Fivetran

How Fivetran fuels Amazon S3 with Iceberg

Table of contents

Introduction 3

The current state of data lake management 4

Simplify your data lake management with Fivetran 6

Get started with Fivetran 8

Introduction

The cloud data lake has come a long way since its origins around 2015. Data lakes provide organizations with flexibility and agility when storing massive amounts of data — two characteristics that are critical to building a data-driven enterprise. But over the years, business use cases and users' expectations for the data lake have evolved, driving a need for modernization.

Many of the elements that once made data lakes a beloved storage vehicle now introduce challenges for the teams that use them. Without proper data curation and management, they can quickly become "data swamps" or "junk drawer" style storage, which — don't get us wrong — serve a purpose and for which there is a place and time. However, the need to modernize is increasing as more data compliance and governance issues arise and more teams want to get their hands on data lakes and run models off them.

As Gartner [1] says, it's also about adopting a thoughtful and thorough system:

"Data and analytics leaders adopting a data lake or modernizing older lakes as part of their data management solutions initiatives must:

- Design data lakes to capture the kind of data and metadata that are necessary for reliable analytical results delivered to a variety of data consumers.
- Design a data lake architecture that supports a multi-step, linear process for data acquisition, insight development and discovery, optimization and governance, and analytics consumption."

But how do you modernize? And how do you do so in a way that maintains what you need from a data lake in terms of volume and flexibility, while still tackling requirements like Europe's GDPR and CCPA? We'll walk you through one way to modernize your data lake with Fivetran fueling Amazon S3 with Iceberg.

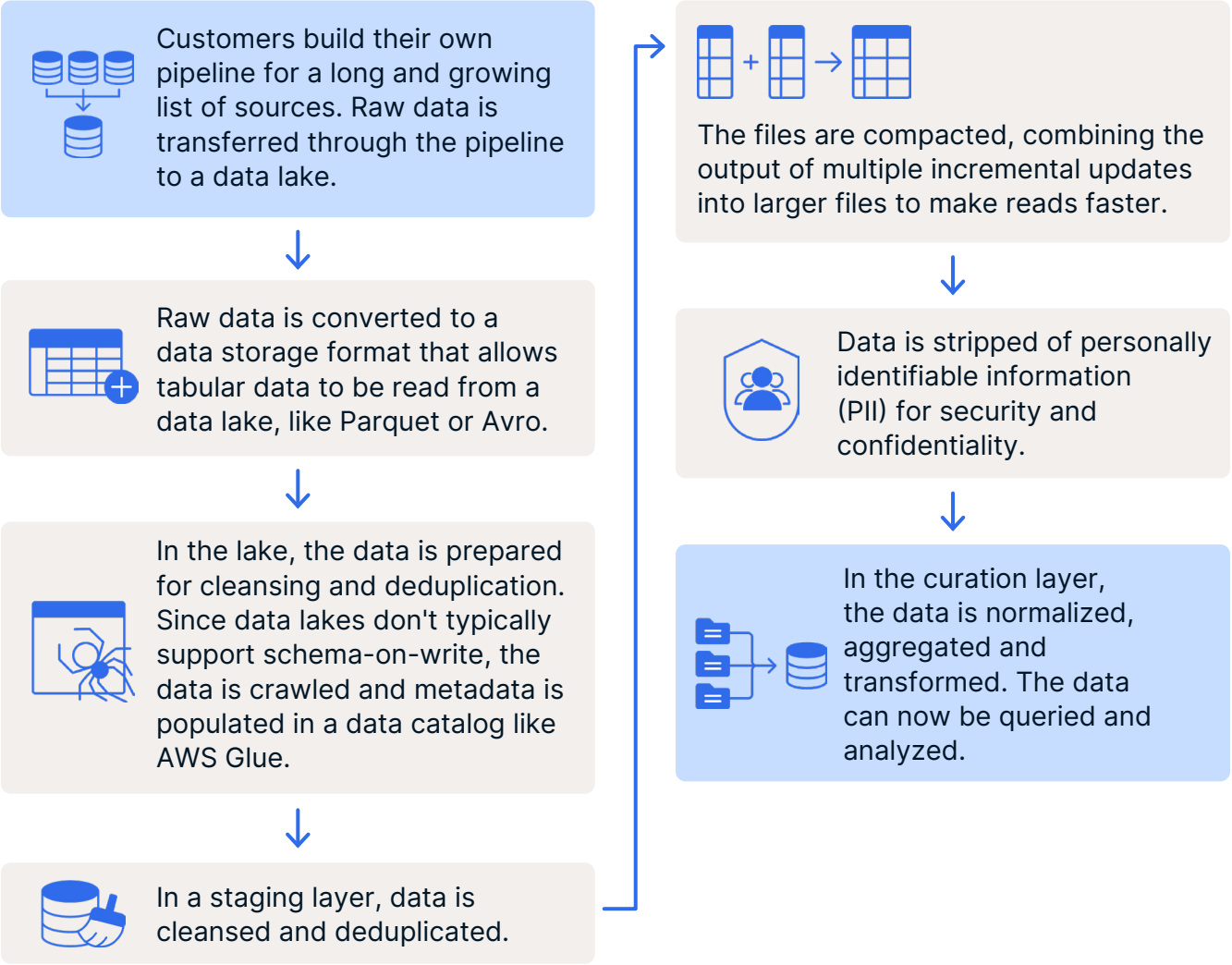
[1] Best Practices for Designing Your Data Lake. Gartner. May 2021.

