The Chief Data Officer's guide to digital transformation

How to push your business's data operations to the forefront of innovation and data-driven value

Fivetran

Table of contents

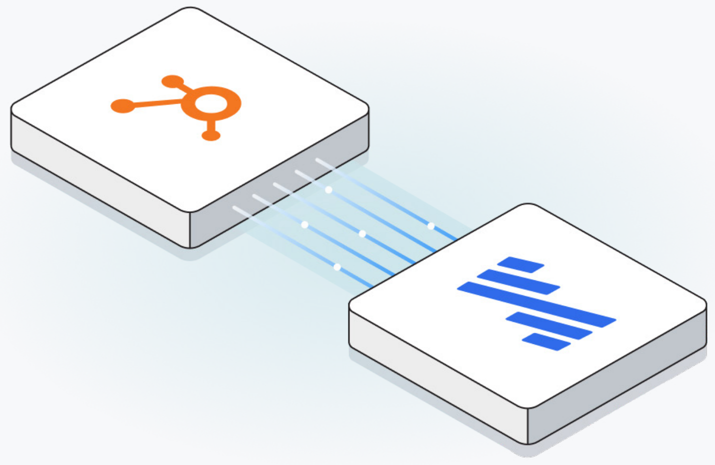
Executive summary

Chief Data Officers (CDOs) have a mandate to bring digital transformation to their organizations. In the context of data operations, digital transformation means moving from crude or reactive uses of data, such as descriptive analytics, to higher-value, more sophisticated and innovative uses of data, such as predictive analytics.

Chief Data Officers face a number of well-known challenges such as institutional inertia and a lack of clear expectations. Moreover, CDOs must contend with the following:

- Teams must make high-stakes decisions that are best supported with data rather than educated guesswork. The more agile the company, the better, but decentralizing decision making carries its own risks. To compound these difficulties, teams are often partly or wholly remote.
- Competition and macroeconomic headwinds mean that organizations must always find ways to do more with less in terms of headcount and budget.
- As the volume of data grows, so does the complexity of its handling. Governments write increasingly stringent regulations, while consumers simultaneously expect greater personalization and stronger security.

These challenges contribute to the notoriously high turnover of CDOs, with an average tenure of fewer than 30 months. The clock for achieving transformational change is always ticking. The goal of this guide is to help CDOs identify critical junctures on their roadmaps to bring their organizations to the forefront of data maturity and innovation, and in the process beat the clock.



There are several stages that a roadmap for digital transformation must progress through. These four stages – data centralization, infrastructure modernization, data democratization and building data solutions – not only constitute a progression but also represent use cases that have a common technological solution.

This solution is the automated data movement platform.

1. **Data centralization** is the classical data integration problem. As SaaS applications for every business operation proliferate, organizations must consolidate data from a growing variety of sources onto a platform that decision makers can easily and performantly access. Data on a single, centralized platform produces a full view of an organization's operations, customers and products, enabling analytics and all further uses of data. Unlike federated analytics, it can also perform at very large scales. However, it is a deceptively complex engineering challenge that is best solved with the use of a fully managed, automated data movement platform.

2. Once data is centralized, **data democratization** brings accessibility and governance to stakeholders and decision makers across an organization. Distributing access to data and decision-making authority enables greater organizational agility and better responsiveness to dynamic, changing markets. However, exposing data to a larger audience also carries security risks that must be addressed using security features and governance tools.

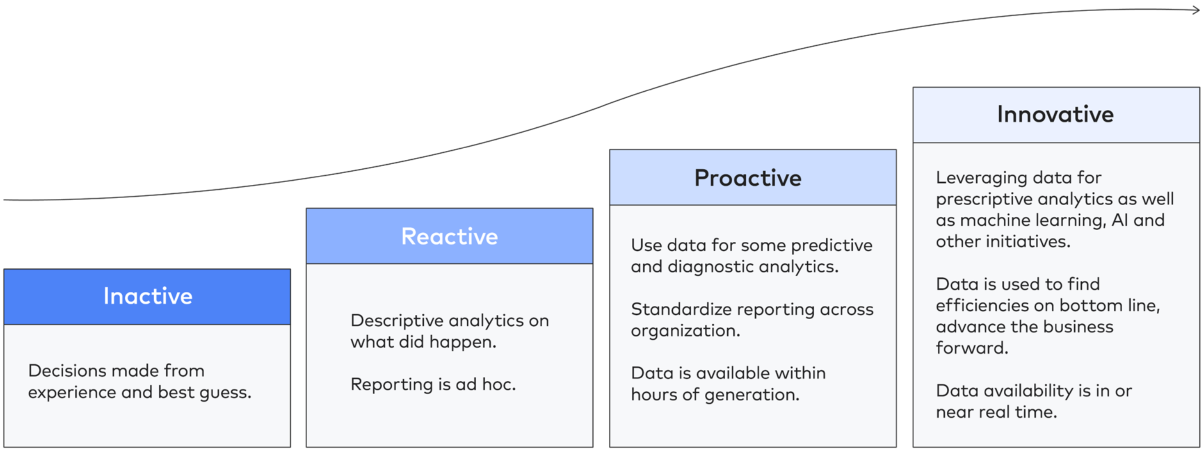
3. **Building data solutions** consists of finding ways to share data internally or with third parties, turn it into products or otherwise monetize it. Raw data, insights and systems fueled by data such as predictive models or artificially intelligent agents are all valuable commodities. Such pursuits require data sharing and extensibility features.

As an organization progresses through the stages listed above, it will also be incentivized to **modernize its infrastructure**, adding and removing tools and platforms to meet emerging needs. A common example is moving from on-premise to cloud or hybrid infrastructure. This is an incremental, continuous progress. Ripping and replacing is rarely practical because of existing commitments.

All roadmaps to digital transformation must begin with data centralization. From there, data democratization is a logical next step. Mastery over data centralization also enables an organization to monetize data by building data solutions. All throughout, an organization should continuously evolve its data infrastructure in response to changing needs.

