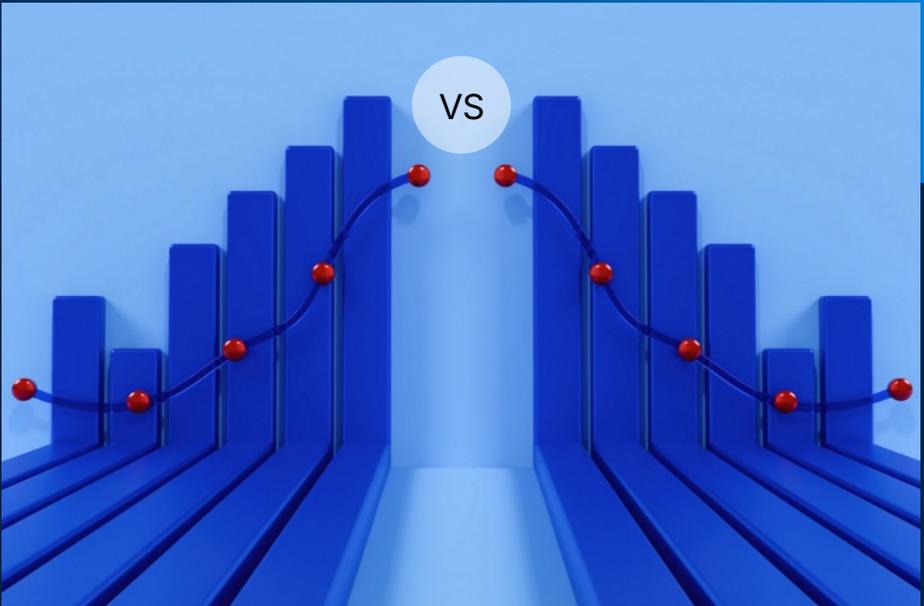


TCO ANALYSIS

Ingestion Costs Compared: Fivetran Managed Data Lake vs. Data Warehouse

William McKnight and Jake Dolezal

DATA, ANALYTICS & AI





Ingestion Costs Compared:

Fivetran Managed Data Lake vs. Data Warehouse

William McKnight and Jake Dolezal

Table of Contents

01	Executive Summary	2
02	Introduction	4
03	Field Study and Results	13
04	Conclusion.....	22
05	Appendix	23
06	Disclaimer	25
	About Fivetran	25
	About William McKnight	26
	About Jake Dolezal.....	27
	About GigaOm	28

01

Executive Summary

This GigaOm TCO report is commissioned by Fivetran

THIS REPORT PRESENTS A COMPREHENSIVE ANALYSIS of the cost and performance benefits of using traditional cloud data warehouse solutions (Snowflake, Databricks, and Amazon Redshift) versus a modern data lake approach using Amazon S3 and Azure Data Lake Storage (ADLS) for data storage and Fivetran Managed Data Lake Service for metadata management, open table format conversion, and automated table maintenance.

Incremental data ingestion is a significant cost driver of a data architecture, underscoring the importance of optimizing data integration and storage strategies.

We saw a substantial reduction in these data ingest cost queries when using the Fivetran Managed Data Lake Service, which incurs the cost of ingest compute when writing to a data lake. Initial sync costs were not included in our calculations.

Key Findings:

- **Significant cost savings:** The data lake approach yields substantial cost savings, reducing total costs from 77% to 95% as compared to traditional data warehouse solutions.
- **Lower incremental sync costs:** The modern data lake approach incurs lower incremental sync compute costs in our test, ranging from \$0.63 to \$2.82 compared to \$6.13 to \$14.17 per incremental sync for data warehouse solutions.
- **Slightly slower incremental sync times:** The modern data lake approach exhibits slightly slower incremental sync times, with an 8-10% increase in sync time compared to data warehouse solutions.

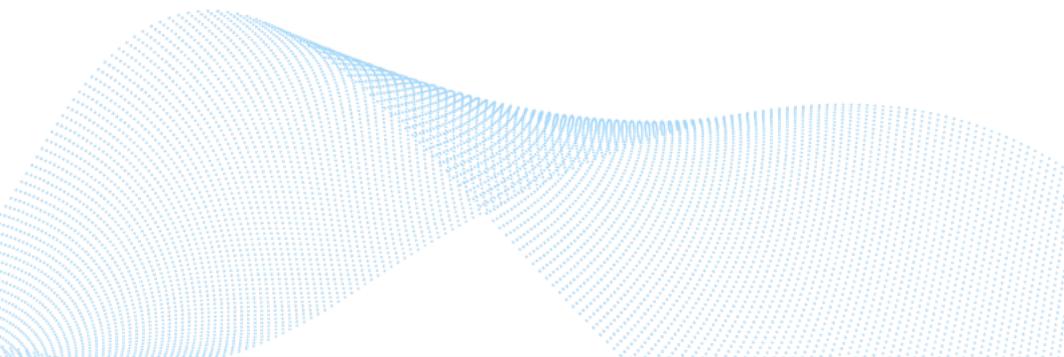
As businesses continue to leverage AI-driven workloads, the anticipated exponential growth in data volumes is expected to significantly increase data management costs. To mitigate this expense, companies should explore opportunities to decrease their overall data management spend.

The modern data lake approach provides a cost-effective and scalable solution for incremental data syncing and storage, making it an attractive alternative to traditional data warehouse solutions. Although it may incur slightly longer incremental sync times, the cost savings make data lakes a viable option for organizations seeking to optimize their data management infrastructure.

The modern data lake approach provides several key benefits, including reduced data management expenses, economies of scale to easily accommodate growing data volumes, and flexibility to support a wide range of data formats and structures in one location. This approach is also built on top of open standards, which helps minimize vendor lock-in and gives organizations the flexibility to use a broader range of tools—choosing whatever best fits their overall tech stack. By adopting a modern data lake approach, businesses can effectively manage increasing data volumes needed for reporting and AI while minimizing costs, ultimately gaining a competitive edge. This helps a company start slow and grow their data platform and architecture as needed.