The current state of data lake management

According to Gartner, the point of a data lake is that its “simplicity enables broad, flexible and unbiased data exploration and discovery.” What makes a data lake uniquely positioned to do this? As Gartner puts it, a data lake “preserves the original details of source data for the richest data exploration, discoveries and analytic correlations possible.”

Yet all of this flexibility and exploration comes at a price — namely the time and attention of data teams. Not to mention the expense of developing a team of engineers who need to focus on building and maintaining bespoke pipelines to fuel the data lake. If you are currently using a data lake, there’s a lot you’re responsible for.

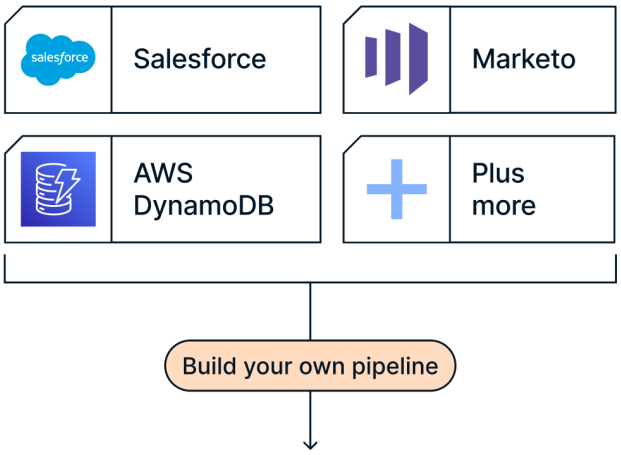
The manual method of maintaining a data lake



A visual overview of the current state

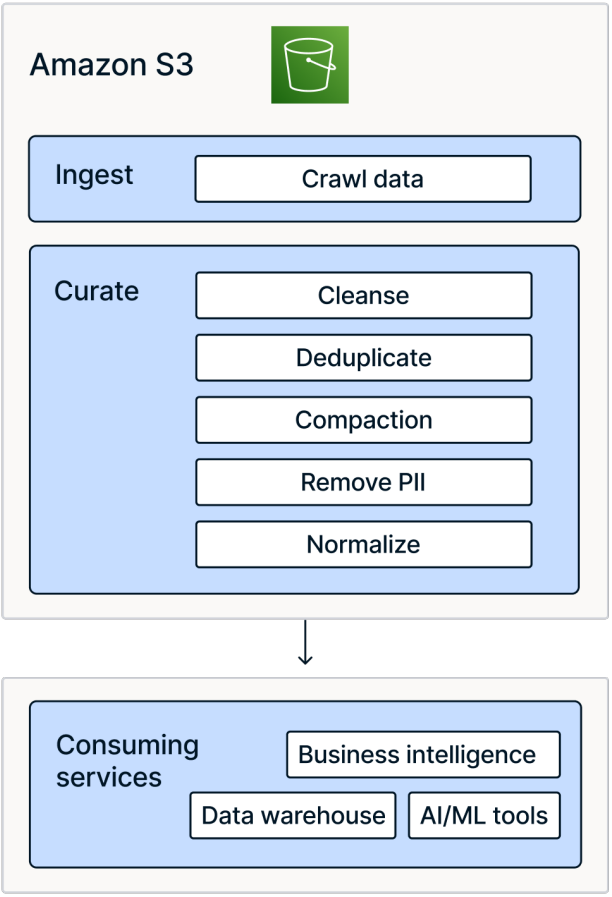
A data team must build its own ELT pipelines for a wide range of sources.

The pipeline must transfer data to the data lake with minimal processing. In the destination, an additional processing layer converts the data into queryable big data file formats like Parquet and Avro.



Without schema-on-write, the team must build a crawler to extract metadata to populate a data catalog like AWS Glue and prepare the data for cleansing and deduplication

The team must cleanse –including removing sensitive data– and deduplicate data for integrity and usability, then compact it, combining it into larger files to make reads faster.



The team must transform and model data as needed for consumption by BI platforms, data warehouses and AI/ML tools.