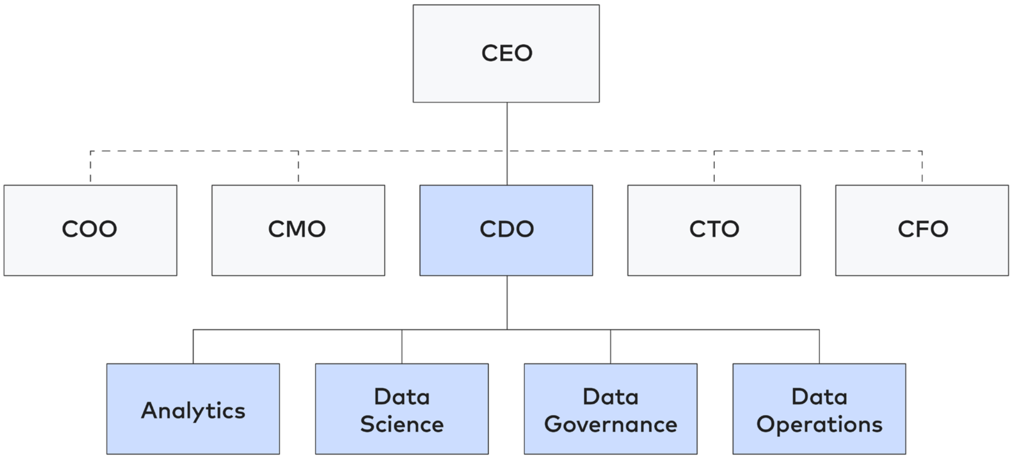
A structured approach to data

Data is the most essential resource for organizations of all kinds, used to support decisions, automate processes and power innovative products. Its value increases with freshness and scale. You can mentally model the use of data as a progression of four steps called the data maturity curve:


Inactive
Decisions made from experience and best guess.
Reactive
Descriptive analytics on what did happen.
Reporting is ad hoc.
Proactive
Use data for some predictive and diagnostic analytics.
Standardize reporting across organization.
Data is available within hours of generation.
Innovative
Leveraging data for prescriptive analytics as well as machine learning, AI and other initiatives.
Data is used to find efficiencies on bottom line, advance the business forward.
Data availability is in or near real time.

Data-driven companies that use data in proactive and innovative ways are far more likely to thrive than companies that rely more on guesswork and ad hoc reporting.

Organizations today use a wide range of systems that produce digital footprints, ranging from applications and databases to event streams and files. All of these digital footprints serve as clues and building blocks to valuable insights. In order to lead and unify the efforts involved in gaining visibility and control over the growing volume, velocity and variety of data created by business activities, many organizations have created the position of the Chief Data Officer (CDO).



The CDO's mandate

As leaders of a company's data organization, CDOs are responsible for the development and management of all data assets as well as their associated teams, technologies and workflows. All data-related operations begin with data movement:

1. Extracting raw data from sources

2. Loading raw data to destinations

3. Transforming raw data into data models for analysis and operational usage

4. Using data models to build products such as dashboards, reports, predictive models, automated processes and more

Many organizations practice ETL, swapping steps 2 and 3, but we strongly believe that ELT is a better approach. More on that later!

In turn, data movement is the foundation of a broader strategy for turning data into value. There are four major use cases for data movement:

Data centralization

Consolidating data on a central platform for advanced analytics

Data democratization

Promoting self-service analytics across an entire organization

Infrastructure modernization

Moving to infrastructure that is cloud-based, more performant, more reliable and more compatible with modern technologies

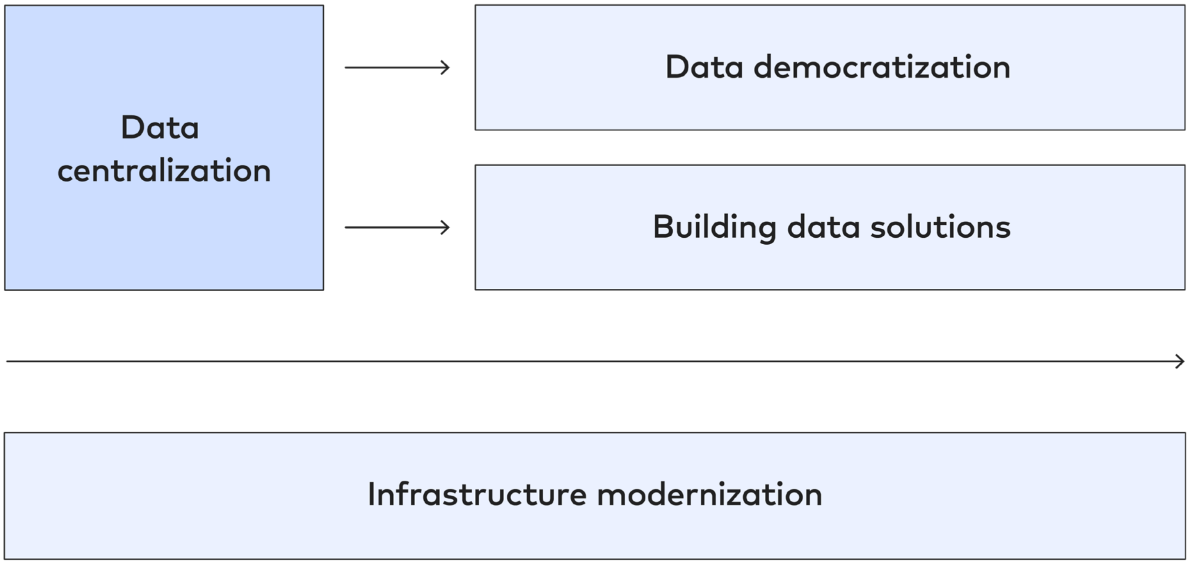
Building data solutions

Sharing data and building systems powered by data, both internally and externally

These use cases form a progression that leads organizations to greater capabilities:

1. **Data centralization** must come first, as without a single source of truth and data residing on a single platform it is difficult to systematically make use of data.
2. Once data is centralized, a profusion of data becomes available to users and **data democratization** is a common next step as organizations grapple with the difficulties inherent to enabling widespread access to data while preventing unsanctioned use.
3. Finally, an organization with a solid grasp of its data may choose to **build data solutions** – that is, share data and **products built using data** (such as predictive models) both internally and externally.

As an organization addresses the challenges of data centralization, data democratization and building data solutions, it will constantly revise its tools, technologies and platforms. **Infrastructure modernization** is a common thread running through all of the other use cases. It is usually impossible to rip and replace all the data infrastructure used by an organization in one fell swoop; rather, an organization should continuously adjust the tools and platforms it uses in response to changing needs.



