



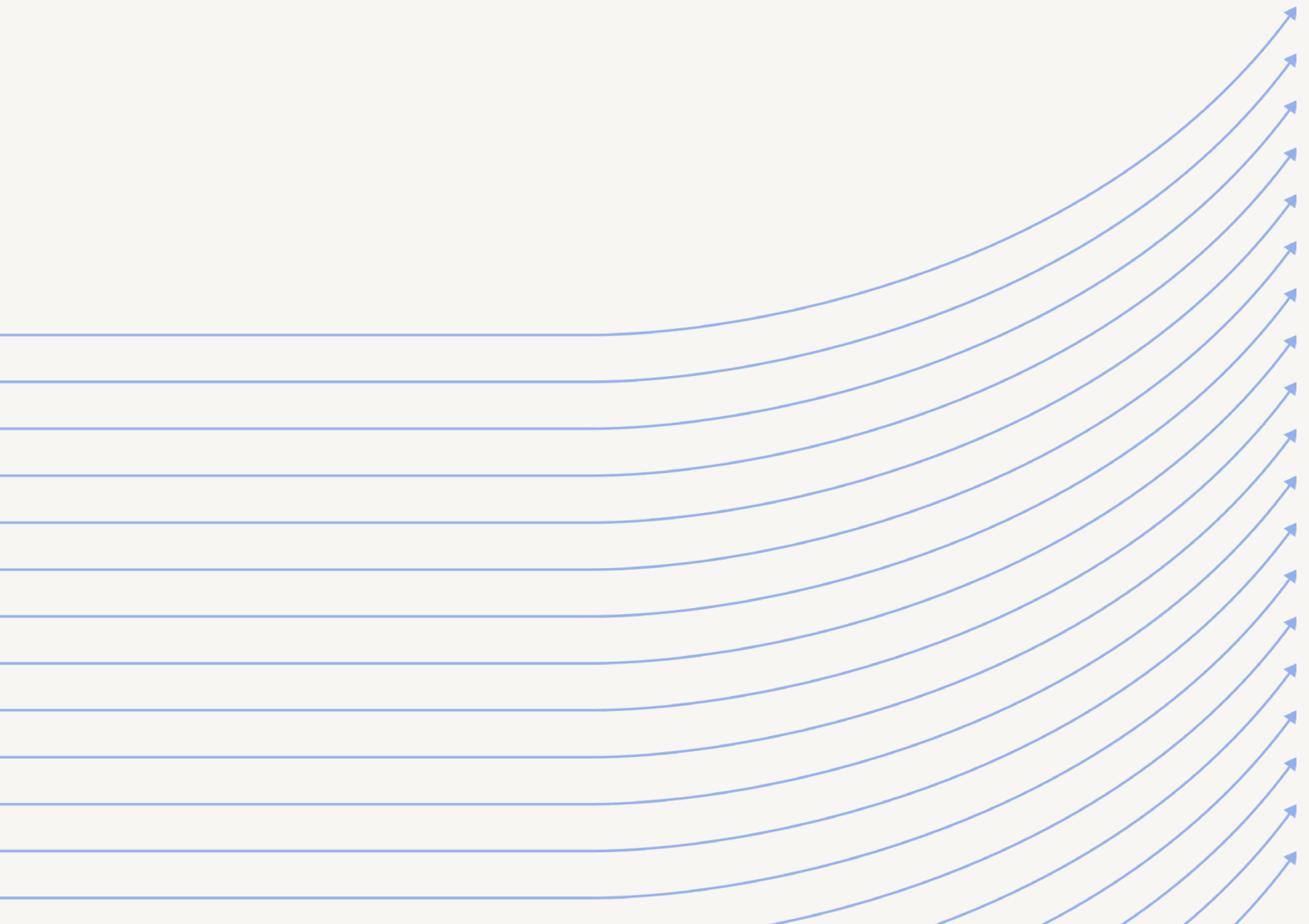
A CIO's guide to data lake management for generative AI

How to build an analytics-ready data foundation
for AI and GenAI success.



