StreamSets
A SOFTWARE AG COMPANY

RESEARCH REPORT

Creating Order from Chaos: Governance in the Data Wild West

Understanding, Governing, and Managing Data Pipelines in the Modern Enterprise

streamsets.com

© 2023, STREAMSETS, INC. ALL RIGHTS RESERVED

Introduction

Taking Control of the Data "Wild West"

Modern data infrastructures are chaotic. Businesses operate in a data "wild west" of complex architectures spanning hybrid and multi-cloud environments, where a patchwork of legacy systems, point solutions, and custom-built tools proliferate. This leaves data decision-makers struggling to understand, govern, and manage a fragmented data supply chain.

At the same time, line of business teams have become increasingly savvy about using data to inform their operations. These teams draw on their domain expertise to create value for the business by working independently to build their own data pipeline integrations. However, this often means pipelines are built outside the purview of IT, creating new data silos and more visibility gaps. Ultimately, businesses can't control what they can't see.

When data ecosystems are governed well, and IT teams have visibility into pipeline creation, the way tools are used to create, process, store, analyze, and share data can be standardized. Standardization allows businesses to meet governance policies as data is transported, processed, and distributed through data pipelines. But many companies today lack the ability to implement and enforce such controls.

When consistent measures to safeguard data are absent, businesses increase the risk of reputational damage, regulatory fines, and security breaches from data leakage. Moreover, without governance policies that establish data uniformity, organizations can't unlock the value of their information assets and be confident that they are making decisions based on reliable data.

To shine a light on the issues modern enterprises face in managing and governing data pipelines and operations, we went directly to the people attempting to wrangle order from the chaotic frontier of their data ecosystems. 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 data governance today. In this report, we dig into the results.

What's inside:

- The challenge of governance in the data wild west
- Confusion over responsibility adds to data visibility challenges
- Blind spots are created when business teams move fast
- Establishing a data "mission control" to create order from chaos

StreamSets
A SOFTWARE AG COMPANY

© 2023, STREAMSETS, INC. ALL RIGHTS RESERVED

The Challenge of Governance in the Data Wild West

As modern data ecosystems have become increasingly diverse, there has been some benefit for businesses. When analytics tools and projects are decentralized and put in the hands of business users, some advanced teams can increase their agility and ability to innovate. But this comes at a cost by creating governance challenges and risks for the whole enterprise. And data leaders and practitioners know it.

More than half (54%) say modern infrastructures that span on-premises and multiple cloud environments, combined with data decentralization between line of business teams, has created a data "wild west." Further, 57% say this fragmentation in the data supply chain has made it harder to understand, govern, and manage data in their organization.

Life on the data frontier

- **54%** say that modern infrastructures and data decentralization have created a data "wild west"
- **57%** say data fragmentation makes it hard to understand, govern, and manage data
- **81%** want consistent security measures to protect data as it flows between on-premises and cloud sources

Connecting many data sources and ensuring uniformity between definitions, formatting, and metadata is hugely complex. It makes initial deployments and later updates a painful management headache for data teams. They need a better way to enforce consistent rules to govern the increasing number of apps, systems, data sources, and tools that get integrated into the ecosystem.

The data leaders and practitioners in our research agree, with 81% saying they want consistent security measures to protect data as it flows between on-premises and cloud sources. Without consistency, visibility and control are lost, significantly increasing the risk of data breaches and resulting in fines.

Five Critical Pillars of Data Governance

Good data governance requires a well-defined strategy. Here are five fundamentals to consider when developing yours.

Identify your data: To design an effective strategy, you need to know your entire data landscape inside out, including types, structures, movements, locations, and points of data transformation.

Establish a governing body: The data governance body is a central control point around which all teams and departments can agree on consistent policies that align with business goals.

Ensure "privacy by design": A privacy-first approach is central to good data governance. It involves collecting only necessary data, masking personally identifiable information (PII), and using data only for intended purposes.

Metadata management: Properly managing metadata makes it easier to track data changes, control data access, and understand relationships between data to fulfill governance requirements.

Data quality management: An effective data governance strategy will establish consistent criteria and scoring to ensure data is high quality and reliable for use in analytics and AI/ML applications.

Once you've designed your strategy following these pillars, you are ready to implement it. Check out this blog to learn how.

Confusion Over Responsibility Adds to the Data Visibility Challenge

One of the factors stopping organizations from embedding consistent data governance is confusion over who is responsible for managing data. This is compounded by the fact that different business stakeholders have different priorities when it comes to pipeline building. Data analysts and line of business teams prioritize the speed of data generation and dissemination. Data leaders often place more focus on compliance and control.