These data movement use cases and the progression they represent require a technological solution in the form of an **automated data movement platform**, which features three pillars:

Automation
The processes and technologies used to move data must minimize the use of engineering time. Labor is the costliest asset for nearly all organizations. Automation further enables organizational agility and faster turnaround for all data-related executions, enabling organizations to assemble the data needed to support decisions in a matter of days rather than quarters.

Reliability
In a similar vein, data movement must involve a minimum amount of maintenance and downtime. Data pipelines must not be disrupted by schema changes or failed syncs.

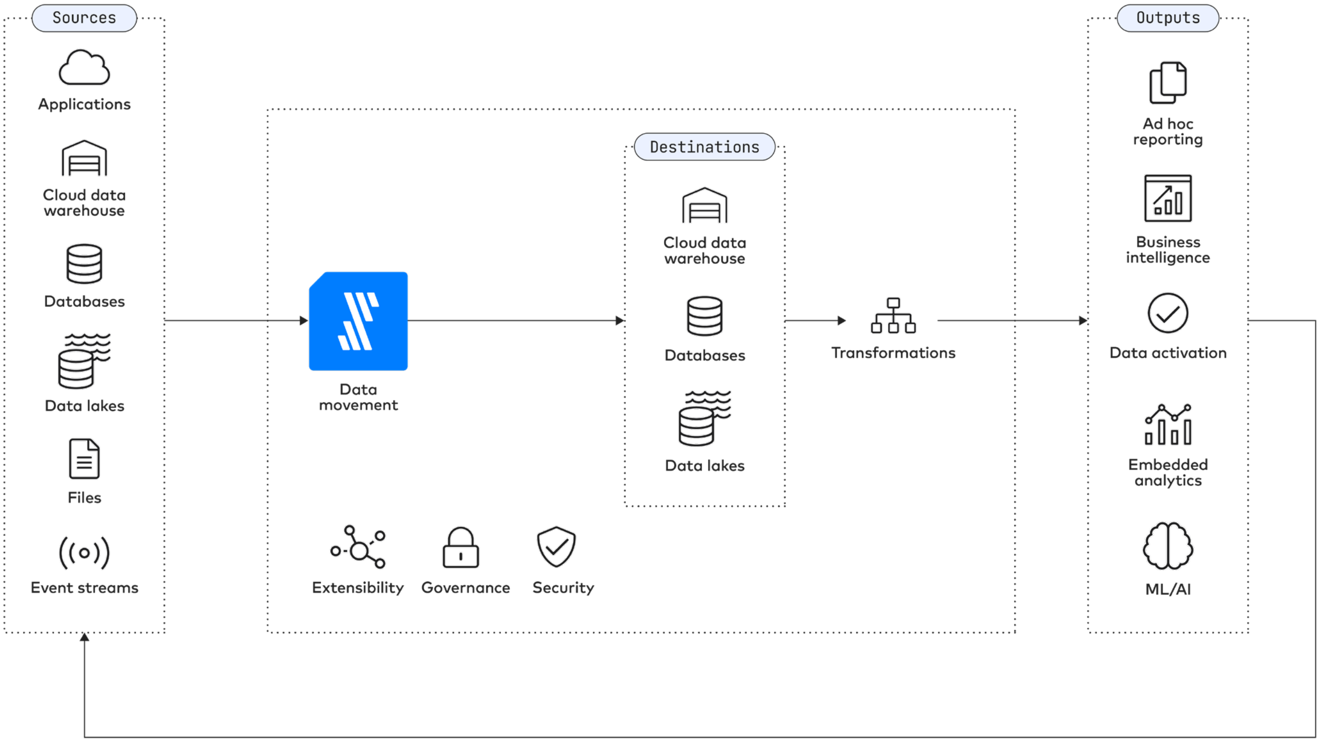
Scalability
Data operations must be able to seamlessly grow in scale. Keys to scalability include the ability to accommodate a huge range of data types and sources, govern the usage of data across departments and programmatically manage administrative tasks associated with data movement.

This guide will discuss the central importance of an automated data movement platform and how CDOs – and executives who share a similar portfolio at some organizations, like CIOs and CTOs – can use it to transform data assets into value. Along the way, we will think through each key use case for data movement, specifically:

1. Why it matters and how it helps an organization progress along the data maturity curve
2. The barriers that stand in the way of accomplishing the use case
3. How to address these barriers
4. Real-world examples of successful implementations

Data centralization

Data centralization means consolidating data from a variety of sources into a central destination. In practice, this destination tends to be a data warehouse in order to best support common analytics needs, though some organizations may use a data lake in place of or in addition to a data warehouse. Once centralized, data has many uses, ranging from business intelligence to machine learning and artificial intelligence.



Why data centralization?

The most immediate goal supported by data centralization is the ability to combine records from across disparate data sources into data models. Most commonly, data teams will use business intelligence platforms to translate the data models into visualizations and dashboards. Visualizations and dashboards support the discovery of insights to support business decisions.

Data centralization is essential to building a full, 360-degree view of an organization's operations, customers and products. A full understanding of an organization's operations means the ability to optimize internal processes. Likewise, understanding the customer lifecycle means a better ability to engage and support customers. Understanding how a product and its features perform (and don't) is essential to improving the core business and profitability of an organization.

Data centralization challenges

Centralizing data involves solving the following problems:

1. Accommodating a wide and growing range of sources
2. Ensuring that syncs run reliably and are resilient to upstream schema changes
3. Maintaining and upkeeping existing data connections as endpoints are updated
4. Guaranteeing data integrity and offering visibility into the status of syncs and data

The fundamental challenge associated with centralizing data is that moving data from a source to a destination is a deceptively complex engineering problem involving such considerations as designing the correct architecture, provisioning the appropriate compute and storage resources, ensuring timely, performant updates with minimal disruption to source systems, building in resilience to failure and more. A do-it-yourself data pipeline is an intensive undertaking that demands considerable investments in time, labor and money to both build and maintain the system.

Time is an especially important dimension of this challenge. There is considerable turnaround time between the need for an insight and the construction of a system to support it. Disruptions, such as pipeline failures due to upstream schema changes, inevitably cause downtime as well. Data engineering delays are the death of organizational agility, specifically the ability to identify and act on decisions in a timely manner.

The complexity, time commitment and expertise required to build a data centralization solution means that throwing engineering time and headcount at the problem turns a data organization into a cost center rather than a source of value and insight.

The inherent difficulty of moving data is multiplied by the fact that the data used by organizations is scattered and siloed across a wide variety of tools and platforms, each with its own idiosyncrasies. To make matters more challenging, the growing complexity of business needs means organizations regularly bring on new tools and platforms that produce valuable data but are siloed.