AI has accelerated modernization projects, but economic uncertainty may impact progress. Companies under pressure may prioritize high-ROI projects, and data readiness initiatives without a clear link to revenue or cost savings may be delayed.

Modern data lakes should be included in an enterprise data architecture to reduce costs and provide a foundation for AI success. By adopting the Fivetran Managed Data Lake Service, organizations should see significant compute cost savings as the ingest query cost is absorbed by Fivetran and automation reduces manual effort.



02

Introduction

DATA LAKES HAVE GROWN IN CAPABILITY thanks to innovative features like open table formats and cataloging for better governance, thus preventing the historical problem of lakes turning into swamps. The modern data lake architecture is thought to be the most scalable, cost-effective, and performant data foundation to support analytics and AI workloads, so we put it to the test. Using open table formats like Apache Iceberg and Delta Lake, organizations can ensure data quality and consistency within their data lakes. These advances help streamline data processing and analysis, ultimately leading to more accurate insights and decision making, and enable new architectures to power AI workloads.

Using cataloging tools also allows for easier data discovery and management within the data lake, further enhancing its efficiency and usability. With a flexible design, you can use multiple catalogs simultaneously, selecting the one that integrates most effectively with your preferred query engines. Overall, these advancements in data lake architecture are revolutionizing the way organizations handle and leverage their data assets.

We looked at the potential benefits of a modern data lake solution as compared to a data warehouse. Here are the key advantages of a modern data lake solution:

- **Universal data storage layer:** Provides a single layer for all data storage, reducing duplicate data and governance concerns.
- **Scalability:** Easily accommodates large volumes of data without extensive restructuring or reformatting.
- **Flexibility and cost-effectiveness:** Offers a more flexible and cost-effective solution when compared to traditional data warehouses.
- **Increased interoperability:** Unbundles compute and storage, reducing vendor lock-in and increasing interoperability.

Fivetran is the data movement platform with the most comprehensive support for data lakes, so it made sense to explore the difference in ingestion costs when using Fivetran to move data to a data lake instead of a data warehouse destination.

Once the data is delivered to its storage location, the data architect can choose how the data will be refined. Some data architects will choose to perform the first levels of refinement in a data lake and then perform the final, curated (or gold) layer in a data warehouse. Others will choose to perform all levels of refinement in a data lake. This benchmark includes both refinement patterns.

Using a data lake requires a separate query engine, which could be a data warehouse (i.e. Databricks, Snowflake, Amazon Redshift) or another purpose-built query engine (i.e. Dremio, Presto). It is important to consider the organization's specific needs and goals when deciding between a modern data lake solution and a traditional data warehouse approach. By evaluating factors such as scalability, performance, and cost, businesses can determine the architectural design that best aligns with their overall data strategy and objectives.

Testing Methods

The GigaOm field test was designed to understand the performance and cost difference between using the Fivetran Managed Data Lake Service to land data into a data lake compared to loading data directly into a cloud data warehouse or lakehouse. The field test covers two data lake use cases:



Test A

Uses the Fivetran Managed Data Lake Service to load data into a data lake destination as primary storage for the data, instead of the cloud data warehouse.



Test B

Uses the Fivetran Managed Data Lake Service to load data into a data lake destination for staging and preprocessing ("bronze/silver" or "staging" layer) with an analytics/BI (or "gold/curated" layer) maintained on the data warehouse.

Note that the bronze, silver, and gold designations refer to a medallion architecture. This decentralized data management architecture organizes data into raw (bronze), processed (silver), and curated (gold) tiers, enabling flexible and scalable data management and analytics.

In both test cases, we compared the cost and performance differences of using the Fivetran Managed Data Lake Service to load data into a data lake to using Fivetran to load data into a cloud data warehouse. The cloud data warehouse is still necessary as the query engine in both cases, utilizing the data warehouse's recommended data catalog integration method.

The working hypothesis is that a data lake reduces compute costs when compared to a data warehouse because it reduces the warehouse ingest query compute consumption. However, we also want to test the potential performance impacts of using a data lake. The following diagram presents the conceptual architecture of our test platforms.

Figures 1 and 2 show the two architecture approaches tested in this benchmark.

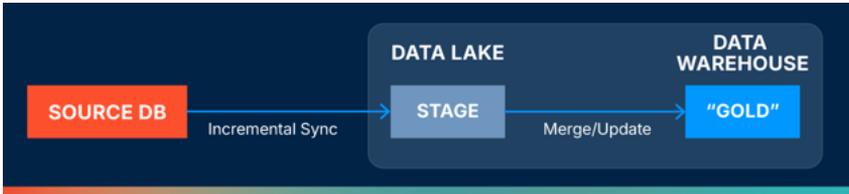


Figure 1. Architecture Without a Data Lake

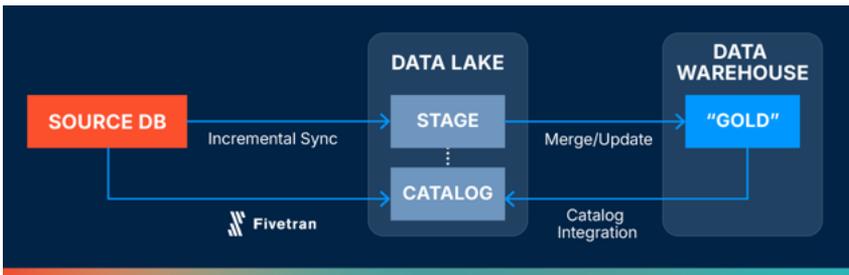


Figure 2. Architecture with a Data Lake

For the sake of simplicity, we are not testing transformations because every use case would have different requirements and parameters. The most straightforward apples-to-apples test case is using the compute to move data to one or multiple locations; i.e., source to staging to destination with no transformation. In this test, the gold layer is a predetermined subset of the staged data.

Platforms Under Test

We performed our tests across three popular data warehouse/lakehouse platforms (Snowflake, Databricks, and Amazon Redshift) and object storage provided by the two major public cloud platforms (Amazon S3 and Azure Data Lake Storage). We performed additional tests across the two predominant open table formats (Apache Iceberg and Delta Lake).