The research results underline this challenge. For nearly half of the businesses (47%), the primary

© 2023, STREAMSETS, INC. ALL RIGHTS RESERVED

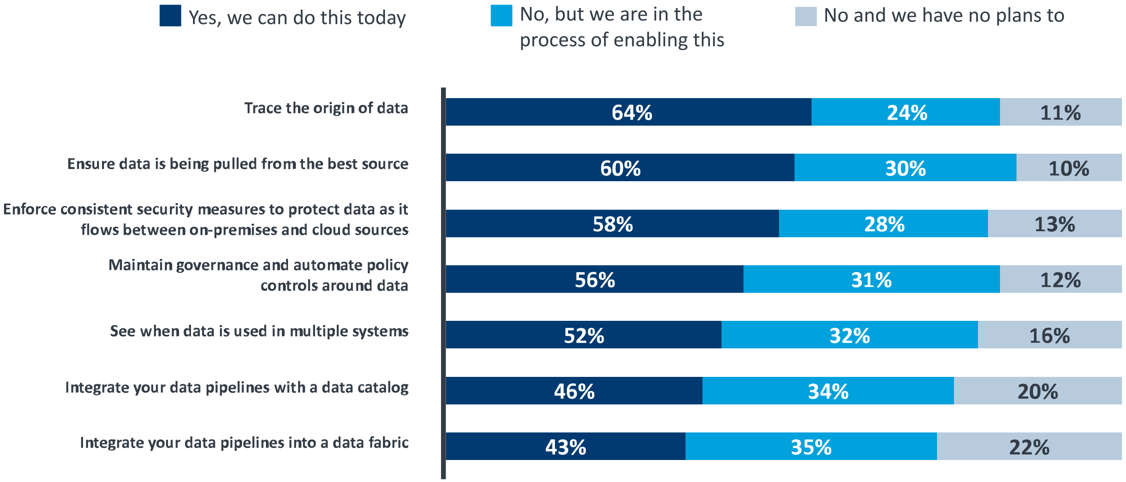
responsibility for managing data sits with the central IT team. However, almost a fifth (18%) say it sits with line of business teams. And more than a third (35%) say the responsibility is split between line of business teams and IT.

Despite this, and contrary to earlier findings that show companies feel they are operating in a data "wild west," 71% of data leaders and practitioners say they are confident they have complete visibility and control over their data. But when we dig deeper into the results, gaps are clear, indicating that many businesses are harboring a ticking time bomb of governance risks they are unprepared for.

The survey finds that 44% of organizations cannot maintain governance and automate policy controls around data, and 42% cannot enforce consistent security measures — a clear vulnerability. Regulatory requirements restrict access to certain types of data, like personally identifiable information (PII), to certain users in specific use cases. An inability to automate policy controls increases the chance that employees who are not compliance experts or are not permitted to view certain data may inadvertently violate regulations.

Similarly, the lack of visibility into where data flows to and from raises the likelihood of a breach. The research reveals that 48% of businesses can't see when data is being used in multiple systems, and 40% cannot ensure data is being pulled from the best source. Moreover, 54% cannot integrate pipelines with a data catalog, and 57% cannot integrate pipelines into a data fabric.

Figure 1. More than half of respondents are currently unable to integrate their data pipelines with a data catalogue (54%) or into a data fabric (57%)



Undoubtedly, ensuring that the many considerations around data access, use, and storage are maintained across all pipelines is challenging. Especially when we consider that many large enterprises can have thousands of data pipelines in operation. But it's clear that businesses must regain control, or they risk a costly and damaging breach.

© 2023, STREAMSETS, INC. ALL RIGHTS RESERVED

CASE STUDY

Ad Astra

Delivering Client Results in Weeks Not Years

Ad Astra is higher education's solution partner in managing the academic enterprise. To help bring schools and students together, Ad Astra pulls large volumes of data from multiple student information systems into a single ERP. Unfortunately, creating their own system meant development teams had to focus primarily on major student information systems (representing 80% of Ad Astra's clients) instead of all systems simultaneously. This left around 20% of Ad Astra's customers unaccounted for and pushed Ad Astra into the data ingestion business — not core to their mission of providing software to higher education.

With StreamSets, Ad Astra found an efficient means to ingest data, simplifying their overall data operations, and reducing the burden placed on their development teams. It has reduced maintenance on dozens of separate XML files down to four basic pipelines that run hundreds of jobs for their clients.

Implementation time for new clients has dropped from months to days, and their ability to start gathering, analyzing, and getting recommendations/results back to clients has been cut from being almost as long as a year to just a few weeks. By implementing StreamSets, Ad Astra has not only become more agile and increased customer value, but they confidently deliver on their top priority: to help students graduate faster.

© 2023, STREAMSETS, INC. ALL RIGHTS RESERVED

Blind Spots Are Created When Business Teams Move Fast

The democratization and decentralization of data, as discussed, has delivered benefits to business. When line of business teams can access and handle data through a governance framework, they can innovate and experiment safely within centralized guardrails. But without those guardrails, blind spots are inadvertently created, especially when business teams are moving fast and developing new integrations and pipelines without IT's oversight.

More than two in three (68%) data leaders and practitioners say that line of business teams and users independently create datasets without telling IT or data teams. Almost all respondents (95%) say this creates problems and business risks.

