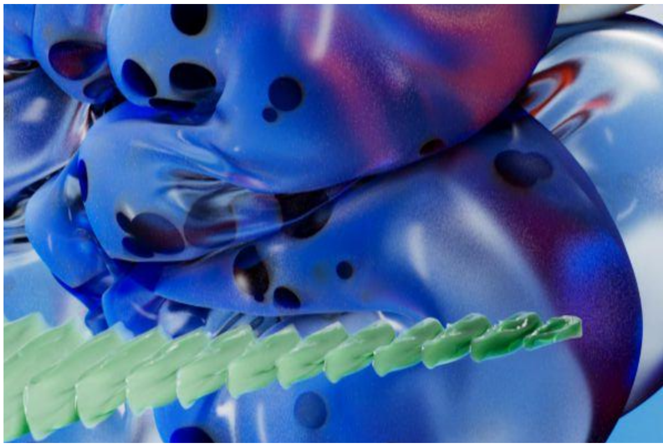
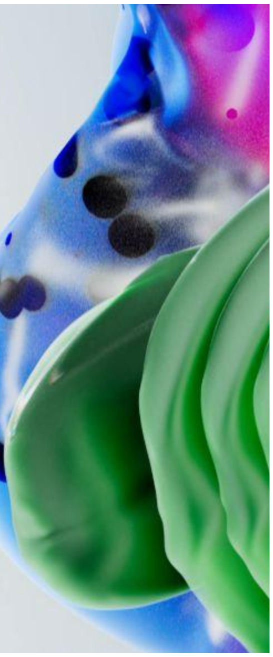
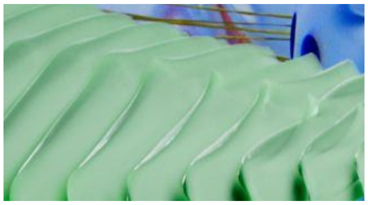
Google Cloud

A practical guide to data lake migration







Unlock your data potential to fuel your AI transformation

Author: Dana Soltani,
Group Product Manager, Google Cloud

Table of contents

Executive summary 03

Chapter 01
The evolution of data lakes 04

It all started with big data 05
Legacy data lakes came with challenges 05
How Cloud changed data lakes 06
Cloud data lakes unlock the potential of data and AI 08

Chapter 02
The economics of data lake migration 10

Cost is a key driver for change 11
Cloud changes your cost structure 13
Planning for upfront migration costs 15
Costs can be optimized in the cloud 16
Migration risks must be carefully managed 18

Chapter 03
Planning and executing a successful data lake migration 21

5 stages of lake migration 22
Organizational planning is essential before you start 23
Phase 1: Discovery 24
Phase 2: Assessment 25
Phase 3: Planning 28
Phase 4: Execution 29
Phase 5: Optimization 31
A checklist to help you plan 33

Chapter 04
How Google can help 34

Executive summary

If you're planning a data lake migration to the cloud, this guide is your companion for success. Based on the deep learnings of countless enterprise data lake migrations over the last five years, **here's how it could help optimize your migration journey**.

Understand the 'why'

Explore the evolution of data lakes and the compelling reasons to modernize with a cloud-based approach. You will learn about the challenges of legacy systems and the benefits of migrating to a scalable, cost-effective, and innovative cloud environment.

Make informed economic decisions

Delve into the economic aspects of cloud migration, including cost optimization strategies and real-world examples of companies achieving significant cost savings. You will learn how to align your migration strategy with business SLAs and choose the right approach for your needs.

Plan your migration

Follow the clear, five-phase migration process in the guide, from discovery and assessment to execution and optimization. Detailed checklists and key considerations for each stage ensure a structured and efficient migration journey.

Migrate with confidence

Adopt best practices for migrating various components of your data lake, including data, metadata, workloads, governance, and workflows. You will learn how to ensure data integrity, minimize downtime, and optimize performance in the cloud.

Tap into Google Cloud's support

Explore the comprehensive suite of services, tools, and expertise on offer from Google Cloud. This includes personalized support from account teams, automated migration tooling, consulting services, technical assistance, and access to a vast partner ecosystem.

By following the guidance and best practices in this guide, you can confidently migrate your data lake to Google Cloud, unlocking scalability, cost-efficiencies, and access to cutting-edge data analytics and AI/ML capabilities.

Chapter 01

The evolution of data lakes

The way that organizations collect, store, and analyze data has evolved over the years. From databases, to data warehouses, to data lakes, this evolution has gone hand-in-hand with other innovations—like the internet, big data analytics, and now AI. To help understand where we are today, let's take a look back at how far data lakes have come.

It all started with big data

In the early days of big data, traditional data warehouses struggled to keep up with the growing volume and variety of information being produced. Designed for structured data, they left unstructured and semi-structured data largely untapped. Data lakes, which emerged in the early 2010s, set out to close the gap. A data lake is a vast repository that stores raw data in its native format, regardless of its structure or intended use—giving organizations new opportunities to collect and analyze all types of data.

Hadoop played a pioneering role in this space. Based on the Google MapReduce paper[1], its distributed file system (HDFS) can store massive amounts of data, and its processing framework (MapReduce) can analyze this data in parallel. Later, Apache Spark emerged as a faster and more efficient processing engine. It wasn't long before organizations began building and maintaining data lakes on-premises, within owned-and-operated data centers. There, they could safely and cost-effectively store and process their proprietary data—leaning on architectures like Hadoop and Spark to help get the job done.

Legacy data lakes came with challenges

While offering control and security, on-premises data lakes made it hard for organizations to get the most from their data assets. Challenges included:

Scalability limitations

Constrained by the physical capacity of hardware infrastructure, it was costly and time-consuming to scale up to accommodate growing data volumes and processing demands. This led to performance bottlenecks and hindered the ability to handle spikes in data or workloads.

$ High upfront and maintenance costs

Building and maintaining an on-premises data lake involved substantial upfront investment in hardware, software licenses, and IT infrastructure. Ongoing maintenance, including hardware upgrades, software updates, and security patching, added to the total cost of ownership.

Management overheads

Specialized expertise and dedicated IT resources were needed for tasks like hardware provisioning, software installation, configuration, performance tuning, and security management. This took time away from higher-value activities like data analysis and insight generation.