Table of contents

Executive summary	3
The CIO's role in data lake management	3
The strategic value of data lakes for AI and ML	4
Data governance for data lakes	5
Strategies to simplify data lake management	5
Customer spotlight: Coupa	8



Executive summary

Generative AI is poised to revolutionize productivity and business growth. [McKinsey](#) predicts it could add trillions to the global economy. Yet, a new [MIT survey](#) reveals that while 82% of C-suite and senior executives prioritize scaling AI use cases, many are grappling with data readiness for AI and generative AI.

The number one reported challenge, cited by 45% of executives in the MIT study, is data integration and pipelines. This data is concerning given that organizations are allocating [13% of their annual revenue](#) on average to AI initiatives. The fact that executives are struggling with such a fundamental aspect of data engineering suggests they are likely unprepared for AI deployment and scalability.

Many organizations struggle with manually built data pipelines that consume excessive time and resources, leading to inefficiency and impeding AI scalability. Moreover, without governance, data lakes become murky, unusable "data swamps" instead of functional data lakes. The opportunity cost of delaying modernization is higher than ever as GenAI offers significant efficiency gains.

This ebook provides CIOs with strategies to accelerate and ensure the success of AI and GenAI initiatives through a managed data lake service. It emphasizes the strategic importance of data lakes, the critical role of data governance and how to streamline operations with modern data management solutions.

"Where [Honeywell] has recently deployed generative AI, it is already seeing a huge uptick in productivity."

– Suresh Venkatarayalu, Chief Technology and Innovation Officer

Source: "AI readiness for C-suite leaders" by MIT Technology Review Insights

The CIO's role in data lake management

As strategic leaders, CIOs must align data strategies with business objectives. CIOs should:



Define missions that connect data strategies to impactful AI/ML use cases. Articulate how a governed, analytics-ready data environment catalyzes innovation through AI and machine learning initiatives.



Manage costs effectively by standardizing platforms and controlling infrastructure sprawl. Avoid the proliferation of multiple infrastructures and tools. Standardizing platforms simplifies operations and ensures more reliable AI deployment.



Facilitate a culture shift towards data stewardship and robust governance across departments. While technology is key, CIOs must also drive the organizational change required for AI transformation.

Building a well-managed, governed data lake is essential both technologically and strategically. MIT reports that 83% of C-suite executives have highlighted the need to consolidate numerous siloed data sources, a task demanding CIO leadership for effective data integration and movement.

Data lakes are evolving into multi-tenant resources supporting broader use cases. They require trustworthy, compliant data that is automatically cataloged, cleaned and prepared. This streamlined approach not only reduces time-to-insight but also boosts the data lake's active usage. By focusing on data movement, quality and governance from the start, CIOs can accelerate AI development and reduce risks.

The strategic value of data lakes for AI and ML

Data lakes are pivotal for unlocking the potential of generative AI, offering significant advantages over traditional databases and data warehouses. Data lakes:



Support diverse data types necessary for training sophisticated AI and ML models.

Unlike these traditional systems, data lakes store and manage large volumes of unstructured, semi-structured and structured data without extensive preprocessing.



Preserve data in its raw format, which provides maximum flexibility for future analytical use cases. Preserving the original details and integrity of data gives businesses maximum flexibility for accommodating any future analytical use cases.



Enable both horizontal and vertical scalability, accommodating vast amounts of data and robust processing capabilities. Horizontal scaling lets businesses store virtually unlimited amounts (petabytes) of data, while vertical scaling enables high-performance processing and analytics on large datasets.



Offer flexible data ingestion from various sources, which helps build accurate AI models and enables informed decision-making. The ability to integrate structured, semi-structured and unstructured data from various sources allows data lakes to support everything from small-scale experiments to enterprise AI deployments.



Empower unified analytics. By consolidating data from disparate sources, data lakes provide the holistic view critical for building accurate AI models. The unified data environment allows AI models to generate cross-functional insights and enable informed decision-making. A consolidated data foundation ultimately powers advanced analytics and machine learning workflows that drive better outcomes and strategic advantages.