The most straightforward apples-to-apples test case is using the compute to move data to one or multiple locations; i.e., source to staging to destination with no transformation.

Table 1. S3 and Snowflake Test 1A

	DATA LAKE	DATA WAREHOUSE
Data Source	TPC-DS	TPC-DS
Data Integration	Fivetran	Fivetran
Storage	S3 (Iceberg) + Fivetran Managed Data Lake Service	Snowflake
Query Engine	Snowflake	Snowflake

Source: GigaOm 2026

Table 2. S3 and Snowflake Test 1B

	DATA LAKE (STAGING)	DATA WAREHOUSE
Data Source	TPC-DS	TPC-DS
Data Integration	Fivetran	Fivetran
Silver/Staging Layer	S3 (Iceberg) + Fivetran Managed Data Lake Service	Snowflake
Gold Layer/Data Mart	Snowflake	Snowflake
Query Engine	Snowflake	Snowflake

Source: GigaOm 2026

Table 3. ADLS and Databricks Test 2A

	DATA LAKE	DATA WAREHOUSE
Data Source	TPC-DS	TPC-DS
Data Integration	Fivetran	Fivetran
Storage	ADLS (Delta Lake) + Fivetran Managed Data Lake Service	Databricks
Query Engine	Databricks	Databricks

Source: GigaOm 2026

Table 4. ADLS and Databricks Test 2B

	DATA LAKE (STAGING)	DATA WAREHOUSE
Data Source	TPC-DS	TPC-DS
Data Integration	Fivetran	Fivetran
Silver/Staging Layer	ADLS (Delta Lake) + Fivetran Managed Data Lake Service	Databricks
Gold Layer/Data Mart	Databricks	Databricks
Query Engine	Databricks	Databricks

Source: GigaOm 2026

Table 5. S3 and Redshift Test 3A

	DATA LAKE	DATA WAREHOUSE
Data Source	TPC-DS	TPC-DS
Data Integration	Fivetran	Fivetran
Storage	S3 (Iceberg) + Fivetran Managed Data Lake Service	Redshift
Query Engine	Redshift	Redshift

Source: GigaOm 2026

Table 6. S3 and Redshift Test 3B

	DATA LAKE (STAGING)	DATA WAREHOUSE
Data Source	TPC-DS	TPC-DS
Data Integration	Fivetran	Fivetran
Silver/Staging Layer	S3 (Iceberg) + Fivetran Managed Data Lake Service	Redshift
Gold Layer/Data Mart	Redshift	Redshift
Query Engine	Redshift	Redshift

Source: GigaOm 2026

Source Data

For our field test, we used a TPC-DS data set. TPC-DS (Transaction Processing Performance Council - Decision Support) is a benchmark data set designed to evaluate the performance of decision support systems, such as data warehouses and analytics platforms. It includes a schema with multiple fact and dimension (star-schema) tables representing a retail business, featuring sales, inventory, customer, and product data. This data set simulates real-world analytical situations.

The TPC-DS data generator has a scale factor parameter that determines the size of the physical, uncompressed, on-disk raw data that it generates. We used a scale factor of 3,000, which generated 3 TB of data. We loaded this data into a cloud-managed AWS Relational Database Service (RDS) for PostgreSQL.

The TPC-DS data set has timestamped transactions. Its data generator produces approximately five calendar years worth of data, regardless of the scale factor you choose. For the purposes of our testing, we predetermined our "gold" layer to contain the latest two years worth of data, which works out to approximately 40% of the fact table data. We did not utilize the query side of TPC-DS.

Table 7. Data Set Summary

Data set	TPC-DS
Data volume	3 TB
Fact tables	6
Total rows (fact tables)	18,224,793,568
Gold layer rows	7,289,917,427 (total rows x 40%)

Source: GigaOm 2026

Initial Data Sync

We conducted the initial data sync from the source to the destination data warehouse or managed data lake with Fivetran. We performed an initial sync of the entire 3 TB database first to bring an exact copy of the data to our "stage" layer. We configured our source PostgreSQL databases to sync via logical replication. Fivetran reads the write-ahead log (WAL) using a logical replication slot and a publication to detect new, changed, or deleted data. Once the data sync was complete, the field test began.

Incremental Data Updates and Inserts

To simulate daily database transactions, we updated a single column within a sample of 2% of each fact table's rows reserved for the 40% "gold" layer. We inserted new rows into each fact table equal to 1% of the 40% "gold" layer. We calculate that four of these incremental changes occur per day for an incremental sync every six hours.

Table 8. Incremental Data Summary

Total rows updated	145,798,349 (gold layer x 2%)
Total rows inserted	72,899,174 (gold layer x 1%)
Total changed rows	218,697,523 (sum of above)
Increments per day	4
Gold layer rows	7,289,917,427 (total rows x 40%)

Source: GigaOm 2026

Incremental Syncs

Every six hours, after we completed the incremental updates and inserts, Fivetran automatically syncs the changes to the destination (data warehouse or data lake). We tracked the consumption costs incurred during each incremental sync by the data warehouse or the cloud object storage being tested. We used their billing metrics queries or APIs to determine the exact amount of usage and isolated the test activity, so there would be no noise from other billable activities. We also tracked incremental sync performance to either destination (data warehouse or data lake) to determine which took longer and by how much.

Gold Layer Merge/Update

After completing the incremental sync, we updated the gold layer in the data warehouse with the newly changed data using a SQL MERGE statement to update changed rows and insert new rows. We tracked the data warehouse costs of this step the same way we tracked the incremental sync, ensuring isolation as well as performance. In the case of the “staging” layer on the data warehouse, we tracked the performance of the native table to native table merge. In the case of the “staging” layer on the managed data lake, we used the data warehouse's preferred method of connecting to Iceberg and Delta Lake tables (Snowflake catalog integration, Redshift Spectrum, etc.) to perform the merge. We were interested to see if there was a significant performance difference with these methods.