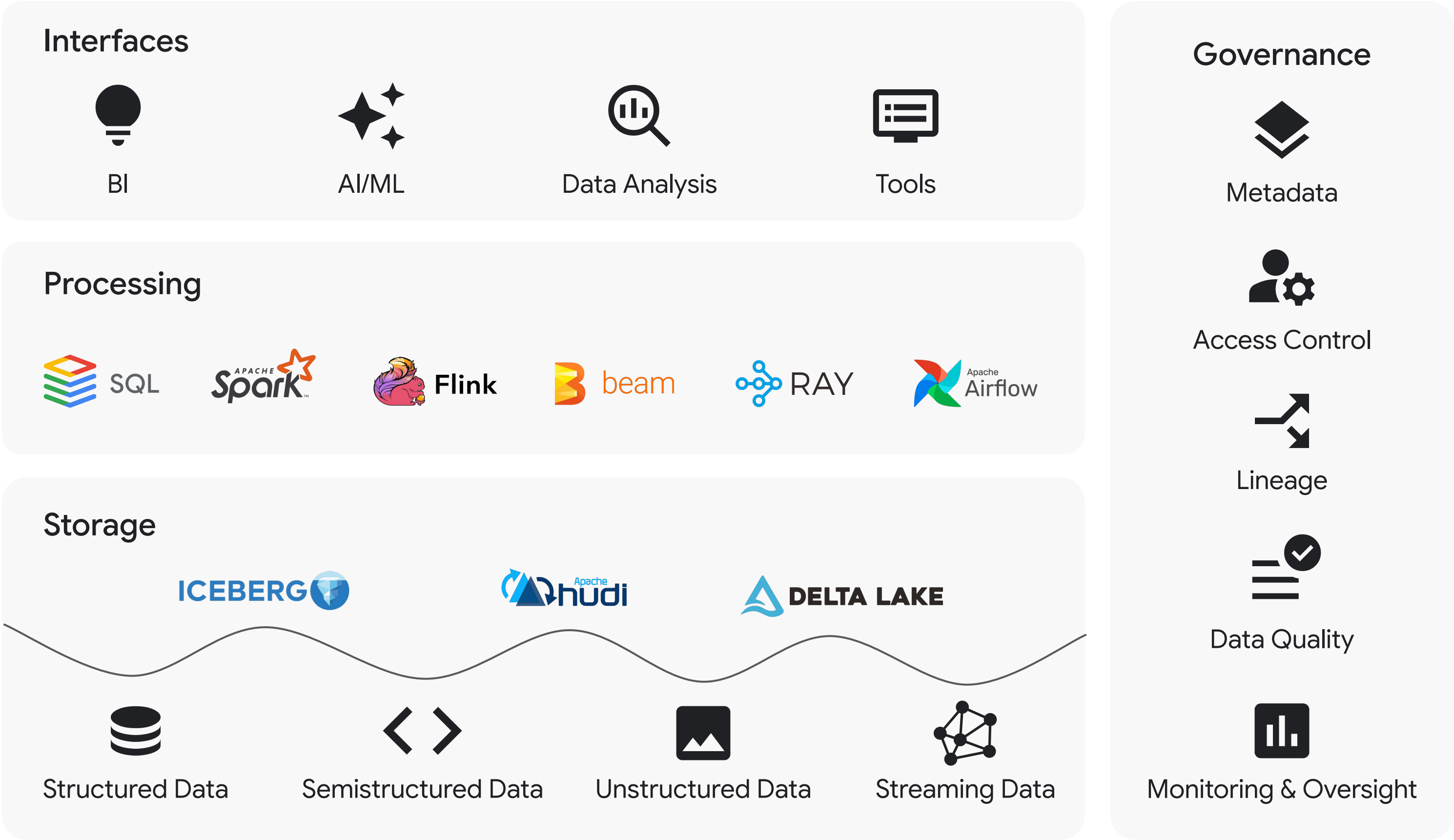
Limited access to innovation

The latest advances in analytics, AI, and cloud-native services—such as serverless computing, AI/ML platforms, and advanced analytics tools—were not compatible with on-premises data lakes. This impeded innovation and made it hard to drive business advantage from data.

[1] Dean, J and Ghemawat, S, 2004, MapReduce: Simplified Data Processing on Large Clusters

How Cloud changed data lakes

Cloud computing took data lakes to the next level. Today's modern, cloud-based data lakes are architected to help solve the big challenges typically associated with traditional data lake deployments. Compared to the static, siloed set-up of on-premises storage and compute, they are highly elastic and ephemeral—built with four layers designed to help organizations unlock new opportunities from their data.



Storage

The base layer stores structured, semi-structured, and unstructured data as files in generic cloud storage. Apache Parquet, Avro, and ORC are the most common file formats; and while the hive table format is commonly used to query data in SQL, it is being supplanted by more modern formats like Apache Iceberg. In addition to atomicity, consistency, isolation, and durability (ACID) transactional support, Iceberg provides efficiencies for petabyte-scale tables and advanced functionalities like schema and partition evolution, time travel, and materialized views.

Processing

The separation of storage and compute carries over from on-premises data lakes and is amplified in the cloud, where processing is ephemeral and compute resources are provisioned only when necessary to minimize costs. Depending on the use cases, multiple processing engines such as Apache Spark, Flink, or Ray might be used to process data in batch or real time.

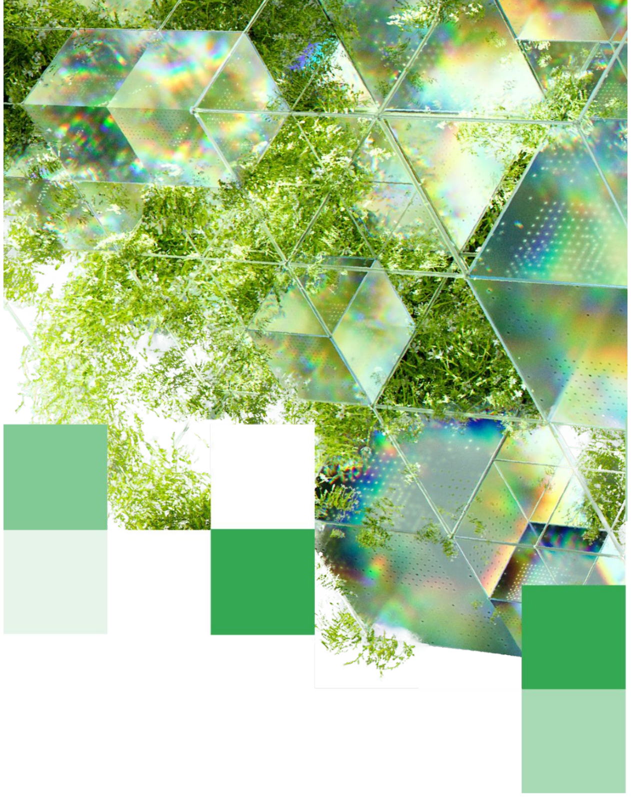
Tools and interfaces

This is the layer where different users come into contact with the data lake, and use cases can range from ad-hoc analysis to building a machine learning pipeline for continuous forecasts in a business application. Users interact with the data lake through things like BI tools, data science notebooks, SQL workbenches, and APIs.

Governance

Data lakes often contain sensitive data and mission critical processes that need to be tightly governed. Diverse functions like metadata management, data quality control, lineage tracking, and access control help reduce both operational costs and the risk of costly data leaks.

This structure accelerates time-to-value for analytics and the delivery of insights across the organization—significantly outpacing the ROI of on-premises data lakes that many organizations still have in play today.



Cloud data lakes unlock the potential of data and AI

Many organizations are finding that their current data infrastructure isn't built for AI. Legacy data lakes cannot efficiently store and process the large volumes of unstructured data that AI thrives upon. Often, data silos persist across warehouses, lakes, and clouds. And many organizations simply lack the compute resources required for building and serving AI models.