How to solve data centralization challenges

Data centralization fundamentally requires a technological solution in the form of a fully managed, automated data movement platform. As you look for an off-the-shelf solution, make sure to:

Find a data platform that is easy to use out of the box, with minimal need for configuration and engineering time to get started.

Ensure that the data pipeline uses an ELT architecture, rather than ETL. This simplifies the data pipeline, enables secure data processing and leverages the scalability of the destination system for transformations.

Consider the product's security features, especially if your organization is in a highly regulated industry where sensitive data must be obscured or entirely excluded.

Carefully note how the platform ensures reliability, including the ability to replay failed syncs without duplication, the ability to cope with upstream schema changes and optimizations for pipeline and network performance.

Check that it supports your current sources and destinations as well as those you are likely to use in the future. This is critical to ensuring that your data infrastructure can keep up with organizational growth and scaling.

The key is to leverage out-of-the-box connectivity to the fullest. The most basic unit of data centralization is the data connector, a discrete data pipeline that connects a source to a destination. Data connectors are deceptively complex to design, build and maintain. A well-functioning connector must be able to update from a source without disrupting operations, be resilient to failed syncs and other stoppages, accommodate changing data models at the source and be highly performant.

Ideally, members of a data team, including non-technical contributors, should be able to activate data connectors at will and largely disregard them thereafter. In practice, organizations that build data connectors end up with an open-ended commitment to support bespoke constructions that require the attention of engineers to deploy, activate and oversee.

Your expensive, highly-skilled engineers and analysts shouldn't spend time reinventing the wheel. Rather, they should leverage the strengths of a product that solves a known problem and an outside team with highly specific expertise dealing with specific data sources and destinations. This will enable your data organization to direct its talents toward modeling or analyzing data, and eventually building operational systems that use data models as inputs.

Using a robust, low-maintenance data movement platform and redirecting your data organization's energies toward analytics should enable a much shorter time to insight from the moment a question is posed about your organization's operations, customers and products.

AUTODESK

One example of successful data centralization in the wake of a sudden profusion of data sources comes from the world of computer-aided design software. In 2018, Autodesk acquired BuildingConnected and BIM 360, in the process introducing a wide-ranging mix of new applications, databases, customer data platforms and other tools to its data sources.

Previously, the BuildingConnected data team relied on a data pipeline that required extensive manual troubleshooting to accommodate column formatting changes and a data warehouse that did not scale easily. Likewise, the former BIM 360 data team struggled with a bespoke system that required constant supervision and manual remediation by analysts.

The answer to Autodesk's problems came in a data stack consisting of Fivetran, Snowflake and dbt, which enabled Autodesk to obviate the engineering-intensive work of supervising and maintaining the data pipelines. Its data team has since moved on to consolidating business intelligence dashboards, systematizing their approach to data modeling and building systems to support machine learning.

Data Stack

PIPELINE: Fivetran

SOURCES: Amplitude, AWS Lambda, Amazon S3, JIRA, Marketo, Mixpanel, MongoDB, NetSuite, PostgreSQL, Salesforce, SFTP, Stripe, Webhooks, Zendesk

DESTINATION: Snowflake

TRANSFORMATIONS: dbt

DocuSign

DocuSign, the electronic agreement and signature provider offers another example of successful data centralization unlocking the ability to greatly expand data operations. Originally, DocuSign made the common engineering choice of using a duplicated SQL Server operational database for analytics. People were confined to analyzing data from within each system or extracting CSVs into spreadsheets.

As the company grew rapidly, the business intelligence team realized it needed a new, systemic approach to sustainably add new data sources.

> "It would take a highly paid engineer anywhere from three to six months to build out a data pipeline and up to 20 hours a week afterward to keep things running. Our team would have to double in size."
>
> — Marcus Laanen, Senior Manager of Business Intelligence

To this end, DocuSign migrated its analytics environment to Snowflake and began using Fivetran. The upshot is that Docusign has been able to triple the number of sources it analyzes data from and maintains over 100 active dashboards that are regularly used across its organization. Its engineers now work on core projects rather than data engineering, and analysts can engage in data modeling and cataloging.

You may have noticed that themes of infrastructure modernization and data democratization are inextricably linked with instances of data centralization. We will discuss those shortly!

Data Stack

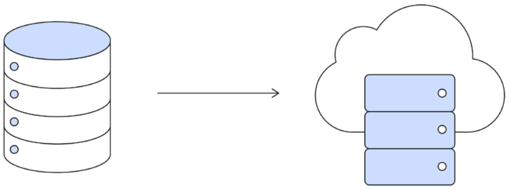
PIPELINE: Fivetran

CONNECTORS: AdRoll, Bing Ads, Google Ads, Google Analytics, Google Sheets, LinkedIn Ads, Mixpanel, SQL Server, Yahoo Gemini

DESTINATION: Snowflake

BI TOOL: Qlik Sense

Infrastructure modernization



Infrastructure modernization includes any instance in which an organization changes the tools, technologies and platforms it relies on for data operations. It can take place along several, non-mutually exclusive dimensions:

- An organization may move from on-premise operational systems, pipelines and destinations to the cloud for greater flexibility and speed along with reduced maintenance.
- Data teams may switch a data pipeline from ETL to a more modular, flexible ELT-based architecture.
- Different destinations differ by cost structure, performance and other attributes, so an organization may **migrate from one type of destination to another**.
- In the pursuit of greater organization agility, a data team might upgrade its data movement capabilities from intermittent, batch updating to real-time or streaming.

For smaller and newer organizations, this is typically a change from a crude, DIY (or nonexistent) data stack to a cloud-based data stack. Without an incumbent solution, it is relatively straightforward to adopt and start using the new solution. For larger and more established organizations rip-and-replace is seldom viable due to the importance of workflows that depend on the existing infrastructure.

Why modernize infrastructure?

Every organization that chooses to modernize infrastructure does so in service of some other important use case. Fundamentally, modernizing infrastructure is about improving the capabilities of your data organization by improving its tools. Common benefits include improved flexibility in terms of compute and storage, lower costs and reduced engineering workloads. Other considerations include interoperability with new technologies and attracting fresh, best-in-class talent. Newer data technologies are all cloud-based, and recent graduates of engineering, machine learning and data science programs have all been trained in cloud-based environments.