Total Cost Differential Calculation

We calculated the total cost differential between directly loading to a data warehouse and to a data lake with the Fivetran Managed Data Lake Service. This is not the total cost of ownership. The cost of using Fivetran is excluded because their pricing model is based on the number of monthly active rows (MAR). In both cases, the number of rows synced (inserted or updated) were exactly the same, so the cost for Fivetran would be identical. We

are more interested in finding the differential in our data warehouse and data lake usage costs. Thus, if one method is significantly less expensive but the performance is acceptably equivalent, then we have a compelling argument for adopting one method over the other.

We did not factor in the initial data sync because Fivetran does not bill for that initial sync of a new destination. Also, the data warehouse or data lake consumption was nominal and one-time. We calculated the costs from the daily operations—incremental syncs four times a day for a 365-day calendar year.

Each incremental sync and gold layer merge only has either data lake costs or data warehouse plus data lake costs, depending on the test. We also used the Fivetran recommended size for each data warehouse tested (**Table 9**). Note that RPU stands for Redshift processing unit and DBU for Databricks unit.

Table 9. Data Warehouse Sizes and Costs

PLATFORM	SIZE	COST UNITS	COST
Snowflake	X-Small	1 Credit	\$3.00 per hour
Databricks	2X-Small	4 DBU	\$2.80 per hour
Redshift	Serverless	8 RPU	\$2.88 per hour

Source: GigaOm 2026

Table 10 shows the computed cost by approach. Note that storage costs are included here for reference, but are not included as part of our cost savings calculation.

Table 10. Compute Cost Table

TEST AND CLOUD	DATA WAREHOUSE/DATA LAKE PLATFORM			OBJECT STORAGE PLATFORM		
	DATA PLATFORM	COST UNIT	COST	STORAGE PLATFORM	COST UNIT	COST
Test 1: AWS	Snowflake	Credits/hour	\$3.00	S3	1,00 PUTs 10,000 GETs	\$0.005 \$0.004
Test 2: Azure	Databricks	DBU	\$0.70	ADLS	10,000 writes 10,000 reads	\$0.065 \$0.005
Test 3: AWS	Redshift	RPU	\$0.36	S3	1,000 PUTs 10,000 GETs	\$0.005 \$0.004

Source: GigaOm 2026

Table 11. Storage Cost Table

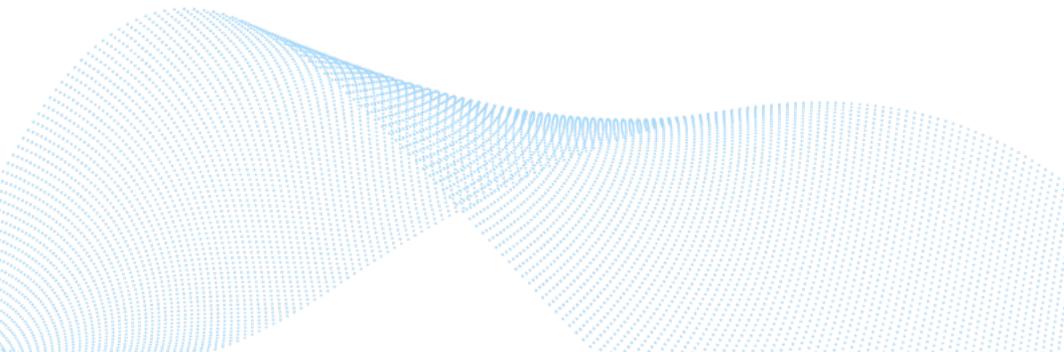
TEST AND CLOUD	DATA WAREHOUSE/DATA LAKE PLATFORM		OBJECT STORAGE PLATFORM	
	DATA PLATFORM	COST	STORAGE PLATFORM	COST
Test 1: AWS	Snowflake	\$23.00	S3	\$23.00
Test 2: Azure	Databricks	\$26.00	ADLS	\$19.00
Test 3: AWS	Redshift	\$24.00	S3	\$23.00

Source: GigaOm 2026

Our cost calculations did not include egress costs, because all resources were hosted in the same region. To calculate the one-year cost differentials, we used the following formulas:

- **Increment Cost** = Incremental Sync Cost + Incremental Gold Merge Cost
- **Annual Cost** = Increment Cost x 4 times/day x 365 days
- **Total Cost Differential** = Data Warehouse Destination Total Annual Cost – Managed Data Lake Stage Destination Total Annual Cost

Fivetran's fully managed approach to data movement lets users focus on higher impact projects and initiatives.



03 | Field Study and Results

S3+Fivetran Managed Data Lake Service vs Snowflake

THIS ANALYSIS (TABLE 12) COMPARES the cost and performance differences of using Snowflake versus a modern data lake (S3 with Iceberg and Fivetran Managed Data Lake Service) for incremental data syncing. The results show significant cost savings using the Fivetran Managed Data Lake Service loading data to a data lake, ranging from 80% (medallion architecture) to 93% (2-layer architecture) reduction in total costs compared to the Data Warehouse approach. The incremental sync cost is also lower with the data lake approach, ranging from \$0.63 to \$1.90 per increment, compared to \$8.51 to \$9.70 per increment for the Data Warehouse approach.

Fivetran's fully managed approach to data movement lets users focus on higher impact projects and initiatives. Says one Fivetran and Amazon S3 customer: "Without Fivetran, managing the data lake would have required five or six full-time engineers."

This customer saw lower compute costs moving raw data to S3 with Iceberg, reducing Snowflake usage and saving an estimated \$100,000 per year.

While the data lake approach has slightly slower incremental sync times, with a 9% increase in sync time compared to the data warehouse approach (2:22:24 versus 2:35:14 at incremental volume of 218 million rows), it should still meet performance requirements for most organizations.