Data lakes provide agility while simplifying data consolidation and empowering AI model accuracy. To capitalize on generative AI, CIOs must lead the charge in architecting data lakes optimized for AI success. Organizations need a future-proof data infrastructure that can flex with evolving requirements.

Data governance for data lakes

Sixty percent of C-suite executives say that addressing data governance, trust and security issues is a prerequisite for achieving their AI goals. Effective data governance is essential for optimizing AI initiatives built on data lakes. Strong governance frameworks ensure:

- ◆ **Data quality:** Clean, accurate and reliable data produces trustworthy AI models.
- ◆ **Compliance:** Adhering to stringent data privacy and security regulations like the GDPR and CCPA mitigates risks of fines and reputational damage.
- ◆ **Security:** Implementing robust measures like encryption and granular access controls protects sensitive data assets.

Strategies to simplify your data lake management

The Fivetran Managed Data Lake Service

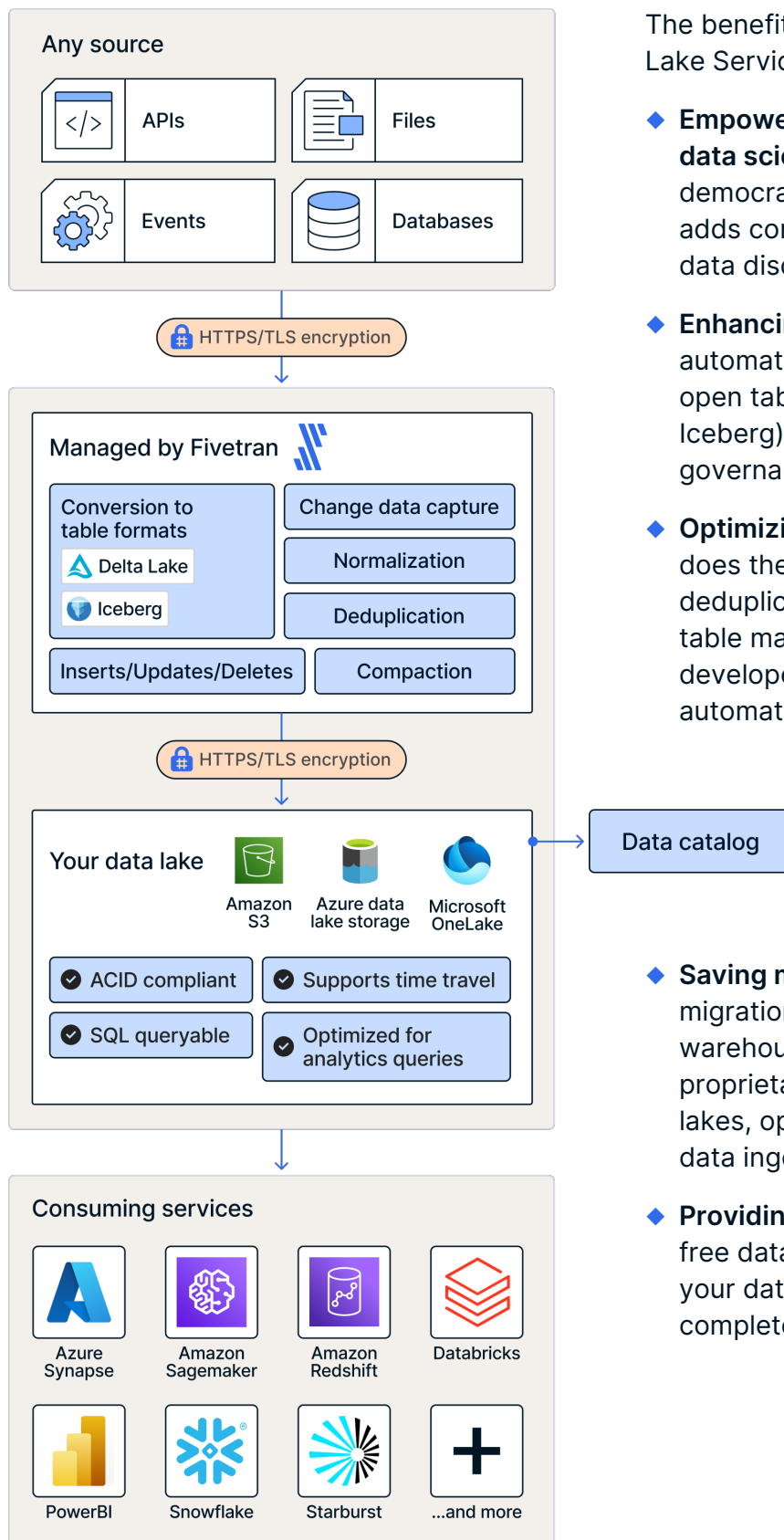
With the ability to land over 600 pre-built or custom data sources into any major

