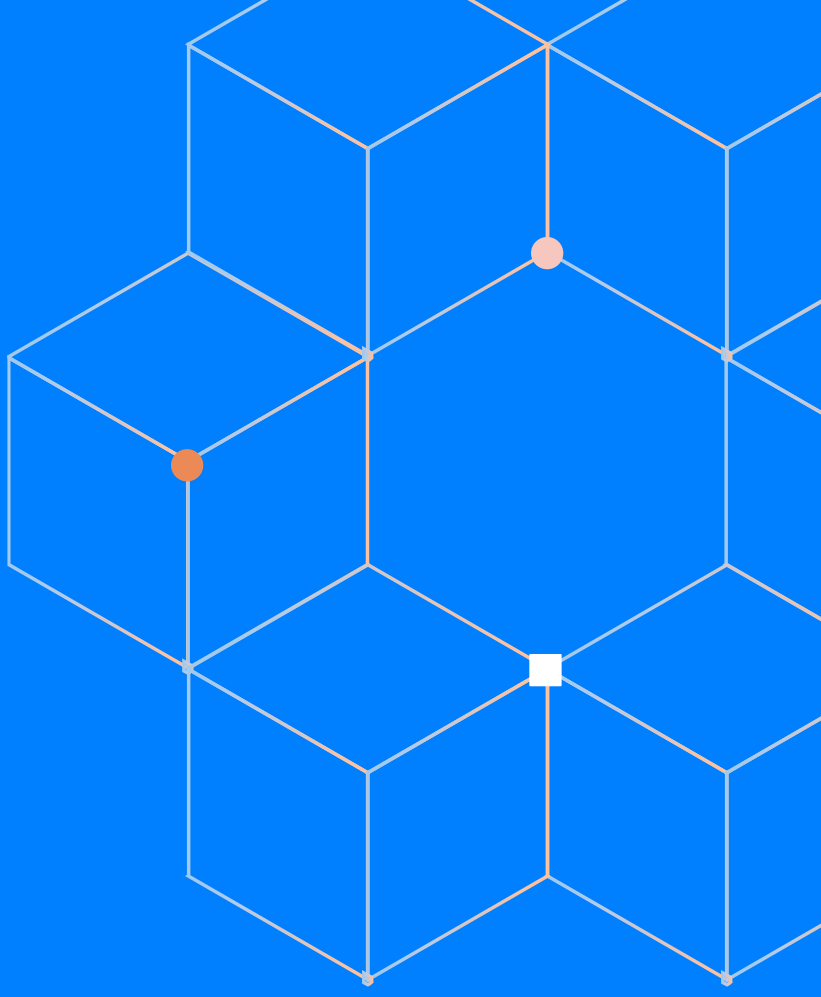


How Fivetran fuels Amazon S3 with Iceberg

Modernizing the data lake

Fivetran

Table of contents

01 Introduction 3

02 The current state of data lake managements 4

- The manual method of maintaining a data lake 4
- A visual overview of the current state 5

03 Modernizing your data lake 6

- The automated way to maintain a modern data lake 7
- A visual overview of the modern state 7

04 Conclusion: It's time to modernize 8

01

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.

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.

02

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."