To overcome these limitations, the natural next step in the evolution of legacy data lakes is modernization—that is, migrating data lakes to the cloud and modernizing the architecture to bring all data together in a unified platform.

A cloud data lake enables organizations to:

Make faster data-driven decisions

In the cloud, data management and processing are unified, making it easier to use data to drive competitive advantage and value for end customers.

Accelerate adoption of AI

Tools for experimentation, monitoring, and governance across structured and unstructured data are readily available, as are machine learning models to accelerate innovation with AI.

Reduce complexity

Cloud data lakes typically come with multi-tiered storage options, a choice of processing engines, and a variety of machine learning and AI models to help simplify data and AI applications and reduce costs.

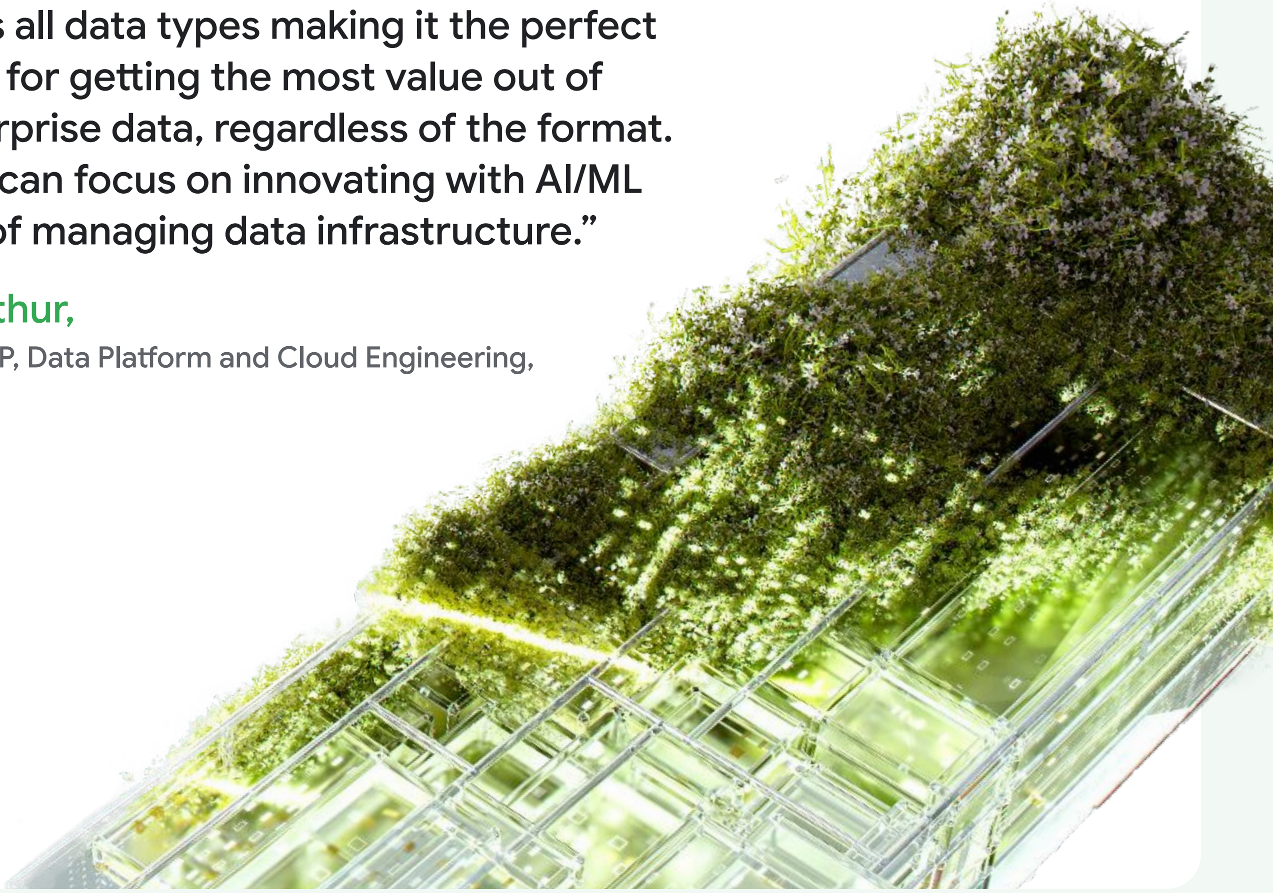
Improve governance

In the cloud, it's easier to ensure data and insights are high-quality, trustworthy, and compliant with regulations.

Definity migrates to BigQuery for cost savings and performance gains

Definity, a leading Canadian P&C insurer with over 150 years of history, knew they needed to modernize their data infrastructure to keep pace with the demands of a rapidly evolving industry. Their legacy on-premises Cloudera platform was hindering their ability to scale, innovate, and fully leverage the power of their data and AI. Definity migrated to Google Cloud, built their data store on BigQuery, and set up Vertex AI as their analytical and AI platform in just 10 months. They unlocked over 30% cost savings in annual infrastructure and operation costs, while seeing a 63% improvement in deployment time and a 10X faster infrastructure setup.

> We chose BigQuery for its scalability, cost-effectiveness, and connectivity to Vertex AI. It's the ideal solution for our unified data analytics platform needs. BigQuery supports all data types making it the perfect platform for getting the most value out of our enterprise data, regardless of the format. Now we can focus on innovating with AI/ML instead of managing data infrastructure."
>
> **Nitin Mathur,**
>
> Associate VP, Data Platform and Cloud Engineering, Definity



Chapter 2

The economics of data lake migration

Migrating your data lake to the cloud pays off in more ways than one. As well as the significant cost savings realized by shifting to a cloud model, the move can lead to greater efficiencies, faster access to insights, better business decision-making, and the ability to tap into the latest AI innovations. And it all translates to more revenue opportunities and a competitive edge.

20-62%
lower costs

Achieved when an organization uses a managed cloud data lake vs self managed or on-premises[2]

80-90%
lower costs

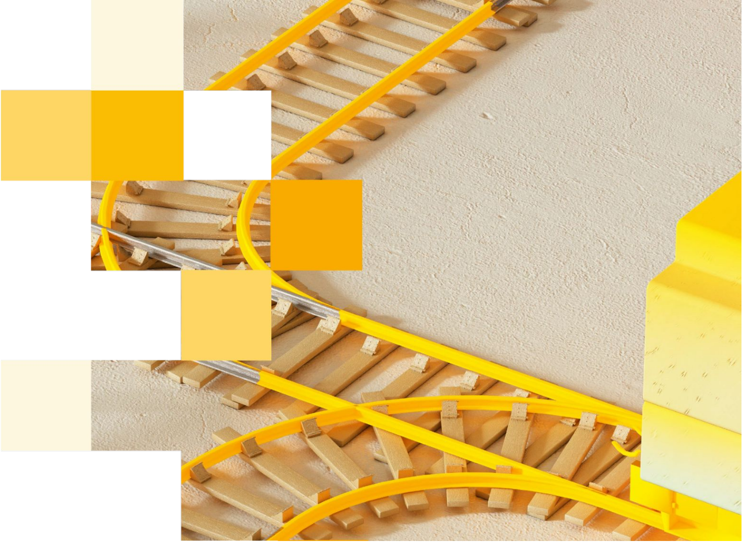
Achieved by moving to a lakehouse architecture, combining data lakes and warehouses on a managed cloud service[2]

Cost is a key driver for change

There's no doubt that on-premises data lakes are becoming increasingly uneconomical. To start with, there are all the actual costs of building and maintaining the infrastructure, which leads to high CapEx and OpEx. Further, to handle the ever-growing volume and variety of data, on-premises data lakes often require costly hardware additions. And finally, managing and securing the infrastructure demands specialized IT expertise, adding to personnel costs.