The main reasons to modernize infrastructure are to enable the centralization of data at infinite scale without upfront capital outlays and to complement data with other cloud-based analytics technologies. As demonstrated by the case studies in the previous chapter, companies are seldom able to change their data centralization capabilities without also adding or changing some elements of their data stack.

As previously discussed, centralizing data is the essential first step to enable further uses of data, including democratizing data and building systems that monetize data. Infrastructure modernization plays a major role in all of these stages.

Infrastructure modernization challenges

A fundamental stumbling block to infrastructure modernization is that ongoing, mission-critical analytical and operational processes depend on existing infrastructure. This infrastructure and the processes that rely on it cannot be shut down without jeopardizing an organization's existing commitments. Furthermore, queries made over the course of the migration itself must not interfere with the performance of operational systems. The upshot is that migrations can be costly, as for some amount of time both the legacy and new environments will be simultaneously active as old data is migrated to the new platform.

In this regard, smaller, leaner organizations without existing obligations have a real agility advantage. Without a need to migrate historical data, it is straightforward to simply start from scratch and transition to a new system with minimal disruption.

Another key challenge is security and compliance. Once data is on the cloud it is no longer confined to proprietary infrastructure (although it may be secured from the public by being routed through a private cloud). Regulatory compliance and brand risk are serious concerns that must be met with strong authentication and authorization protocols, end-to-end encryption and other security features. Relatedly, data privacy and access control are important considerations. Data leaders must prevent unauthorized access while data is migrated from on-premise to cloud.

Finally, a longer-term cost-related consideration that applies once the cloud environment is active is that it is easy to offer too much access to too many parties too quickly and accidentally spend too much on compute and storage, leading to ballooning expenses. Fortunately, cloud-based data stacks do offer means to get these expenses under control. This is a governance issue that we will address in the next section on data democratization!

How to solve infrastructure modernization challenges

Unless your organization has no existing obligations downstream of your infrastructure, you must keep existing processes running while you set up the new ones. The good news is that there are several ways to contain costs in the new environment as it ramps up.

You don't need to migrate all parts of your infrastructure at once and it can make sense to modernize infrastructure piecemeal. Hybrid cloud architectures in which some elements of the infrastructure remain on-premises or in private clouds are possible. This is especially relevant in sensitive, highly-regulated industries. As a starting point, however, many organizations choose a managed, cloud-based data platform as a destination.

As with data centralization, the key is to abstract as much complexity and custom engineering (and with it, labor costs) away from your data team as possible, specifically the design, construction and maintenance of data connectors. A team saddled with such responsibility must periodically rebuild pipelines in response to bugs, performance issues, changing upstream data models and downstream business needs, leading to downtime and stale data. The solution is to adopt a data movement platform that features off-the-shelf data connectors and supports both on-premises and cloud-based data sources.

It can be a good idea to start a migration effort with applications, which produce valuable data (especially in sales and marketing) but tend to be less demanding than operational databases in terms of technical complexity and security requirements. Migrating applications is a low-risk, low-cost approach to building a strong proof of concept. Building on the strength of initial success, your team will be able to approach subsequent efforts with more confidence, setting up your team to activate connectors to more data sources, including operational databases, over time.

For the sake of security, compliance, privacy and access control, look for regulatory compliance features, end-to-end encryption and the ability to designate different levels of access by role (i.e., **role-based access control** or RBAC) within the organization.

Oldcastle Infrastructure™
A CRH COMPANY

The building materials company Oldcastle Infrastructure once maintained a separate on-premises operational database and cloud NetSuite ERP. It could not view transactional, manufacturing and production data in a single environment. Initially, the company attempted a cloud migration from SQL Server to Azure. This effort took five months of planning and was three months into execution when Nick Heigerick, IT Manager of BI, realized "We didn't have the expertise to manage so many moving parts in-house, so we had to decide: Are we going to pay consultants to do this forever or is there a better way?" In short, Nick and his team realized that this approach would struggle to accommodate new data sources and scale the company's usage of data.

On the advice of systems integrator Interworks, Oldcastle Infrastructure swapped Azure out for Snowflake and connected Fivetran to both its on-premise SQL Server and its NetSuite instances. Within 10 days, the entire data set had been migrated to the cloud. The migration and new capabilities led to the following returns:

- $360K in annual infrastructure savings, and $25M ROI directly attributed to replatform efforts
- Operating profit growth of 21.5% in one year
- ROI in the next 12-18 months is estimated to be a minimum of $25 million
- The ability to sync data from Salesforce, Coupa, Box and other SaaS applications

Data Stack

PIPELINE: Fivetran

CONNECTORS: SQL Server, NetSuite, Salesforce

DESTINATION: Snowflake

BI TOOL: Tableau

SI: InterWorks

Another example of a migration story from a legacy ETL solution to modern ELT comes from the AI-powered driver and fleet safety provider Nauto. For years, Nauto used a proprietary data repository and maintained a patchwork of point-to-point ETL integrations through Informatica to stitch together operational and analytical systems. Broken integrations could cause days of downtime that affected orders, payment processing, subscriptions and hardware shipments. Every month, Nauto's IT team spent an average of three days and 80 engineering hours debugging data inconsistencies.

Moving to fully managed services including Databricks as a destination and Fivetran as a data pipeline enabled Nauto to gain total access to all of its data and control over formatting. To complete the loop, Nauto also engaged the help of the data activation tool Hightouch, enabling the flow of data from Databricks back to NetSuite and Salesforce and streamlining operations such as device returns. Ultimately, Nauto has optimized its data engineering spend by 75% since its Informatica days.