data lake destination, Fivetran automatically normalizes, compacts and deduplicates your data before standardizing it in Iceberg or Delta Lake table format. From securely and reliably moving data into the data lake to continuous table maintenance and updates, Fivetran automates your data lake management.

The Fivetran Managed Data Lake Service:

- ◆ **Intelligently adapts** to changes in your data,
- ◆ **Automatically evolves** your changing schemas,
- ◆ **Strategically optimizes** your data so that it's clean, cataloged and standardized in a query-ready table format for analytics and
- ◆ **Lowens ingestion costs** by leveraging the Fivetran managed service to cover the compute costs resulting from data ingestion into the data lake.

Even before your data lands in the lake, Fivetran provides robust data privacy and quality controls. Customers can block or hash PII fields before ingestion, while Fivetran automatically cleanses, deduplicates and normalizes data to analytic standards. It then writes the cleansed data to cloud storage like Amazon S3, OneLake and ADLS in the Delta Lake and Iceberg open table formats, populating metadata into data catalogs such as AWS Glue, Polaris and Unity, enabling immediate query readiness. Furthermore, Fivetran covers the cost of ingestion, freeing more resources for AI projects.



The benefits of the Fivetran Managed Data Lake Service include:

- ◆ **Empowering business users and data scientists alike** with centralized, democratized, query-ready data that adds context, invites insights and drives data discovery.
- ◆ **Enhancing operational efficiency** by automatically converting your data to open table formats (Delta Lake / Apache Iceberg) with robust data cataloging and governance features.
- ◆ **Optimizing developer time:** Fivetran does the heavy lifting of table updates, deduplication and other low-level table maintenance so you don't waste developer time on work that can be automated.
- ◆ **Saving money** by automating data migration away from expensive data warehouses that lock you in with proprietary data formats and toward data lakes, open data formats and covering data ingestion costs.
- ◆ **Providing peace of mind** with worry-free data replication that ensures your datasets always arrive clean and complete, with every change captured.

Manual data lake management	Benefits with Fivetran Managed Data Lake Service
Time and labor intensive for data engineers	Set and forget, easy for data engineers
Complexity and fragmentation of bespoke pipelines and integrations	Data sources automated and managed with simple credentials
Low quality and inconsistent data, prone to errors	Data quality and consistency built-in
Scalability challenges as data volumes grow, leading to pipeline problems	Maintenance-free scaling as data volumes grow
Compliance and security risks from lack of governance and data protection	Compliance and security taken care of with robust governance and data protection (masking, hashing)
Maintenance overhead of keeping up with data source changes	Automated schema management and data source connector functionality
Lack of real-time capabilities	Near real-time data ingestion with 5-minute syncs
Skill gap in pipeline design, implementation and delivery	Easy-to-use interface that anyone can learn
Leads to more technical debt	Managed service with no technical debt costs

Fivetran fulfills all your requirements for a well-governed, analytics-ready data lake:

- ◆ **Comprehensive connectivity:** Over 600 pre-built connectors.
- ◆ **Cloud cross-platform:** AWS, Azure and GCP support.
- ◆ **Lakehouse format output:** Data conversion to Iceberg or Delta Lake, with underlying Parquet files.
- ◆ **Continuous table management and maintenance:** In-transit cleansing, normalization, schema evolution, inserts/updates/deletes, file compaction, performance optimization.
- ◆ **Data quality:** Deduplication, schema evolution and management.
- ◆ **Governance integration:** AWS Glue, Databricks Unity Catalog.
- ◆ **Cost-effective:** Ingestion costs paid by Fivetran.