Table 12. Cost Breakout with Snowflake and S3 (Test 1A and Test 1B)

DESTINATION	TEST 1A		TEST 1B	
	DATA WAREHOUSE	DATA LAKE	DATA WAREHOUSE	DATA LAKE
Data Integration	Fivetran	Fivetran	Fivetran	Fivetran
Silver Layer	Snowflake	S3 (Iceberg)	Snowflake	S3 (Iceberg)
Gold Layer	n/a	n/a	Snowflake	Snowflake
Query Engine	Snowflake	Snowflake	Snowflake	Snowflake

	TEST 1A		TEST 1B	
INCREMENTAL SYNC	DATA WAREHOUSE	DATA LAKE	DATA WAREHOUSE	DATA LAKE
Resource	Snowflake	S3	Snowflake	S3
Type	X-Small	N/A	X-Small	N/A
Write Usage	2.836	85.893	2.836	85.893
Write Usage Units	Credits	1,000 PUTs	Credits	1,000 PUTs
Write Rate	\$3.00	\$0.005	\$3.00	\$0.005
Read Usage		51.259		51.259
Read Usage Units		10,000 GETs		10,000 GETs
Read Rate		\$0.004		\$0.004
Read + Write Cost	\$8.51	\$0.63	\$8.51	\$0.63
INCREMENTAL GOLD UPDATE	DATA WAREHOUSE	DATA LAKE	DATA WAREHOUSE	DATA LAKE
Resource	n/a	n/a	Snowflake	Snowflake
Type			X-Small	X-Small
Write Usage			0.397	0.422
Write Usage Units			Credits	Credits
Write Rate			\$3.00	\$3.00
Write Cost			\$1.19	\$1.27
PER INCREMENT COST	\$8.51	\$0.63	\$9.70	\$1.90

Source: GigaOm 2026

Figure 3 offers a visual comparison of the per-increment cost of a data warehouse approach and a data lake approach. We compare the cost of utilizing the Fivetran Managed Data Lake Service to load data directly into a data lake instead of using the cloud data warehouse.

We also offer a cost comparison for using the Fivetran Managed Data Lake Service to load data into a data lake destination (instead of the cloud data warehouse) as primary storage for the data (Test 1A), and using the service to load data into a data lake destination for staging and preprocessing maintained on the data warehouse (Test 1B).

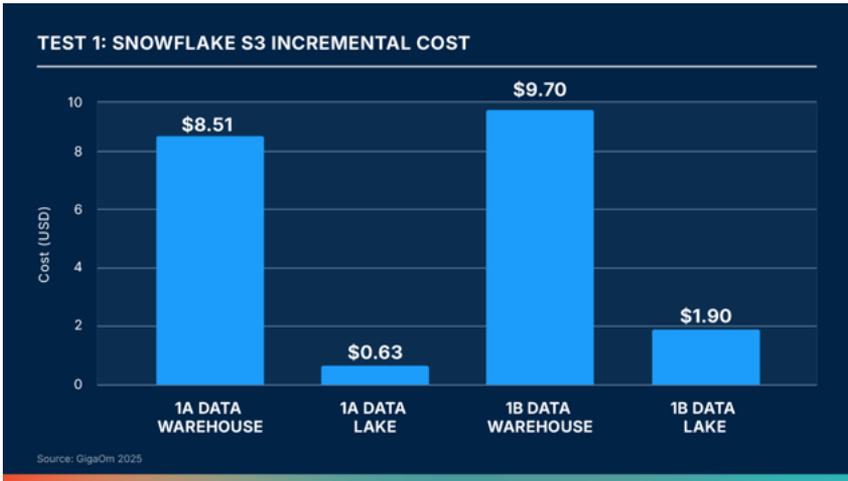


Figure 3. Total Incremental Cost with Snowflake and S3

Figure 4 applies the per-increment cost over the posited annual use case, where 1,460 updates are executed over 365 days. This provides a look at the scale of the spend for each approach.

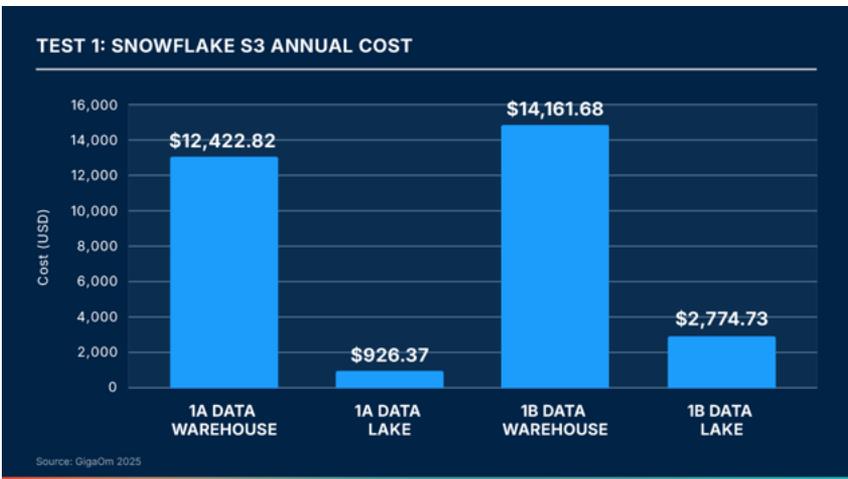


Figure 4. Total Annual Cost with Snowflake and S3

ADLS+Fivetran Managed Data Lake Service vs Databricks

This analysis compares the cost and performance differences of using Databricks versus a modern data lake (ADLS with Delta Lake and the Fivetran Managed Data Lake Service) for incremental data syncing. The results show significant cost savings with the data lake approach, ranging from a 95% (2-layer architecture) to 80% (medallion architecture) reduction in total costs compared to the data warehouse approach.

The incremental sync cost is lower with the data lake approach, ranging from \$0.65 to \$2.82 per increment, compared to \$12.12 to \$14.17 per increment for the data warehouse approach. See **Table 13**.

While the data lake approach has slightly slower incremental sync times, with a 9% to 10% increase in sync time compared to the data warehouse approach, it should still meet performance requirements for most organizations.



The results show significant cost savings with the data lake approach, ranging from a 95% to 80% reduction in total costs compared to the data warehouse approach.