> "With Databricks, Fivetran, and Hightouch, everyone can bring their own data in any format and compare it with everyone else's to get to the truth so that we can make the right decisions for the company."
>
> — Ernest Prabhakar, Business Data Lead, Nauto

Data Stack

ELT: Fivetran

TARGET: Databricks Lakehouse

SOLUTION: Predictive AI

PLATFORM USE CASE: Lakehouse, Machine Learning, Ingestion, Reverse ETL

CLOUD PLATFORM: AWS

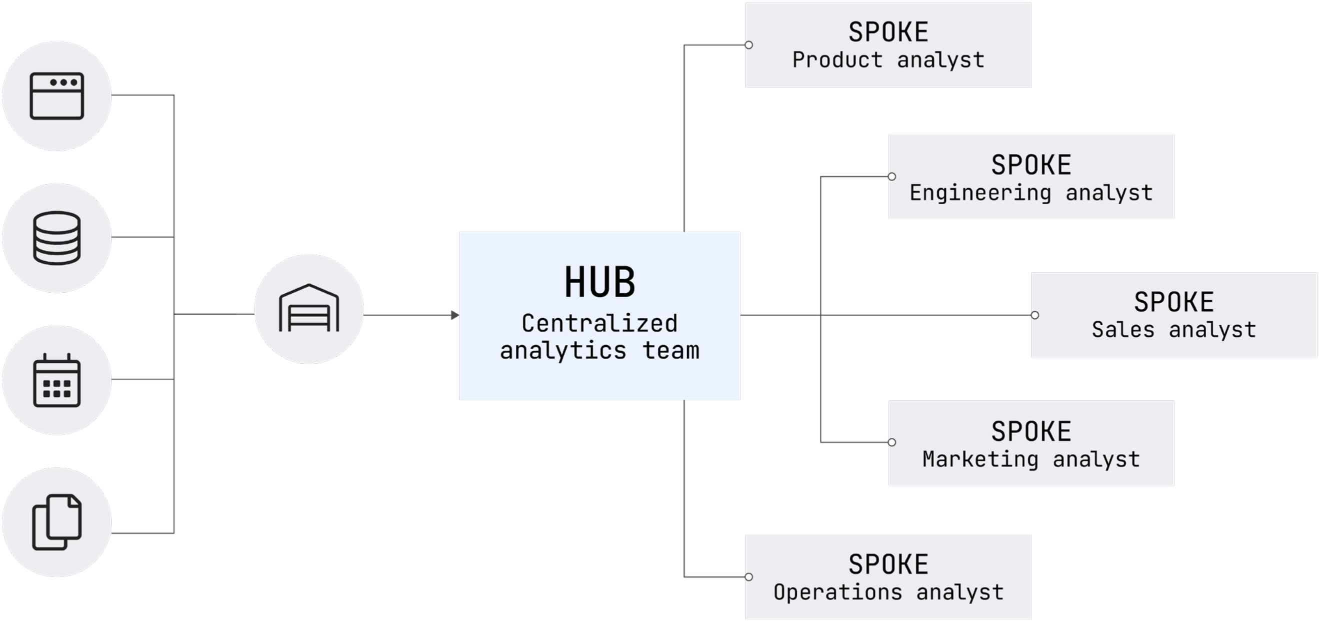
ADDITIONAL COMPONENTS: Hightouch

Data democratization

Data democratization is a driving ambition of CDOs and consists of promoting data literacy and self-service analytics across an organization to become a truly data-driven enterprise. To get there organizations need to build trust in the quality, accuracy and reliability of their data to get adoption and buy-in from end users. With the appropriate security and governance measures in place, different teams across an organization are empowered to own their pipelines and load their data back into a centralized warehouse. Data democratization enables people ranging from individual contributors to executives to routinely consult and use data to influence decisions.

Why democratize data?

The key benefit of data democratization is the ability to scale critical data-driven decision making by business users who know their domain best, enabling an organization to become more agile and responsive to changing market conditions. This approach is most suited to companies that have built decentralized data teams (as in a hub-and-spoke model) in industries with less stringent data regulations.



Data democratization challenges

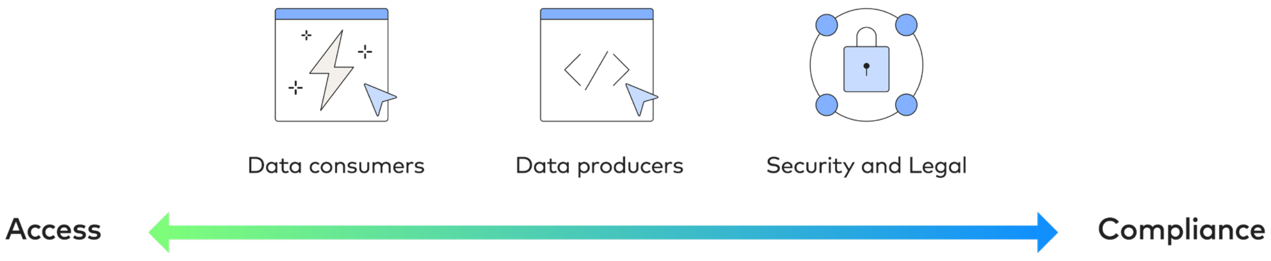
The core challenge of data democratization is governance, and the core tension that data governance must solve is between access and compliance.

Every organization contains three key groups of stakeholders:

Data consumers
Such as analysts want unrestricted access to data to facilitate their projects. They are frustrated by slow turnaround and stale data and often end up finding ways to circumvent existing data service processes, accessing and producing data products in unsanctioned ways.

Data producers
Typically teams of data engineers, are responsible for managing a growing queue of data integration projects. They are forced to balance the competing interests of the security and legal teams, on the one hand, and analysts on the other. For security and legal teams, they must have answers for audits. For analysts, they must be able to onboard and deploy new data sources quickly and reliably.

Security and legal teams
Prioritize regulatory compliance, especially in the face of continued regulatory changes. Their main goal is to minimize data misuse and the risk of data breaches. The worst case scenario for security and legal teams is to be the subject of data mishandling resulting in fines and reputational damage that can be hard to recover from.



Given the high stakes involved with improper exposure and misuse of data, an organization might impose stringent limitations on access in order to guarantee compliance. The added layer of red tape causes analysts and business users to miss out on opportunities as they are blocked by long turnaround times and stale data.

How to solve data democratization challenges

The solution to simultaneously ensuring access and compliance comes down to knowing, protecting and controlling authorized access to data. These challenges require the use of technology for a scalable solution. Look for a data movement platform with the features described below.

Knowing data requires full visibility of all the data that is being generated across the organization. It is impossible, however, for one person or even one data team to easily know all the data that is being created and integrated across the enterprise. Consolidating a view of all the data in a data catalog via metadata sharing allows the data stewards to track column-level lineage and schema changes to see what data exists and how it changes over time.

To protect data, an organization needs to be able to enforce policy and compliance requirements. This requires basic **security features** as well as role-based access control to enforce a hierarchy of permissions to prevent insecure or non-compliant actions on pipelines or destinations.