Then there are the opportunity losses. On-premises solutions lack the flexibility and agility of cloud-based data lakes, hindering rapid innovation and AI adoption—thus making it harder to respond to changing business needs. As a result, many enterprises find themselves on an unsustainable path with their on-premises data lakes.

[2] ESG, 2022, Analyzing the Economic Benefits of Google Cloud Dataproc



Cloud changes your cost structure

The cost structure of a cloud data lake resembles a pay-as-you-go utility model, with costs scaling based on usage.

The key components include:

Storage

This is a fundamental cost, charged per gigabyte stored and influenced by factors like data volume, frequency of access, and chosen storage tiers (hot, cold, archive).

Compute

These costs are incurred for processing and analyzing data, with charges based on the type and duration of compute instances used for tasks like extract, transform, and load (ETL), querying, and machine learning.

Managed services

Costs of using managed analytics and data warehousing engines, BI tools, streaming services, etc.

Data transfer

These costs arise when moving data into, out of, or within the cloud environment.

Orchestration and management

These are the costs associated with scheduling and monitoring data and AI pipelines.

Data governance, security, and compliance

Features like access control, encryption, and auditing can contribute to the overall cost.

Understanding and optimizing each of these cost components is essential for managing the total cost of ownership of your cloud data lake.

Eureka reduces downtime and accelerates app deployment on Google Cloud

Technology company Eureka helps organizations derive intelligence at scale by applying AI and machine learning to large data sets. To better run its data analytics services, the company has migrated to Google Cloud and is using BigQuery and Dataproc—reducing server downtime and accelerating the deployment of new applications. They have also realized cost savings, with BigQuery storage delivering cost reductions of as much as 18%.

> Improving our data analytics capabilities has been refreshing. Also, since migrating to Google Cloud, we've been able to significantly reduce downtime and outages compared to when we were on our on-prem solution.”

Michael Hawkins
CMO,
Eureka

Read the full story →

Planning for upfront migration costs

There are various upfront costs involved in migrating an on-premises data lake to the cloud. By understanding these costs from the outset, you can avoid unexpected surprises or budget blow-outs.

Upfront migration costs typically include:

Cloud infrastructure

Costs of provisioning virtual machines, storage, and networking components during the migration process while the on-premises data lake is still in use and cannot be decommissioned.

Professional services

Most enterprise organizations use consulting, implementation support, and training to assist with their cloud migration.

Data transfer

This can include egress fees from your current data center and potential bandwidth costs.

Data transformation and cleansing

This can be a significant expense, as you adapt data to the cloud environment and ensure compatibility.

Licensing and subscription fees

You may need to pay for cloud-based software and tools required for your data lake migration.

Costs can be optimized in the cloud

Optimizing cloud data lake costs requires proactive and continuous effort. Here are some specific actions you can take before, during, and after migration to significantly reduce costs while maintaining performance and scalability.

Right-size your compute and storage resources

Analyze your workload patterns and adjust compute capacity based on actual needs. Use autoscaling to dynamically adjust resources based on demand. Choose the appropriate storage tiers (hot, cold, archive) based on data access frequency.

Use cost-effective pricing models

Take advantage of reserved instances or spot instances for predictable or fault-tolerant workloads. Explore sustained use discounts and other pricing options offered by the cloud provider.

Optimize data storage

Compress data to reduce storage footprints and utilize data lifecycle management policies to automatically move data to cheaper storage tiers as it ages. Regularly delete or archive unused data.

Optimize data pipelines

Design efficient data ingestion and processing pipelines. Minimize data movement across regions and availability zones to reduce network costs.
Use serverless computing for data transformation tasks to pay only for actual usage.

Monitor and analyze cloud spending

Utilize cloud provider cost management tools to track spending, identify cost drivers, and set budget alerts. Regularly review cost reports and analyze spending patterns to identify areas for optimization.

Regularly review and optimize

Continuously monitor your data lake environment and adapt your strategies as your needs evolve. Stay informed about new cloud offerings and cost optimization best practices.

Implement data governance policies

Establish clear data retention policies and enforce data quality standards to minimize storage costs and improve data usability.

Use cloud-native tools and services

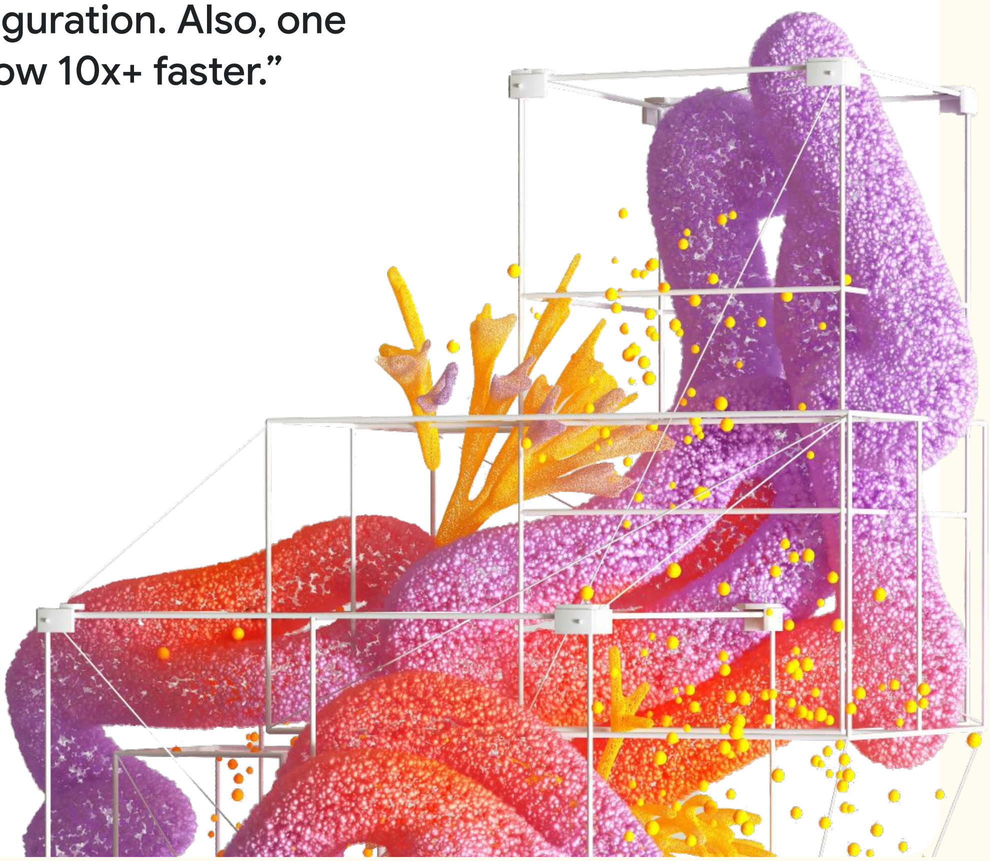
Take advantage of managed services like BigQuery or serverless Spark on Dataproc for ELT/ETL processes, which is more cost-effective than self-managed solutions.