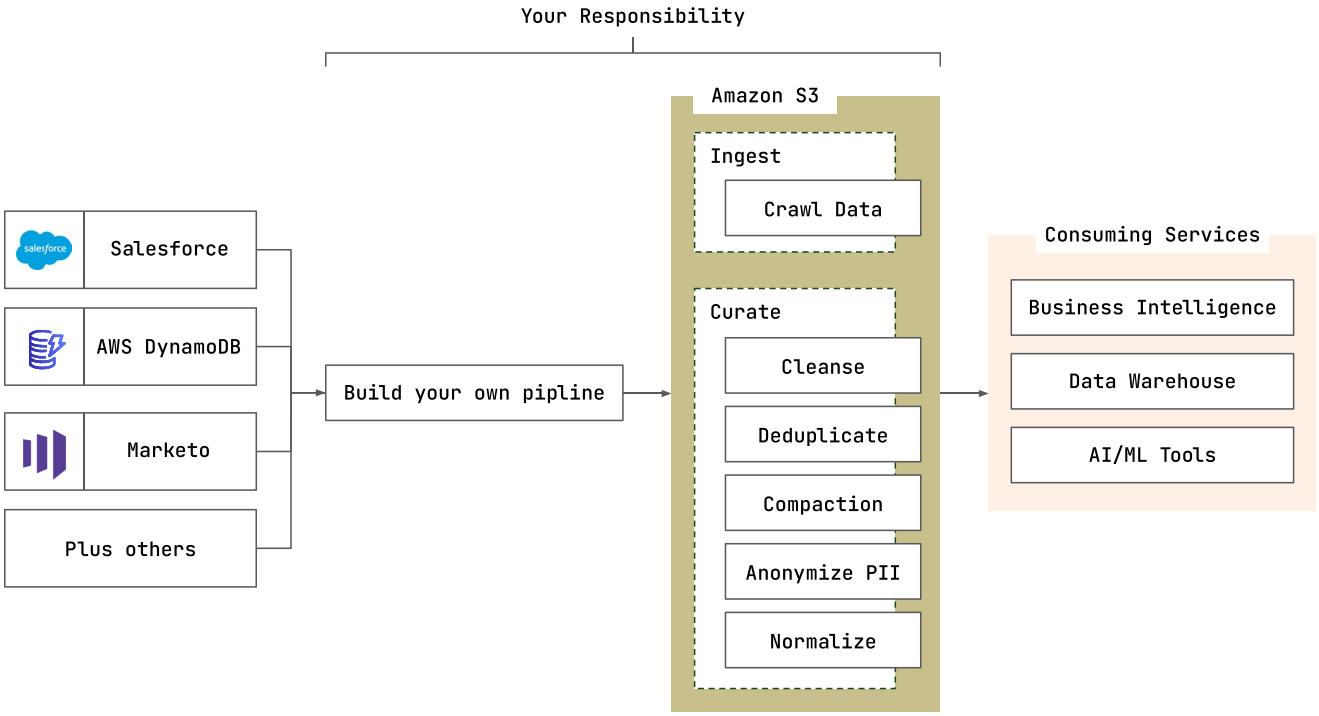
Yet all of this flexibility and exploration comes at a price — namely the time and attention of data teams. Not to mention the expense of developing a team of engineers **who need** to focus on building and maintaining bespoke pipelines to fuel the data lake. If you are currently using a data lake, there's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

The manual method of maintaining a data lake

- Typically, to get data into the data lake, customers are required to build their own pipeline for a variety of different sources, and that list only continues to grow.
- The data is transferred through the pipeline, in a raw format, to the lake. No stringent format or heavy transformations are required of the data that would make it favorable to customers who don't want to maintain load-intensive transformations in the pipeline.
- To be queried in a data lake, data has to be converted into a compatible format. Data lakes support several popular formats, including Parquet and Avro.
- Once it lands in the lake, the data has to be prepped for cleansing and deduplicating. Since the data lake typically doesn't contain a schema-on-write, the data is crawled and the metadata is populated in a data catalog like AWS Glue.
- In the staging layer, the data is cleansed and deduplicated to ensure clean data.
- The next step is to compact the files. Data in the lake, due to incremental updates, can result in several smaller files. Frequent compaction into larger files makes reads from the lake much faster.

- The data is then stripped of personally identifiable information (PII), removing any details that can identify the specific customer associated with it.
- In the curation layer, the data is normalized, aggregated and transformed for consumption by services like Amazon Redshift, Amazon EMR and Amazon Athena.

A visual overview of the current state



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.

03

Modernizing your data lake

Look, we know data lakes. We love data lakes. But clearly the current iteration is having trouble serving a wider range of users and their needs.

First of all, the way most folks use their data lakes, it's tough to satisfy compliance regulations found in the General Data Protection Regulation (GDPR) and similar privacy regulations adopted in places like California and Canada. As a part of GDPR and other laws like it, people have a "right to be forgotten." This means, if they want their data deleted, you have to comply or else you may face costly fines. If that data is stored in a data lake, you effectively have to pull everything out, find that one piece of data, delete it and put everything back in. And all of this is done manually, with custom scripts, and with a lot of time from your engineering team.

What's more, as data lakes have become more multi-tenant (serving more user types and use cases), it's time we expect a data lake to operate and look like all our other trusted destinations. In a word, it's time to modernize. What does this mean? It means creating a data lake environment where your data is automatically cleaned, prepped and deposited — with less effort on your behalf. The kind of experience you've come to enjoy with a data warehouse. With a modern data lake, more users can access and make use of your data, decreasing time to insight and time spent on maintenance of the lake. This allows you to spend more time discovering new and exciting things within it.

Add to this the ability to find more in your data lake faster. While data lakes are revered for their ability to hold an immense amount of unstructured data, they can quickly become like bottomless pits where you can't easily find what you need — or it takes a whole team to get that data ready to go.