From manually building, re-building and maintaining pipelines for all of your sources to the work you need to do within the landing, staging and curation processes, you're spending hours of precious time just getting a data lake up and running. Only after all of this cleansing, compaction and normalizing can teams finally get to the important work of finding new and exciting insights. But if a pipeline breaks or you need to bring in a new data source, the tedious work begins again and the insights remain waiting until your team has the time.

Simplify your data lake management with Fivetran

Fivetran offers a pathway to simplifying data lake management:

Connectivity:
Easy integration with data sources with pre-built connectors.

Data management:
Automated cleansing, deduplication and normalization.

Compliance:
Ensures compliance with data protection standards, such as the GDPR and CCPA.

The Fivetran Managed Data Lake Service

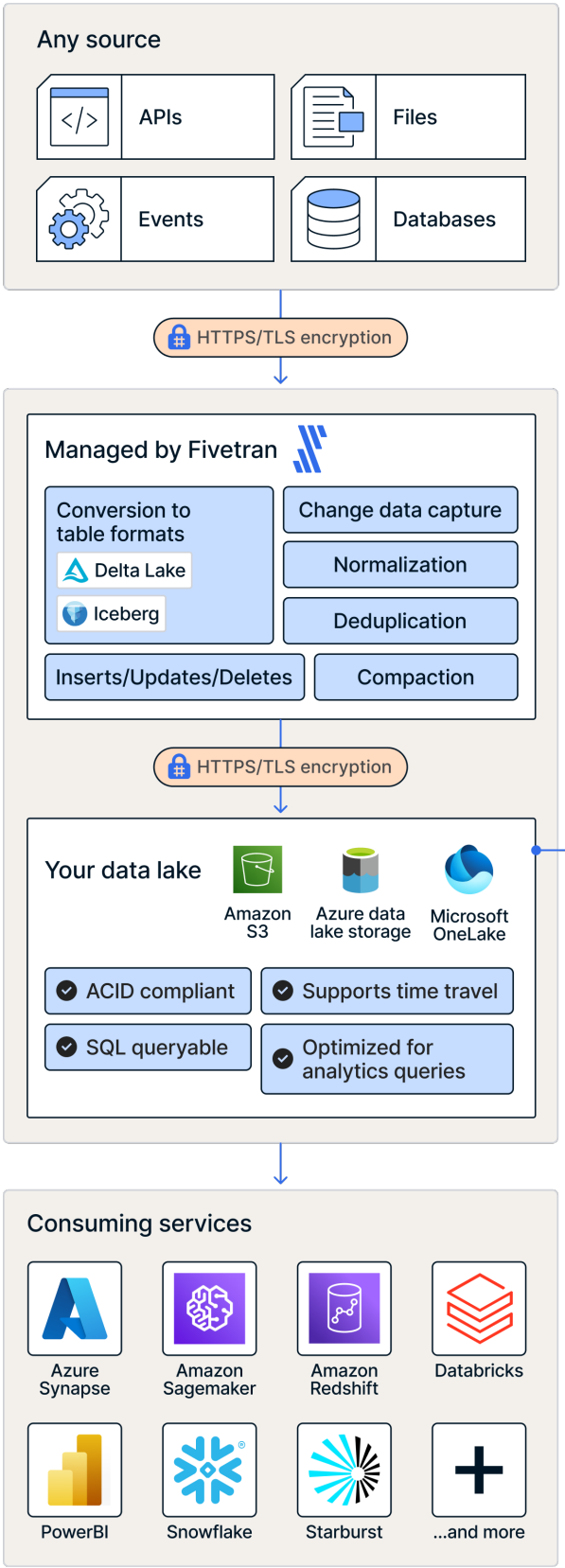
With the ability to land over 500 pre-built or custom data sources into any major data lake destination, Fivetran automatically normalizes, compacts and deduplicates your data before standardizing it in Iceberg or Delta Lake table format. From securely and reliably moving data into the data lake to continuous 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 and
- Strategically optimizes your data so that it's clean, cataloged and standardized in a query-ready table format for analytics.

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 in optimized Parquet file formats and Iceberg table structures, with metadata populated in the AWS Glue data catalog, enabling immediate query readiness.

Fivetran's Managed Data Lake Service



Fivetran automatically extracts data from sources. The customer must input credentials and set a schedule.

For an extra layer of security, customers can block or hash PII and other sensitive data before it even enters the pipeline.

Fivetran cleanses, deduplicates and normalizes the data before it lands in the lake. It is then written to Parquet files, then Iceberg tables.

Data catalog

Fivetran loads Iceberg tables to Amazon S3 and sends the metadata to Glue. The data team can now explore and transform its data.

AWS services such as Athena can natively consume the Iceberg table format as well.

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, PII detection, 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.
- Providing peace of mind with worry-free data replication that ensures your datasets always arrive clean and complete, with every change captured.

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.

Get started with Fivetran

Sign up for a 14-day free trial of Fivetran

Book a live demo →

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.

Start your free trial

©2024 Fivetran Inc.