LiveRamp migrates Hadoop to Google Cloud

With its business scaling, LiveRamp was running into constraints relating to data center space and power in its on-premises environment, which restricted its ability to meet business objectives. A strategic decision was made to leverage elastic environments and migrate from Hadoop to Dataproc—and it has paid off.

> We have achieved around 30% cost savings in certain clusters by achieving the right balance between on-demand and spot VMs. The cost savings were a result of our engineers building efficient A/B testing frameworks that helped us run the clusters/jobs in several configurations to arrive at the most reliable, maintainable, and cost efficient configuration. Also, one of the applications is now 10x+ faster."
>
> Mithun Bondugula
>
> Sr Engineering Manager,
> LiveRamp

Read the full story →



Migration risks must be carefully managed

When migrating a data lake to the cloud, there are three broad areas that can pose a risk to your organization. Here's what to watch out for and how to mitigate the risks.

01 Data-centric risks

Data loss

Data can be lost during transfer due to network issues, errors in extraction/ loading, or even accidental deletion in the new environment. Mitigation should involve robust data validation, backup/recovery mechanisms, and potentially using data quality tools to ensure completeness.

Data corruption

Data can become corrupted in transit or during transformation processes. This might be due to incompatible formats, encoding issues, or bugs in the migration scripts. Thorough data profiling before and after migration, along with schema mapping validation, is essential.

Data security

Maintaining data security during every phase of migration is paramount. Risks include unauthorized access, data breaches in transit, or misconfigurations in the cloud environment that expose data. Encryption, strong access controls, and security audits throughout the project are crucial.

02 Project management and execution risks

Unrealistic timelines

Underestimating the complexity and time required for migration can lead to rushed decisions, increased errors, and ultimately project failure. Thorough planning, including realistic timelines with buffers
for unexpected issues, is key.

Skill gaps

A lack of cloud expertise—including skills in cloud platforms, data migration tools, and security best practices—can significantly hamper the project. Training existing staff or bringing in experienced cloud professionals is often necessary.

Poor communication

Migration projects involve various stakeholders. Poor communication can lead to misunderstandings, missed deadlines, and conflicts. Establish clear communication channels and ensure all stakeholders are informed and aligned.

Insufficient testing

Testing is crucial to identify and address potential issues before they impact production. If it's rushed or inadequate, costly post-migration problems can arise. Comprehensive testing strategies, including performance and security testing, are a must.

03 Cloud-specific risks

Vendor lock-in

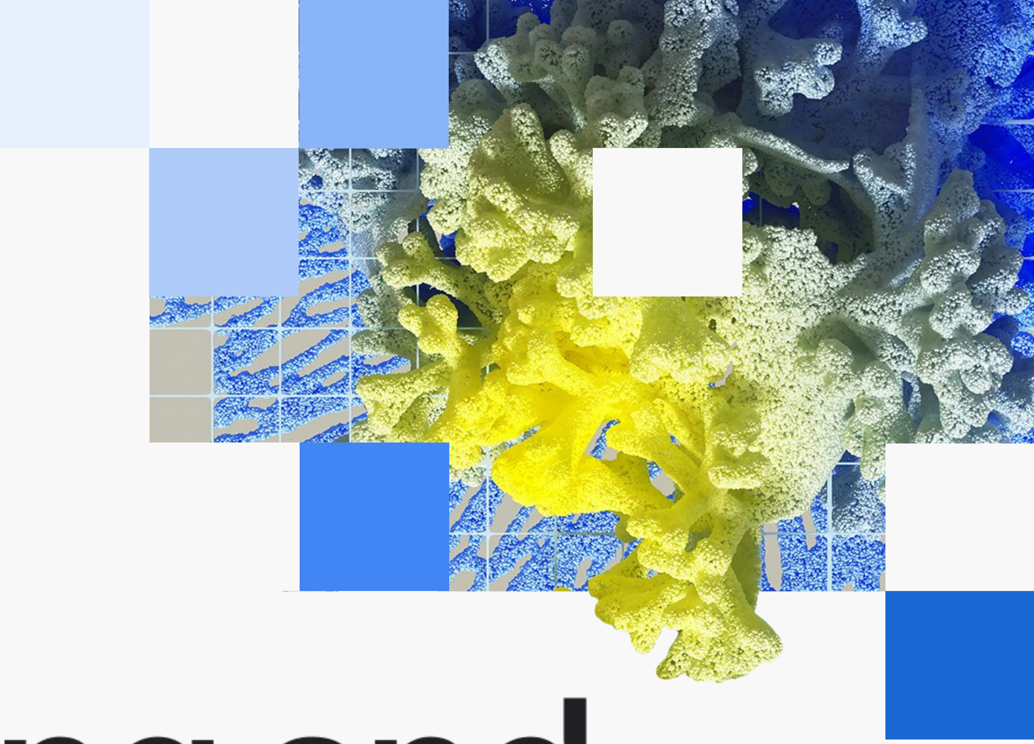
Becoming overly reliant on a single technology or provider can limit future flexibility and potentially lead to higher costs. To mitigate this risk, evaluate multiple technical options and providers, and consider a strategy that emphasizes interoperability.

Unexpected cloud costs

Cloud pricing can be complex. Unforeseen costs related to data transfer, storage, or compute can take you by surprise if not properly planned for. Thorough cost estimation, ongoing cost monitoring, and optimization strategies are essential.

Integration challenges

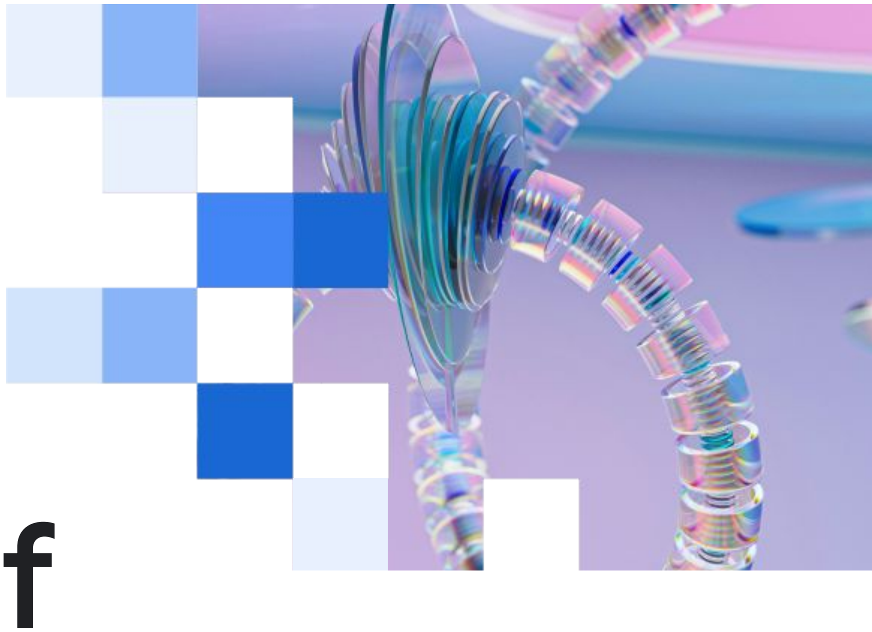
Integrating the cloud data lake with existing on-premises systems, applications, or data sources can be complex. Ensuring seamless data flow and compatibility across the hybrid environment requires careful planning and integration testing.