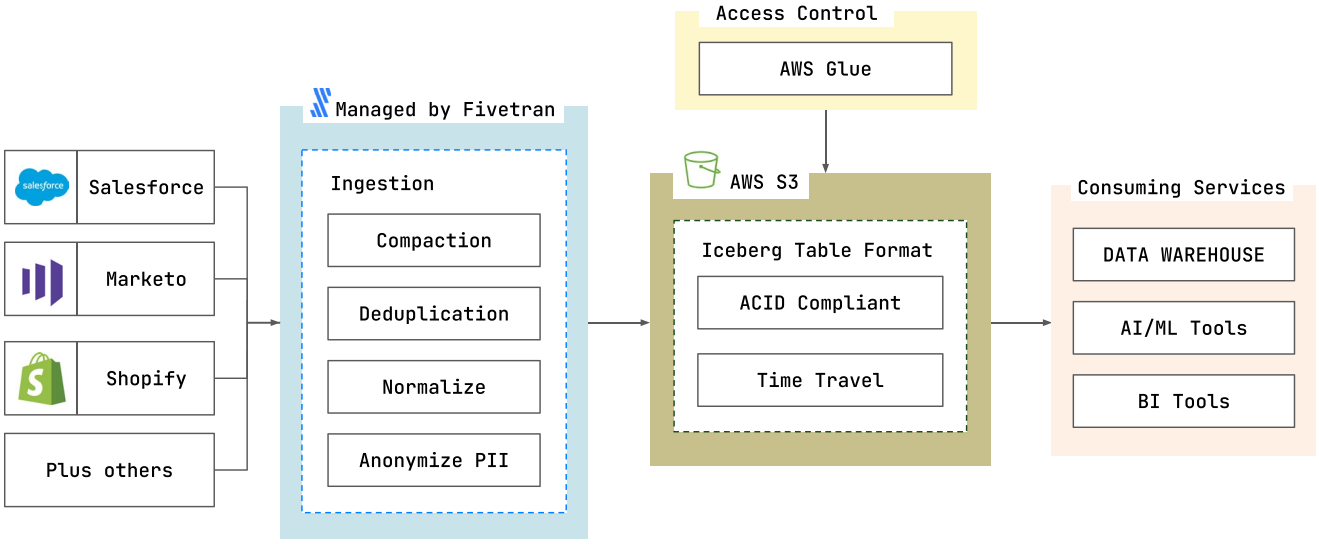
A modern data lake brings together the compliance and ease of use you get in a data warehouse, the large and inexpensive storage that you love in a data lake, and the effortless data integration you expect from an automated ELT vendor.

The automated way to maintain a modern data lake

With Fivetran, Amazon S3 and Iceberg, the current process we reviewed previously becomes a lot more automated:

- Data is extracted from sources using our native connectors, with no work needed from you except for authorization credential input. Fivetran natively supports hundreds of different data sources that can be moved to the data lake.
- Customers can block or hash their PII before it even enters the pipeline, making data even more secure.
- Within Fivetran, before the data lands in the lake, the data is cleansed, deduplicated and normalized.
- The data is written to Parquet files in our Fivetran pipeline and using the Iceberg table format, and we insert that data into Amazon S3 with the metadata immediately populated into Glue.
- Your data is then ready for immediate querying or, depending on your needs, aggregation and transformation.
- Iceberg table format allows the data to be natively consumed by AWS services such as Athena.

A visual overview of the modern state



Thanks to the automation provided by Fivetran, Amazon S3 and Iceberg, data engineers no longer need to spend hours on painstaking pipeline work or preparing the data to be analyzed. Instead, data lands immediately, ready to be used. Your manual responsibility shrinks dramatically. Now, you can focus on much more interesting work — finding new insights from your data lake that will transform your business.

04

Conclusion: It's time to modernize

Modernizing your data lake isn't just about automation — though that will be a key component. 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."[1]

With this call to modernize, Fivetran pioneered a new solution.

This modernization unlocks time and resources for your teams insofar as they will no longer have to be singularly focused on the creation and maintenance of pipelines that feed your lake or on all the data preparation necessary to use the massive amounts of data found therein. Instead, through the automation provided by technologies like Fivetran, Amazon S3 and Iceberg, teams and companies can begin to enjoy a data lake that is more accessible to more people, easier and faster to use and get value out of, and capable of transforming more businesses into agile competitors.

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

Fivetran is the global leader in modern data integration. Our mission is to make access to data as simple and reliable as electricity. Built for the cloud, Fivetran enables data teams to effortlessly centralize and transform data from hundreds of SaaS and on-prem data sources into high-performance cloud destinations. Fast-moving startups to the world's largest companies use Fivetran to accelerate modern analytics and operational efficiency, fueling data-driven business growth. For more info, visit Fivetran.com.