An IDC study found Fivetran customers realized an average \$1.5 million in annual benefits and 459% three-year ROI — savings that you can reinvest into deriving more value from your data lake.

Source: "The Business Value of Fivetran" by IDC

Coupa accelerates the adoption of its S3 data lake with Fivetran

Business challenge

Coupa faced challenges with ingesting and centralizing data from multiple applications and platforms in the cloud. Data silos hindered its ability to quickly generate insights on platform usage and customer behavior, resulting in missed opportunities and outdated insights. Coupa recognized the need to consolidate data sources to gain a holistic view of customer interactions and purchase patterns, prompting the search for a solution to break down these data silos and enable more effective analytics.

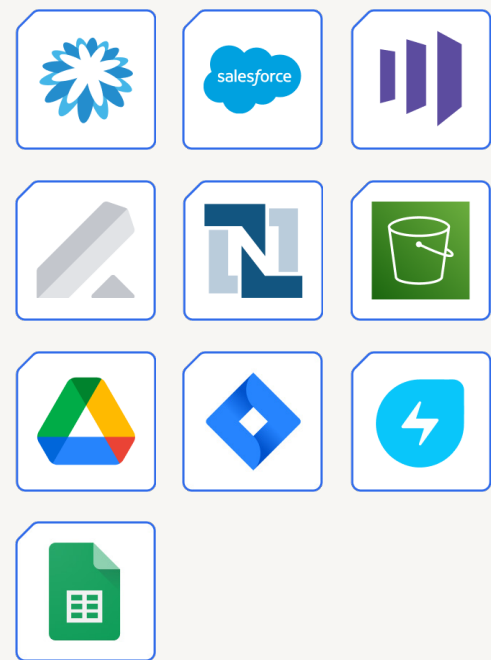
Outcome

Fivetran enabled Coupa to automate data movement into an Amazon S3 data lake, reducing time-to-value from 3-6 months to just weeks. Coupa tripled its data volume and integrations, quadrupled the number of data users, and enhanced its strategic decision-making capabilities.

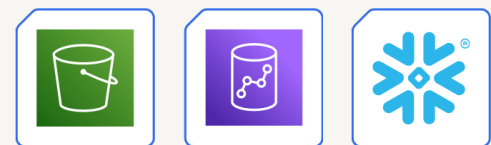
Key results:

- ◆ 3X increase in data volume and integrations
- ◆ 4X more data users across the organization
- ◆ Significantly improved data quality and accuracy
- ◆ Reduced time-to-value from 6 months to weeks
- ◆ Accelerated, data-driven customer insights

Connectors



Destination and cloud



With the Fivetran Managed Data Lake Service, we're making data as accessible and reliable as electricity, empowering businesses to unlock new opportunities and drive innovation.



Sign up for a 14-day
free trial of Fivetran



Fivetran, the global leader in data movement, helps customers use their data to power everything from AI applications and ML models, to predictive analytics and operational workloads. The Fivetran platform reliably and securely centralizes data from hundreds of SaaS applications and databases into any cloud destination — whether deployed on-premises, in the cloud or in a hybrid environment. Thousands of global brands, including Autodesk, Condé Nast, JetBlue and Morgan Stanley, trust Fivetran to move their most valuable data assets to fuel analytics, drive operational efficiencies and power innovation.

For more info, visit [Fivetran.com](https://fivetran.com).



Start your free trial

©2024 Fivetran Inc.