Table 13: Cost Breakout with Databricks and ADLS

	TEST 1A		TEST 1B	
DESTINATION	DATA WAREHOUSE	DATA LAKE	DATA WAREHOUSE	DATA LAKE
Data Integration	Fivetran	Fivetran	Fivetran	Fivetran
Silver Layer	Databricks	ADLS	Databricks	ADLS
Gold Layer	n/a	n/a	Databricks	Databricks
Query Engine	Databricks	Databricks	Databricks	Databricks
INCREMENTAL SYNC	DATA WAREHOUSE	DATA LAKE	DATA WAREHOUSE	DATA LAKE
Resource	Databricks	ADLS	Databricks	ADLS
Type	2X-Small	N/A	2X-Small	N/A
Write Usage	17.317	5.467	17.317	5.467
Write Usage Units	DBU	10,000 Writes (4MB)	DBU	10,000 Writes (4MB)
Write Rate	\$0.70	\$0.065	\$0.70	\$0.065
Read Usage		56.174		56.174
Read Usage Units		10,000 Reads (4MB)		10,000 Reads (4MB)
Read Rate		\$0.005		\$0.005
Read + Write Cost	\$12.12	\$0.65	\$12.12	\$0.65
INCREMENTAL GOLD UPDATE	DATA WAREHOUSE	DATA LAKE	DATA WAREHOUSE	DATA LAKE
Resource	n/a	n/a	Databricks	Databricks
Type			2X-Small	2X-Small
Write Usage			2.922	3.104
Write Usage Units			DBU	DBU
Write Rate			\$0.70	\$0.70
Write Cost			\$2.05	\$2.17
PER INCREMENT COST	\$12.12	\$0.65	\$14.17	\$2.82

Source: GigaOm 2026

Figure 5 shows the per-increment cost for each tested scenario, while Figure 6 extrapolates this cost over a full year.

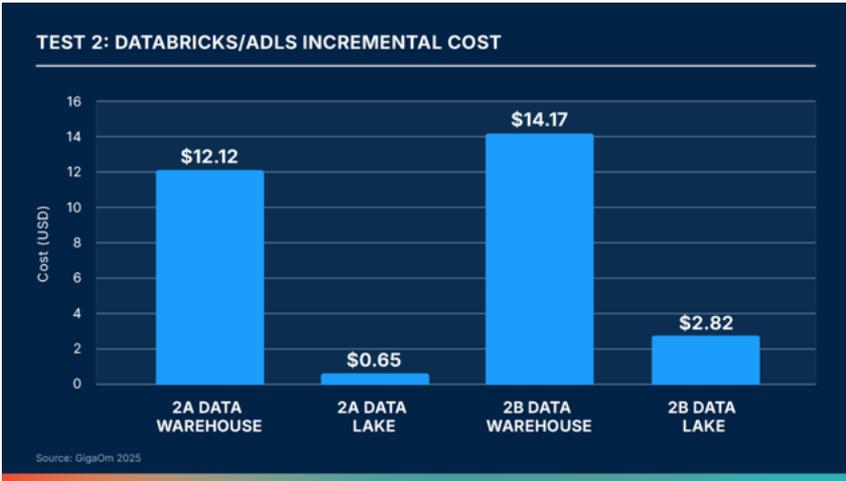


Figure 5. Total Incremental Cost with Databricks and S3



Figure 6. Total Annual Cost with Databricks and S3

S3/Fivetran Managed Data Lake Service vs Amazon Redshift

This analysis compares the cost and performance differences of using Amazon Redshift versus a modern data lake (S3 with Iceberg and Fivetran Managed Data Lake Service) for incremental data syncing and storage. The results show significant cost savings with the data lake approach, ranging from 90% (2-layer architecture) to 77% (medallion architecture) reduction in total costs compared to the data warehouse approach.

The incremental sync cost is also lower with the data lake approach, ranging from \$0.63 to \$1.64 per increment, compared to \$6.13 to \$7.09 per increment for the data warehouse approach. See **Table 14**.

A director and enterprise data architect at a multinational pharmaceutical and biotechnology company had this to say about the approach.

“Cost is a good benefit to the Iceberg approach. Along with the abstraction, open formats bring the ability to interact across technical landscapes. You can have your data sitting in Iceberg and use Databricks to read or Starburst and so on. Being more interoperable is another benefit of the modern data lake architecture,” he said.

While the data lake approach has slightly slower incremental sync times, with an 8% to 9% increase in sync time compared to the data warehouse approach, it should still meet performance requirements for many organizations.



Table 14. Cost Breakout with Redshift and S3

DESTINATION	TEST 1A		TEST 1B	
	DATA WAREHOUSE	DATA LAKE	DATA WAREHOUSE	DATA LAKE
Data integration	Redshift	Redshift	Redshift	Redshift
Silver Layer	Redshift	S3 (Iceberg)	Redshift	S3 (Iceberg)
Gold Layer	n/a	n/a	Redshift	Redshift
Query Engine	Redshift	Redshift	Redshift	Redshift
INCREMENTAL SYNC	DATA WAREHOUSE	DATA LAKE	DATA WAREHOUSE	DATA LAKE
Resource	Redshift	S3	Redshift	S3
Type	8 RPU	N/A	8 RPU	N/A
Write Usage	17,031	85,893	17,031	85,893
Write Usage Units	RPU	1,000 PUTs	RPU	1,000 PUTs
Write Rate	\$0.36	\$0.005	\$0.36	\$0.005
Read Usage		51,259		51,259
Read Usage Units		10,000 GETs		10,000 GETs
Read Rate		\$0.004		\$0.004
Read + Write Cost	\$6.13	\$0.63	\$6.13	\$0.63
INCREMENTAL GOLD UPDATE	\$0.63	DATA LAKE	DATA WAREHOUSE	DATA LAKE
Resource	n/a	n/a	Redshift	Redshift
Type			8 RPU	8 RPU
Write Usage			2,651	2,803
Write Usage Units			RPU	RPU
Write Rate			\$0.36	\$0.36
Write Cost			\$0.95	\$1.01
PER INCREMENT COST	\$6.13	\$0.63	\$7.09	\$1.64

Figure 7 shows the per-increment cost for each tested scenario, while Figure 8 extrapolates this cost over a full year.

