration friction is holding technical teams back and stopping them from keeping up with "need-it-now" business demands. Data leaders are aware of the scale of the problem, with 70% of respondents believing that smarter data pipelines would enable them to deliver data to the business at pace.

Instead, the complex and brittle architectures that most businesses are working with are driving up costs and impeding agility. This leaves enterprises struggling to meet the demand for data and empower all lines of business with data insights that accelerate innovation and real-time decision making.

Businesses can reduce the load on already stretched technical data teams by enabling individual business units and end-users to do more "last mile" data collection and analysis themselves. The clear majority of respondents agree; 70% of data decision makers and practitioners today are currently responsible for the last mile of data delivery, but 86% would prefer line of business teams be empowered to do this independently.

CASE STUDY

(Formerly GlaxoSmithKline plc)

Powering Drug Discovery with Self Service Data

SCENARIO

10,000+

scientists consuming data

6Pb

of stored data

1,000+

data sources

SOLUTION

Bringing a new drug to market costs billions of dollars and can take up to 20 years. At every phase of drug discovery research, from compound investigation to post-market monitoring, data is critical. Scientists require multidisciplinary data that is accurate, relevant and trusted, to inform their research. GSK wanted to give its 10,000+ scientists engaged in R&D around the world access to such data. To achieve this, it needed to de-silo its data sources and deliver it in a single enterprise-wide platform that allowed individual teams to consume data-on-demand. GSK worked with StreamSets to bring this vision to fruition, building a Data Center of Excellence to accelerate delivery of clean data from 1,000s of data sources. Using StreamSets, GSK has automated data pipeline creation and data drift handling without interrupting the critical flow of self-service data for scientists.

RESULT

- Onboarding time for new data sources was **reduced by 98%.**
- New product discovery time was **reduced by 96%.**
- Accelerated time to market for new drugs **by almost 3 years.**

StreamSets
A SOFTWARE AG COMPANY

Demographics

Total respondents: 653

Country of residence

🇺🇸	🇬🇧	🇩🇪	🇫🇷	🇪🇸	🇮🇹	🇦🇺
200	100	102	100	51	50	50

Role type

Decision makers: 50% | Practitioners: 50%

Role (IT/ LOB)

IT: 73% | Line of Business: 27%

Size of company

# of employees	1,000-4,999	5,000-9,999	10,000+
% of respondents	35%	26%	40%

Business sector

Company Sectors Top 3: Financial Services: 21% | IT: 12% | Local authority & Public sector: 11%

Gender & age

- **52%** of respondents were female and **48%** male
- **57%** of respondents were aged between 25-44

Methodology and demographics

The survey was commissioned by StreamSets; it was conducted among 653 decision makers for data tools and practitioners who use data tools in the UK, US, Germany, France, Spain, Italy and Australia. The interviews were conducted online by Sapio Research using an email invitation and an online survey.

StreamSets
A SOFTWARE AG COMPANY

About StreamSets

StreamSets, a Software AG company, eliminates data integration friction in complex hybrid and multi-cloud environments to keep pace with need-it-now business data demands. Our platform lets data teams unlock data—without ceding control—to enable a data-driven enterprise. Resilient and repeatable pipelines deliver analytics-ready data that improve real-time decision-making and reduce the costs and risks associated with data flow across an organization. That's why the largest companies in the world trust StreamSets to power millions of data pipelines for modern analytics, data science, smart applications, and hybrid integration.

streamsets.com