Finally, controlling access determines who can access what data with a fine degree of granularity. At scale, this requires the aforementioned role-based access control as well as repeatable, programmatic control over approval workflows for a data platform, such as through an API to reduce the risk of human error. Another feature that enables granular control at scale over data is automated user provisioning using System for Cross-domain Identity Management (SCIM) such as Okta or Azure AD, allowing an organization to quickly and safely onboard many users and manage the user lifecycle journey.

The goal of data democratization is ultimately advanced by using technological solutions to ensure both access and compliance. This resolves the conflicting needs of data consumers, data producers and security and legal teams, allowing organizations to safely provide self-service analytics to decision makers at all levels.

wework

Shortly after going public in 2021, office space provider WeWork was forced to contend with a stricter compliance landscape as a publicly traded company. At the same time, the company needed visibility into occupancy, turnover, outstanding renewals, member growth and profit margins per location, per region and holistically across the company. These metrics would enable stakeholders at all levels – community managers, the senior leadership team and even investors – to determine the factors influencing the business's success.

Fivetran offered WeWork the ability to both set up hundreds of new data connectors as well as ensure security at scale. Each of WeWork's hundreds of global locations uses a Postgres operational database, while the company also uses many applications such as Salesforce, Pardot, Mandrill, Zendesk, Stripe and more. Instead of spending hundreds of engineering hours per month building and maintaining connectors to these data sources, Fivetran enabled WeWork to outsource this basic legwork, enabling the data engineering team to engage in the higher-value work of building data models.

With the help of Fivetran's metadata API and auditing and logging features, WeWork was able to gain end-to-end visibility into each of its hundreds of data connectors. With the help of Fivetran, WeWork can not only synchronize hundreds of data sources but also ensure that employees at all levels are able to safely and responsibly access the data they need to get business critical insights.

Data Stack

PIPELINE: Fivetran

CONNECTORS: Postgres, MySQL, Salesforce, Pardot, Mandrill, Zendesk, Bing Ads, Appsflyer, Intercom, Google Sheets, Confluent, Airtable, Stripe, Aurora and more

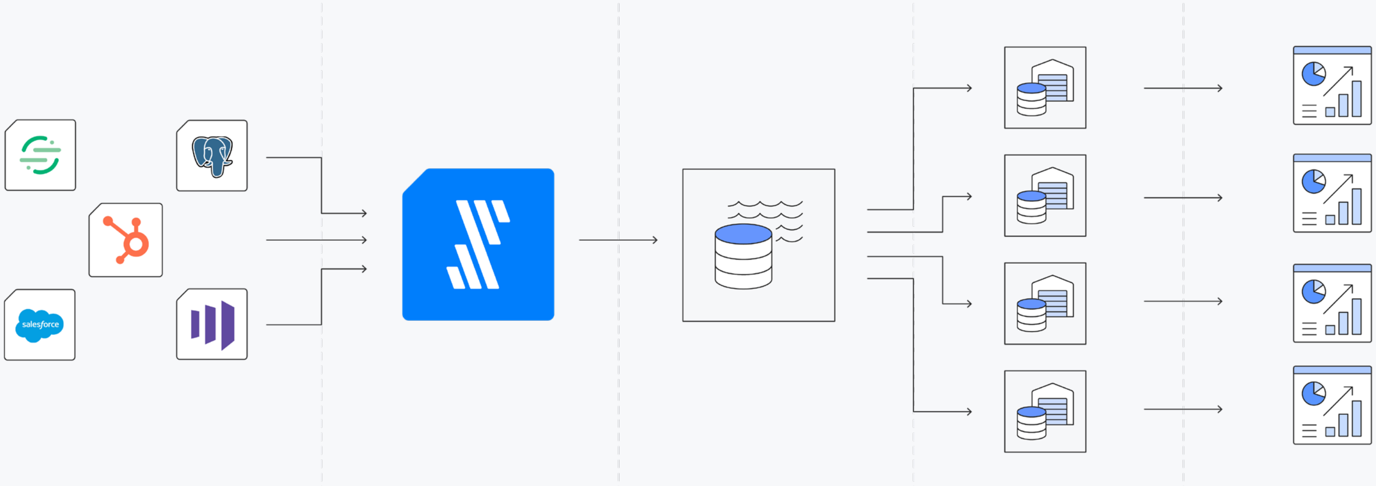
DESTINATION: Snowflake

CLOUD SERVICE PROVIDER: AWS, GCP

Building data solutions

Aside from data democratization, data centralization can also aid the creation of a wide range of novel products and services. At Fivetran, we often liken data to electricity – an enabling technology with unlimited potential for innovation. These innovative data solutions fundamentally involve adding value to data and sharing it internally or externally. Roughly speaking, there are three ways to think about data solutions:

1. **Enterprise pipeline management** involves management of data assets at a massive scale and complexity, requiring programmatic control. Large organizations can have data travel in a dizzying range of directions for both analytical and operational purposes.
2. **Analytics products** are derived from raw data and can range from dashboards and reports all the way to productionized artificial intelligence/machine learning models.
3. **Data sharing systems** make real-time data available both internally to business units within an organization and externally to customers and other third parties.



Why build data solutions?

The huge volumes of data any organization produces can be extremely valuable in the right hands. Data solutions are about finding ways to monetize data by making it available to the appropriate parties at the appropriate levels of refinement, ranging from views of key metrics all the way to systems that use data to produce recommendations and automate decisions.

Data solution challenges

Enterprises face several challenges related to building data solutions.

1. **Integrating technologies** can be difficult, as enterprises often use dozens or hundreds of data sources, tools and platforms.
2. Data solutions may involve handling data on behalf of customers, which requires **onboarding customer data** that an enterprise doesn't produce or directly own.
3. Data solutions must simultaneously **ensure access and compliance** so that valuable data is only exposed to the appropriate parties.
4. In order to **monetize data assets**, an enterprise needs some mechanism for pricing and selling data. This requires visibility into the unit economics of the data.