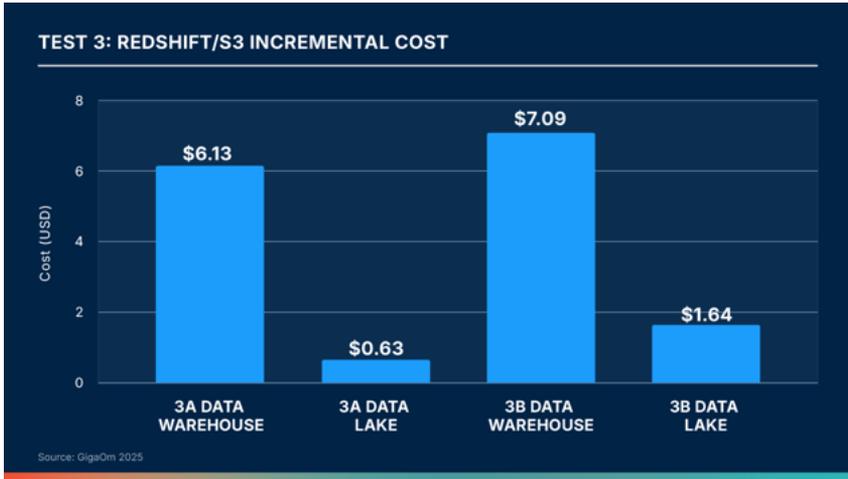


Figure 7. Total Incremental Cost with Redshift and S3



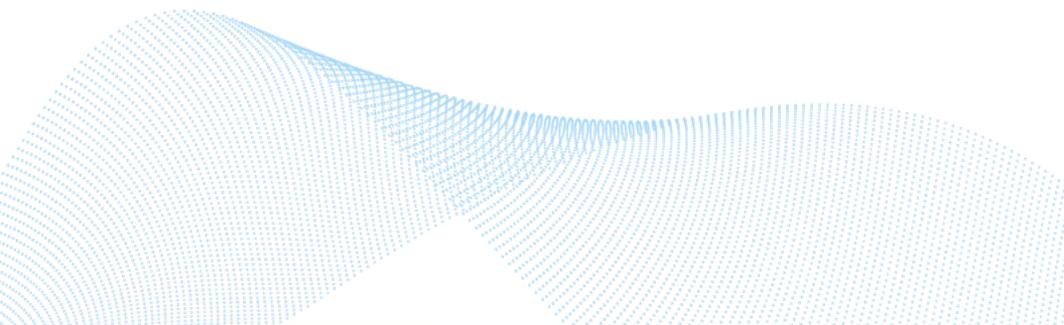
Figure 8. Total Annual Cost with Redshift and S3

04 | Conclusion

THIS REPORT DEMONSTRATES that a modern data lake approach, using S3 or ADLS with Iceberg/Delta Lake and the Fivetran Managed Data Lake Service, provides a cost-effective and scalable solution for incremental data syncing. Our analysis shows that this approach yields significant cost savings, ranging from a 77% to 95% reduction in total costs compared to traditional data warehouse solutions such as Snowflake, Databricks, and Amazon Redshift.

The cost savings are significant and the performance differentials are not a concern for most use cases.

Overall, our analysis suggests that organizations considering a modern data management solution should strongly consider a data lake approach. With its significant cost savings, scalability, and competitive storage costs, a data lake solution can provide a strong foundation for organizations seeking to derive business value from their data.



05

Appendix

A Customer Case Study of Fivetran and Data Lakes

A GLOBAL MOBILE PAYMENT SOLUTIONS COMPANY faced challenges scaling its data infrastructure to handle high-volume transactions across multiple countries. Initially relying on custom-built scripts, the company encountered scalability issues such as high latency, compute costs, and engineering overhead.

To address these issues, the customer adopted Fivetran's Managed Data Lake Service. This helped them replicate data into Iceberg on S3, resulting in several key benefits:

- **Lower compute costs:** By moving raw data to S3 with Iceberg, they could use Athena for ad hoc queries, reducing Snowflake usage and saving an estimated \$100,000 per year.
- **Automated, real-time ingestion:** This helped detect payment issues more quickly.
- **Improved governance and compliance:** Built-in catalog integration provided better data discovery and security, which was crucial for their data protection officer.
- **Greater flexibility:** Iceberg provided an open, interoperable format; reducing vendor lock-in and enabling future AI workloads.

By using Fivetran, the customer can now focus on AI innovation, leveraging AI-powered agents for contract management, predictive analytics, and customer support automation. The company has built a flexible data lake where they can run queries with tools like Athena, Snowflake, or Redshift, thus optimizing costs and performance. The customer plans to expand its Iceberg-based data lake across more databases, further optimizing costs, scalability, and AI readiness.

We spoke with the customer's head of Analytics about implementing Fivetran's Managed Data Lake Service and the benchmark results.

"The way Fivetran commits data to Iceberg is very interesting. It doesn't do any merge, doesn't do any complex query. It's really working on the Parquet file and updating the data. Without Fivetran, managing the data lake would have required five or six full-time engineers," he said. This efficient data updating mechanism provides significant advantage, as it avoids the need for data merging and reduces cost and performance issues during ingestion.

He is convinced that "Iceberg is the future for storing data." And it's hard to argue with that assessment. With its ability to offer significant performance benefits once properly engineered and tuned, Iceberg is poised to revolutionize the way the customer stores and manages data.

Despite its many advantages, Iceberg is not without its challenges. “The main problem is really the management of the permission and authorization. It’s a total mess to handle. It’s very complex to understand,” according to the head of Analytics. This complexity is particularly evident when using AWS Lake Formation, and can be further exacerbated when integrating Iceberg with other platforms like Snowflake.

So how to overcome these challenges and realize the full potential of Iceberg? According to the head of Analytics, benchmarks such as these are crucial. “By seeing our experiences and findings in this benchmark, we can reinforce our confidence in the technology and move forward with greater certainty,” he says.

As the customer continues to navigate the evolving landscape of data storage, it’s clear that Iceberg is a key player. With its efficient data ingestion, significant performance benefits, and potential to revolutionize data storage and management data, Iceberg and data lakes are worth keeping an eye on.