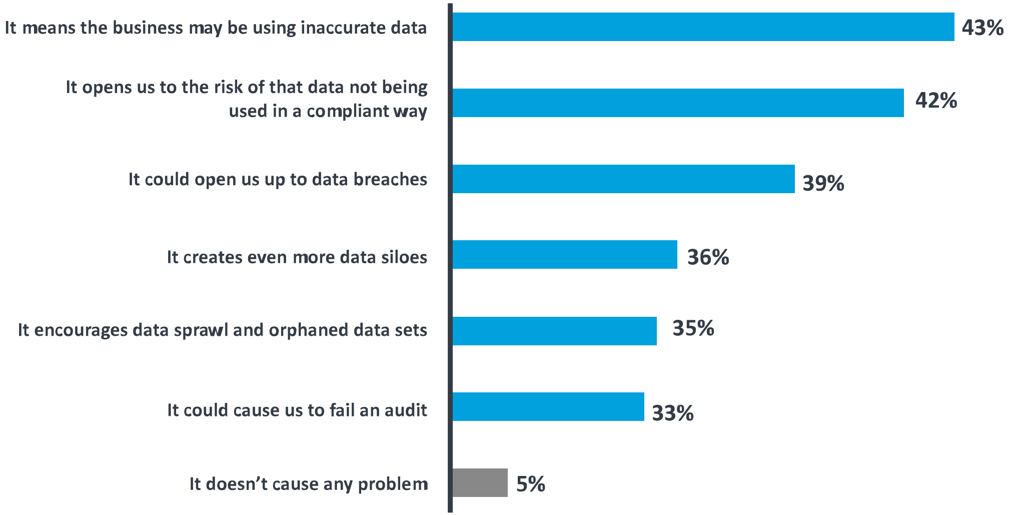
The most cited risk stemming from data users operating outside of IT's field of vision is that it causes the organization to use inaccurate data (43% of respondents). This was followed by opening the business up to the risk of data being used in a non-compliant way (42%) and data breaches (39%).

The creation of data silos was also identified as a challenge (36%), as were the spread of data sprawl and orphaned datasets (35%). In fact, orphaned datasets that users have created and forgotten about were identified by more than half of respondents (56%) as one of the most significant risks to their business.

Organizations must strike a balance between control and enabling transformation. Everyone in the business wants access to more data. So those with the keys to the data kingdom must give users the opportunity for innovation, prototyping, and experimentation — but they must do so safely. Data leaders and practitioners want to enable this way of working. The research finds that 72% want to empower line of business teams to use data while maintaining control, and 80% want to enable a self-service data model for end-users. But given what we've learned about the lack of visibility businesses currently have over their data ecosystem, this is not possible in many instances.

To empower business users to independently access and utilize data while ensuring good governance policies are in place that reduce the organization's exposure, IT teams need transparency. This means seeing how systems are connected and how data flows across the enterprise, including where new integrations or pipelines are built, when, and by whom. And crucially, it will shine a light on blind spots that cause risk to the business while unleashing line of business teams to innovate rapidly.

Figure 2. LoB teams creating their own datasets without notifying the data / IT team creates problems for 95% of respondents, the biggest being the possibility of using inaccurate data (43%)



© 2023, STREAMSETS, INC. ALL RIGHTS RESERVED

Establishing a Data “Mission Control” to Create Order from Chaos

To achieve the level of visibility needed to empower safe innovation, data leaders and practitioners must be armed with a centralized management console that acts as a data “mission control.” And given the chaos of modern data ecosystems, any solution needs the capability to “see” across all environments.

End-to-end visibility is crucial because every environment has unique deployment and governance challenges. Data teams must ensure that cloud-based applications can securely leverage data from a different environment and vice versa. Respondents in our survey agree, with 74% saying a single platform that can handle the complexity of data spanning across cloud and on-premises worlds would be a huge benefit.

Data leaders and practitioners who can establish a centralized data console that embeds good governance throughout the organization will ensure line of business teams can extract maximum value from data. At the same time, they will lower the costs and reduce the headache of managing a fragmented data supply chain.

StreamSets helps businesses create order from chaos. Our single, fully managed, end-to-end platform becomes an organization’s mission control. Sophisticated topologies deliver deep visibility into how systems are connected, allowing data leaders and practitioners to see how data flows across the enterprise.

With StreamSets you can be certain policies and procedures governing how data is created, processed, and distributed are in place throughout the entire data lifecycle, ensuring access to reliable data and complying with privacy and data safety laws at all times. With StreamSets, businesses have centralized guardrails that allow line of business users to explore the art of the possible, with the confidence that data is compliant and secure.

© 2023, STREAMSETS, INC. ALL RIGHTS RESERVED

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 in December 2022 using an email invitation and an online survey.

© 2023, STREAMSETS, INC. ALL RIGHTS RESERVED

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

© 2023, STREAMSETS, INC. ALL RIGHTS RESERVED