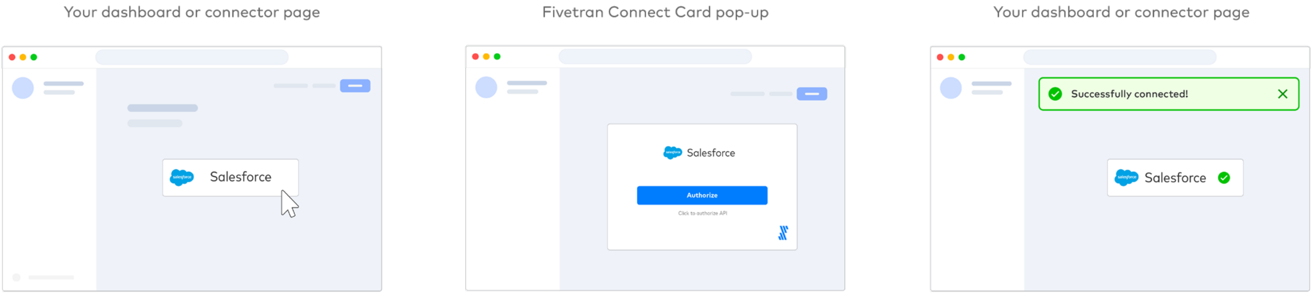
How to solve data solution challenges

The aforementioned challenges can be addressed with the implementation of the right technology.

The ability to integrate disparate technologies largely depends on finding a data platform with the following characteristics:

- It is compatible with a wide range of vendors and solutions, especially orchestration technologies (such as Terraform and Airflow) for executing custom workflows
- It supports a wide range of data sources and destinations
- It features an API that enables large-scale, programmatic management of data connectors and users

The best way to onboard customer data is to enable self-service for the customers in question so that there is minimal exposure to inappropriate parties. Self-service data sharing should be enabled through a graphical interface, such as embeddable authentication popups. At Fivetran, we call this capability "connect cards."


Your dashboard or connector page
Salesforce
Fivetran Connect Card pop-up
Salesforce
Authorize
Click to authorize API
Your dashboard or connector page
Successfully connected!
Salesforce

1. User clicks logo in your UI
2. User enters credentials for chosen data source
3. User has successfully connected to data source and a new connector is created

Access and compliance depend on data governance and security features. Some we have previously discussed, such as various certifications and role-based access control. Another important security feature includes localized data residency. It is also important f or a data platform to avoid comingling of customers' data, through process isolation and multitenancy.

Finally, features that support gaining visibility into the unit economics of data in order to monetize data assets include metadata APIs and the ability to programmatically ingest logs. These should enable a user to observe consumption, spend and the general health of data connectors.

groupm

Global media agency group **GroupM** serves over 200 clients with top-tier advertising campaigns. Data sharing with clients is a core concern for GroupM. In the past, the GroupM data team would pull marketing data directly into spreadsheets but was forced to contend with pipeline failures, formatting issues and human error of all kinds. Eventually, a major client began asking GroupM for dashboards for historical data analysis and day-to-day reports. The data team realized that other clients would likely have similar needs and that their existing data movement workflow could not support it.

In order to support the needs of their clients, GroupM needed to centralize its large volumes of data in a common repository and grant secure access to every client at scale. Fivetran addressed this problem in two ways. First, Fivetran data connectors provide the capability to extract, load and transform data from a wide range of marketing analytics data sources. Second, Connect Cards offer GroupM's clients an embeddable interface to set up connectors without exposing their proprietary data to GroupM.

All told, GroupM now saves up to 75 engineering hours every month. A further benefit is that GroupM's clients now have full and easy access to historical data. Longer term, GroupM plans to build a sales model using machine learning.

Data Stack

PIPELINE: Fivetran

SOURCES: Marketing (Facebook Ads, Linkedin Ads, Snapchat Ads, Google Ads, Google Analytics, Google Campaign Manager, Twitter Analytics)

DESTINATION: Google BigQuery

CLOUD PLATFORM: Google Cloud Platform

BI TOOLS: Google Data Studio

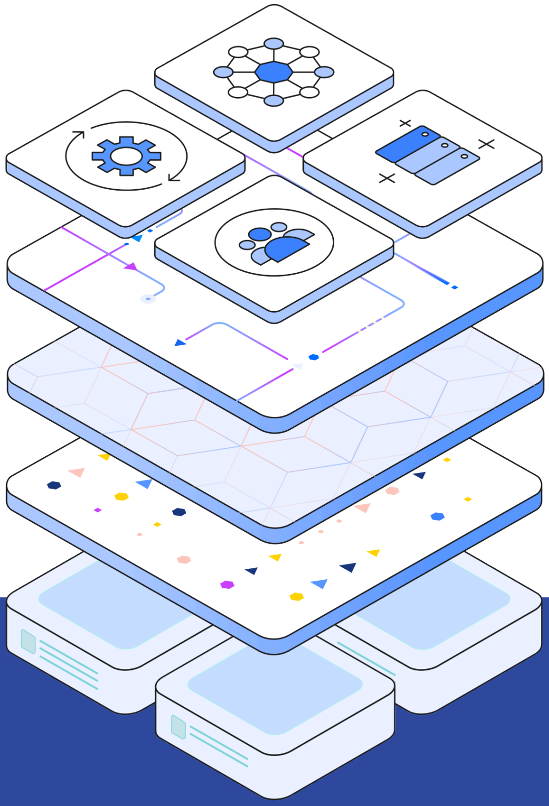
A roadmap for innovation and data-driven value

The four use cases we have discussed – data centralization, infrastructure modernization, data democratization and building data solutions – are all essential elements of a cohesive strategy for digital transformation. Data centralization is the essential first step to accomplishing anything with data. The ability to easily and reliably move data from sources to destinations is upstream of democratizing data and using data to build products. As an organization progresses through different stages of its strategy, infrastructure modernization remains a persistent concern as teams experiment with, adopt and sunset tools and technologies.

Each of these elements can be addressed using a modern cloud data movement platform that leverages automation, security, governance and extensibility features to provide reliable, secure and scalable data movement. Often organizations can lean on technology to solve the problems that arise from addressing competing concerns with data.

With continuing developments in artificial intelligence and machine learning, the most innovative uses for data have yet to be invented. In the meantime, your organization stands to benefit from following a roadmap predicated on progressing through the four data use cases.

Good luck!



Fivetran

Fivetran automates data movement out of, into and across cloud data platforms. We automate the most time-consuming parts of the ELT process from extracts to schema drift handling to transformations, so data engineers can focus on higher-impact projects with total pipeline peace of mind. With 99.9% uptime and self-healing pipelines, Fivetran enables hundreds of leading brands across the globe, including Autodesk, Conagra Brands, JetBlue, Lionsgate, Morgan Stanley, and Ziff Davis, to accelerate data-driven decisions and drive business growth. Fivetran is headquartered in Oakland, California, with offices around the world.

For more info, visit Fivetran.com.

Start your free trial