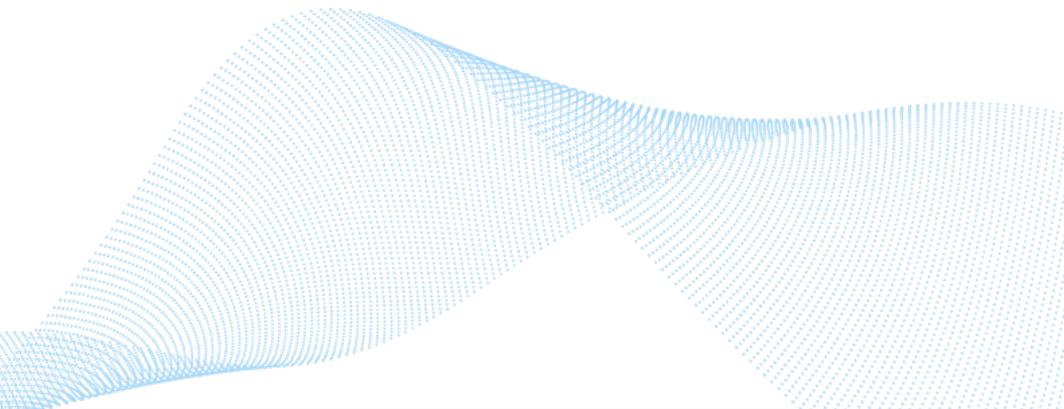
A modern data lake architecture offers numerous benefits, including lower TCO, reduced data ingestion costs, and increased flexibility. It also provides time savings through minimized maintenance efforts and enables a scalable, cost-effective, and high-performance data foundation for analytics and AI workloads.

Storage

Storage Costs are based on usage times and rates, and are relatively consistent.

The annual costs range from a low of \$313.02 to a high of \$544.08. Snowflake’s costs are the lowest, ranging from \$313.02 to \$414. Databricks’ costs fall in the middle, ranging from \$357.96 to \$539.76. Redshift’s costs are \$446.4 and \$414, while using Fivetran with a data lake approach is consistently the highest storage costs for our test, ranging from \$443.10 to \$544.08.

Overall, the storage costs were largely similar, suggesting that usage patterns and rates had a more significant impact on costs than the platforms themselves.



06 | Disclaimer

TCO IS IMPORTANT but is only one criterion in a data lake selection. The tests applied in this report offer a point-in-time check into specific performance aspects of the solution. There are numerous other factors to consider across performance, administration, features and functionality, workload management, user interface, scalability, reliability, and numerous other criteria. It is also our experience that TCO changes over time and is competitively different for different workloads. Moreover, a TCO leader can reach a point of diminishing returns, enabling contenders to quickly close the gap.

GigaOm runs all its performance tests to strict ethical standards. The results of the report are the objective results of the application of the tests to the simulations described in the report. The report clearly defines the selected criteria and process used to establish the field test. The report also clearly describes the tools and workloads used. The reader is left to determine for themselves how to qualify the information for their individual needs. The report does not make any claim regarding third-party certification and presents the objective results received from the application of the process to the criteria as described in the report. The report strictly measures TCO and does not purport to evaluate other factors that potential customers may find relevant when making a purchase decision.

This is a sponsored report. Fivetran chose the competitors and the test. GigaOm chose the most compatible configurations as-is out of the box and ran the testing workloads. Choosing compatible configurations is subject to judgment. We have attempted to describe our decisions fully in this report.

About Fivetran

FIVETRAN, the industry leader in data integration, enables enterprises to power AI workloads like predictive analytics, AI/ML applications and generative AI, and accelerate cloud migration. The Fivetran platform reliably and securely centralizes data from hundreds of SaaS applications and databases into a variety of destinations—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





About William McKnight

William McKnight is a former Fortune 50 technology executive and database engineer. An Ernst & Young Entrepreneur of the Year finalist and frequent best practices judge, he helps enterprise clients with action plans, architectures, strategies, and technology tools to manage information.

Currently, William is an analyst for GigaOm Research who takes corporate information and turns it into a bottom-line-enhancing asset. He has worked with Dong Energy, France Telecom, Pfizer, Samba Bank, ScotiaBank, Teva Pharmaceuticals, and Verizon, among many others. William focuses on delivering business value and solving business problems utilizing proven approaches in information management.



About Jake Dolezal

Jake Dolezal is a contributing analyst at GigaOm. He has two decades of experience in the information management field, with expertise in analytics, data warehousing, master data management, data governance, business intelligence, statistics, data modeling and integration, and visualization. Jake has solved technical problems across a broad range of industries, including healthcare, education, government, manufacturing, engineering, hospitality, and restaurants. He has a doctorate in information management from Syracuse University.

GIGAOM

About GigaOm

GigaOm provides technical, operational, and business advice for IT's strategic digital enterprise and business initiatives. Enterprise business leaders, CIOs, and technology organizations partner with GigaOm for practical, actionable, strategic, and visionary advice for modernizing and transforming their business. GigaOm's advice empowers enterprises to successfully compete in an increasingly complicated business atmosphere that requires a solid understanding of constantly changing customer demands.

GigaOm works directly with enterprises both inside and outside of the IT organization to apply proven research and methodologies designed to avoid pitfalls and roadblocks while balancing risk and innovation. Research methodologies include but are not limited to adoption and benchmarking surveys, use cases, interviews, ROI/TCO, market landscapes, strategic trends, and technical benchmarks. Our analysts possess 20+ years of experience advising a spectrum of clients from early adopters to mainstream enterprises.

GigaOm's perspective is that of the unbiased enterprise practitioner. Through this perspective, GigaOm connects with engaged and loyal subscribers on a deep and meaningful level.



Copyright

© [Knowingly, Inc.](#) 2025 *"Ingestion Costs Compared: Fivetran Managed Data Lake vs. Data Warehouse"* is a trademark of [Knowingly, Inc.](#) For permission to reproduce this report, please contact sales@gigaom.com.



GIGAOM

GigaOm is built on the perspective of professionals who have operated, deployed, and managed the technologies we assess. Our practitioner lens brings nuance, realism, and foresight, ensuring our research is trusted, actionable, and aligned to the real pressures facing modern enterprises.