Chapter 3

Planning and executing a successful data lake migration

Migrating a data lake to the cloud can be a complex process—one that requires meticulous planning and flawless execution. This chapter delves into the critical steps involved, from assessing your current environment and defining your cloud strategy, to orchestrating the migration and optimizing for ongoing success.



5 stages of lake migrations

Phase 01

Discovery

Build an understanding of the current state of the data lake by taking inventory of all data assets, workloads, users, permissions, and workflows. This information is crucial for determining the scope of the migration and identifying any potential challenges.

Phase 02

Assessment

Define the optimal cloud architecture, which includes selecting the right cloud services, establishing governance policies, determining cost controls, and outlining processes and protocols for data management and security.

Phase 03

Planning

Create a detailed project plan for the migration, including defining tasks and subtasks, assigning roles and responsibilities, establishing timelines, and setting a budget.

Phase 04

Execution

Migrate tables and data, users and permissions, and workloads and workflows to the cloud. Test and validate the migrated data and applications to ensure they perform as expected in the new environment.

Phase 05

Optimization

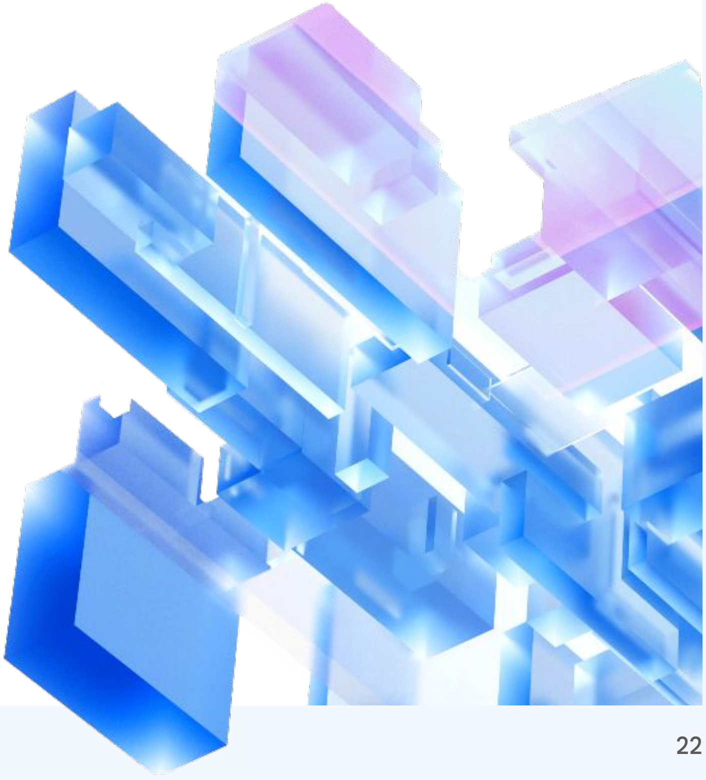
Once the migration is complete, continue to fine-tune performance, ensure cost efficiency, and enhance oversight and governance processes. During this phase, you may use cloud-native features and services to improve scalability, reliability, and maintainability.

Squarespace reduces number of escalations by 87% with analytics lakehouse on Google Cloud

To enhance scalability, reduce maintenance overheads, and foster innovation within their data operations, Squarespace migrated its on-premises Hadoop infrastructure to Google Cloud, primarily leveraging BigQuery, Dataflow, and Cloud Composer to build a unified data platform.

> Following the successful migration to Google Cloud of our Hadoop ecosystem, we have seen the significant maintenance burden of the infrastructure disappear. From the months before our migration compared to the months immediately after, we've seen an 87% drop in the number of escalations. The data platform and data infrastructure teams have turned their attention away from monitoring the health of various services/ filesystems, and are now focused on delivering new features and better software that our internal users need to move our business forward."
>
> **Douglas O'Connor**
> Senior Software Engineering Manager,
> Squarespace
>
> **Constantinos Sevdinoglou**
> Senior Staff Software Engineer,
> Squarespace

Read the full story →



Organizational planning is essential before you start

To ensure the success of your data lake migration, you should start by securing C-level sponsorship and assembling a cross-functional team with representatives from various departments, including IT, data engineering, data science, BI, and business units. Clearly define roles and responsibilities for each team member, including a dedicated platform team responsible for the underlying cloud infrastructure.

Then, to keep everyone informed and facilitate collaboration, establish clear communication channels and reporting structures.

Regular meetings, progress reports, and a centralized communication platform can help maintain transparency and promptly address any roadblocks. Encourage teams to work together, share their expertise, and learn from each other throughout the migration process.

To solve for any skill gaps, offer training and upskilling on cloud technologies and best practices. This empowers teams to effectively navigate the new cloud environment and make the most of its capabilities. Consider collaborating with experienced partners to supplement your team's skills and accelerate the migration process.

And, finally, be proactive in addressing any potential concerns or resistance to change within the organization. Communicating the benefits of the cloud migration and providing support as team members adapt to the new environment can help here.

Phase 01

Discovery

During the discovery phase, your goal is to build a comprehensive understanding of your current data lake environment. This will help you determine the scope of your migration, identify potential challenges and risks, and estimate the resources and timeline required for a successful migration.

At the same time, you should catalog the needs of your organization, including business service-level agreements (SLAs), to ensure your migration is a success—not only technically, but tactically as well. SLAs define the expected performance, availability, and security levels for your data and workloads, and these should be central to your migration strategy.

What elements of your current state should you be mapping out? Your inventory should include:

Data assets

Identify and catalog all data sources, types, volumes, and their locations; as well as downstream consumers (systems and users) that are dependent on the data. Data discovery tools can help automate this process. Pay particular attention to critical data that supports your core business functions and has the most stringent SLAs. This way, you can prioritize migrating this data first to minimize disruption to essential operations. Also catalog your metadata—that is, the information that describes your data, such as schemas, data lineage, data quality metrics, and business context.

Governance

Thoroughly document your current data governance policies, including data classification, data retention, access control rules, and compliance requirements. Identify all users and groups with access to the data lake, along with their respective roles and permissions. And, finally, check whether your existing data governance tools and processes are compatible with your chosen cloud platform—as you may need to adapt them to align with the cloud provider's security and governance frameworks.

Workloads

Identify all processes and applications that interact with your data lake, including ETL pipelines, data transformation jobs, analytical queries, and machine learning models. Assess their dependencies on other systems and data sources, and check their compatibility with your chosen cloud platform so you can plan for any necessary code modifications or architectural changes. As you did with your data assets, identify and prioritize any business-critical workloads.

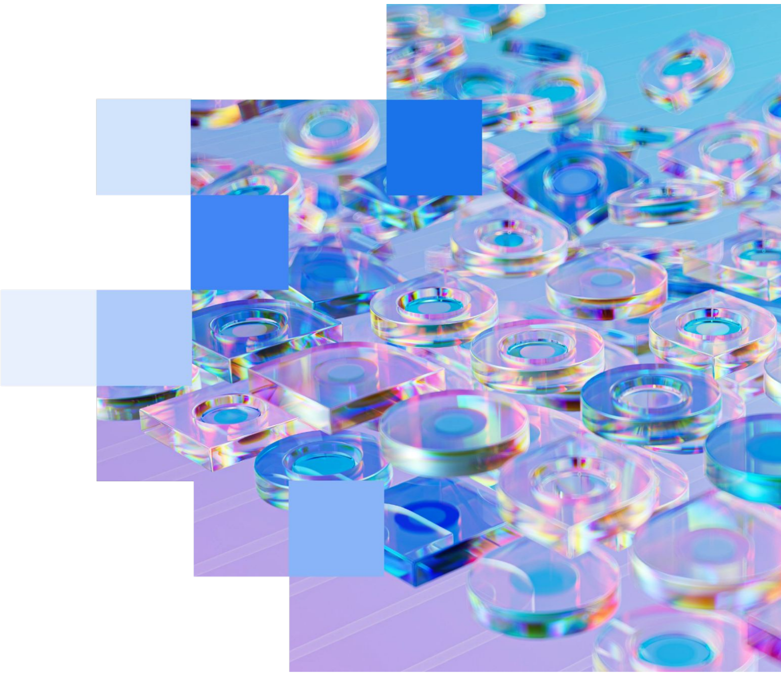
Workflows

Identify all workflows and directed acyclic graphs (DAGs) associated with your data lake, including ETL pipelines, data transformation jobs, and machine learning workflows. Analyze their dependencies, schedules, and resource requirements. As with workloads and governance, check the compatibility of your current workflow orchestration tools with your chosen cloud platform so you're clear on any changes required.

Phase 02

Assessment

Now that you have a better understanding of your data lake, you can define your optimal cloud architecture. This includes selecting the right cloud services, establishing governance policies, determining cost controls, and outlining processes and protocols for data management and security.



Defining your migration strategy

Choosing the right migration strategy depends on factors like the complexity of your data lake, your budget and timeline constraints, your desired level of cloud optimization, and your risk tolerance. A well-defined strategy ensures a smooth and efficient migration process, minimizing disruption and maximizing the value of your cloud investment.

Typically, organizations will choose from one of three strategies for data lake migrations.

Lift-and-shift

Migrate your existing data lake to the cloud 'as-is', with minimal changes. Suitable for tight timelines or when you want to minimize initial disruption, this approach could hamper your ability to tap into cloud benefits like scalability, cost efficiency, and advanced analytics.

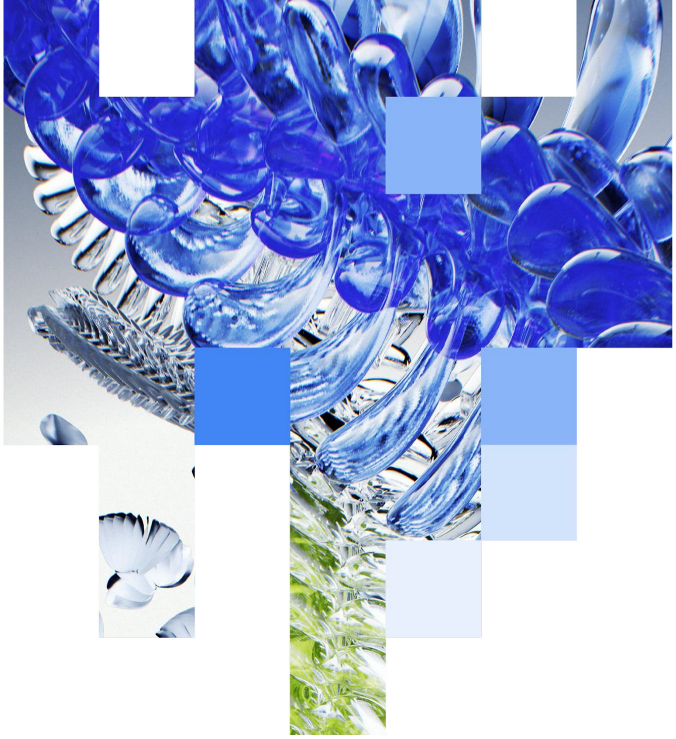
Lift-and-optimize

Migrate to the cloud and then incrementally improve your data lake architecture, using managed cloud services for greater efficiency. Striking a balance between speed and optimization, this strategy helps you realize some cloud benefits early on, while iteratively optimizing for cost and performance.

Modernization

Re-architect your data lake to fully leverage cloud-native services and capabilities. While this comprehensive approach offers the greatest potential for long-term benefits, it comes with significant upfront investment and a longer implementation timeline.

Designing your architecture



To optimize for performance, scalability, and cost-efficiency in the new environment, pay close attention to the design of your cloud architecture. As part of this, ensure that it can meet or exceed your business SLAs—leveraging cloud-native services that offer built-in SLA guarantees where possible, such as managed data warehouses or serverless data processing engines.

The architecture of a cloud-based data lake includes six key elements.

01 Storage

Object storage

Cloud object stores provide scalable and cost-effective storage for raw as well as processed data, especially unstructured data like images, videos, and log files.

Table formats

Modern table formats like Apache Iceberg offer significant advantages over traditional table formats like Apache Hive. They provide ACID properties, schema evolution, and time travel capabilities, crucial for data lakes with frequent updates and evolving schemas.

02 Compute

Serverless streaming

Cloud providers offer serverless streaming platforms such as Pub/Sub and Dataflow, and managed services for Apache Flink and Apache Kafka. These products are ideal for event-driven, real-time data processing and transformations.

Managed engines

Cloud providers offer managed, serverless Spark and Ray services for large-scale data processing, analytics, and machine learning. These services simplify infrastructure management and provide optimized performance.

Data warehousing

For structured data and analytical workloads, consider cloud data warehousing services, which offer high performance and scalability for complex queries. You can also consider a unified data platform like BigQuery to power your analytical and ML workloads on structured and unstructured data using open formats like Apache Iceberg.

03 Networking

VPC

Use Virtual Private Clouds (VPCs) to isolate your data lake environment and control network access.

Data transfer

Optimize data transfer between your on-premises environment and the cloud using dedicated connectivity services offered by your cloud provider.

04 Security

Access control

Implement fine-grained access control using Identity and Access Management (IAM) roles and policies to restrict access to sensitive data.

Encryption

Encrypt data at rest and in transit using cloud-provided encryption services.

Threat detection

Use cloud-native security tools for threat detection and monitoring.

05 Data governance and metadata unification

Data governance

Establish clear data governance policies and utilize cloud-based tools for fine-grained access control and compliance. Consider the full data-to-AI governance capabilities of your cloud provider as you think long-term about your modernization strategy.

Cloud-based metastore

Your cloud provider may offer a centralized metadata repository for your data lake, enabling data discovery, governance, lineage tracking, and quality monitoring in a unified, integrated component.

06 Migration tools

Data migration services

Cloud providers offer various data migration services with features like data validation, transformation, and schema mapping. Explore open-source tools or commercial ETL solutions for other migration scenarios—while weighing up the volume and types of data you need to migrate, along with your ideal timeframe and acceptable level of downtime.

Bulk data transfer

For large datasets, consider bulk data transfer tools provided by your cloud provider. Incremental data migration, where only changes are transferred, can be useful for minimizing downtime and maintaining data consistency.

Phase 03

Planning

Thorough planning is crucial for a smooth and successful migration. It's now time to create a detailed project plan—which will serve as a living document that gets updated as you go through the execution phase, documenting what actually happened.

The key elements of a project plan include:

Timelines and milestones

Define your key milestones for the migration, and the expected timeline for each. A phased approach can help minimize disruption. Document the critical data and workloads that should be migrated first, and plan subsequent phases of the migration according to urgency and criticality.

Tasks and dependencies

Break down the migration process into manageable steps, documenting dependencies at each stage.

Roles and responsibilities

Identify key stakeholders and their roles in the migration. Document communication channels for stakeholders to connect, and define escalation procedures should they become necessary.

Budget

Outline the estimated costs associated with the migration, including cloud resources, migration tools, and professional services.

Rollback plan

Document steps to take if something goes wrong and impacts business-critical processes.

Phase 04

Execution

You're now ready to transfer your data lake components to the cloud. Key considerations to keep in mind during this phase include:

- ✓ How to ensure data integrity during the migration
- ✓ How to minimize downtime and disruption to critical business processes
- ✓ How to validate the migrated data and workloads
- ✓ How to handle any issues that may arise

Throughout the migration process, continuously monitor performance and availability to ensure adherence to SLAs. Use the communication channels established during the planning phase to alert stakeholders of any potential issues or disruptions. By aligning your migration efforts with business SLAs, you can ensure a smooth transition, minimize downtime, and maintain the integrity of your critical data and applications in the cloud.

Let's break down the key components of the migration.

Data

To ensure accuracy and completeness, verify data integrity before, during, and after the migration. This might involve comparing checksums, validating data against schemas, and reconciling record counts and other stats. It's also important to choose the right data migration tools—as cloud providers offer various data migration services with features like data validation, transformation, and schema mapping. Finally, remember that data migration is not just about moving bits and bytes. It's also about transforming data to fit the new cloud environment and optimizing it for cloud-native tools and services. This might involve data cleansing, schema changes, and format conversions so your data can be used efficiently in your new data lake.

Metadata

Migrating your metadata is critical for ongoing data discovery and data governance, and to maintain the value and purpose of your data in its new home. Use cloud-native metadata management tools offered by your cloud provider, which often provide features for automated metadata discovery, classification, and lineage tracking. During the migration, maintain clear mapping between your source and target metadata to help preserve data lineage and ensure data can be accurately interpreted and used in the cloud. Don't forget to validate the migrated metadata for completeness and accuracy.

Governance

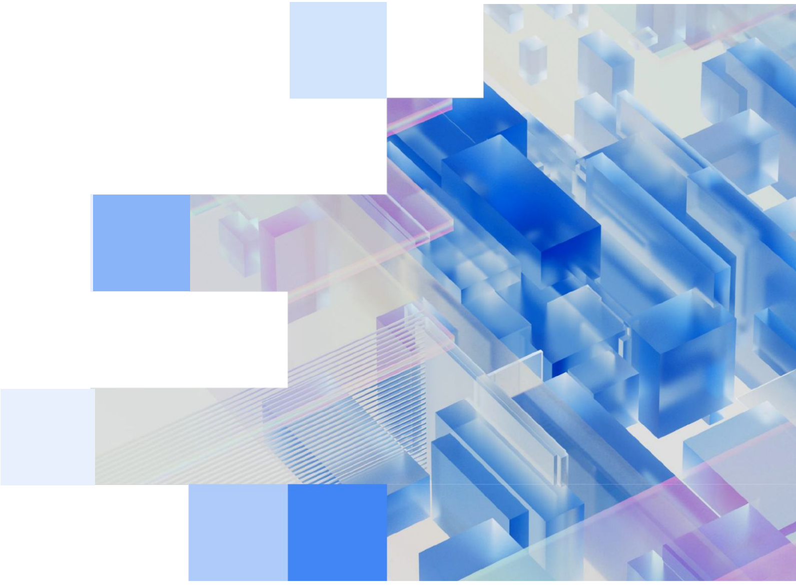
To help you define roles, create user groups, and enforce access control policies for your cloud-based data lake—while ensuring a smooth transition and minimal disruption—use IAM services that provide fine-grained control over user access and permissions. Ensure that data access policies are consistently enforced throughout the migration process and that sensitive data is protected according to your governance policies; and regularly audit user access and permissions to identify and address any potential security risks. By taking these steps during migration, you can create a more secure and compliant environment, while empowering authorized users to access the data they need to get work done effectively.

Workflow migration

Migrating and adapting your workflows and DAGs will help ensure the continuity of your data processing and ML pipelines, and unlock the scalability and efficiency of cloud-based orchestration services. Consider using workflow orchestration services such as Cloud Composer, which provide managed services for defining, scheduling, and monitoring complex data pipelines. These services often offer features like visual DAG editors, version control, and monitoring capabilities. At the same time, you may want to consider refactoring workflows to optimize them for the cloud. Thoroughly test your migrated workflows and DAGs in the cloud environment to ensure they function as expected and meet your performance requirements.

Workloads

Refactoring your workloads could help you take advantage of cloud-native services and optimize performance. You should consider serverless computing, using managed Spark, Flink, and Ray services, or serverless data warehousing for large-scale data processing. Prioritize workload migration based on business criticality and dependencies, adopting a phased approach to reduce risk and minimize downtime. Thoroughly test migrated workloads in the cloud environment to ensure functionality, performance, and data integrity; and use cloud monitoring tools to help identify any bottlenecks or optimization opportunities.



Phase 05

Optimization

The job doesn't end once you've successfully migrated your data lake. Once in the cloud, there's the ongoing opportunity to optimize the environment for both cost and performance. Here are some of the key ways you can do this:

Performance tuning

Continuously monitor and fine-tune your data lake's performance, using cloud-native tools to identify bottlenecks and optimize queries.

Cost optimization

Tactics like adopting serverless products, using appropriate instance types, leveraging spot instances, and optimizing data storage will help you cut costs.

Enhancing oversight and governance

Strengthen data governance with cloud-native tools for data quality monitoring, lineage tracking, and compliance automation.

Using cloud-native services

Continuously explore and adopt new cloud services and features to further enhance your data lake's scalability, reliability, security, and capabilities.

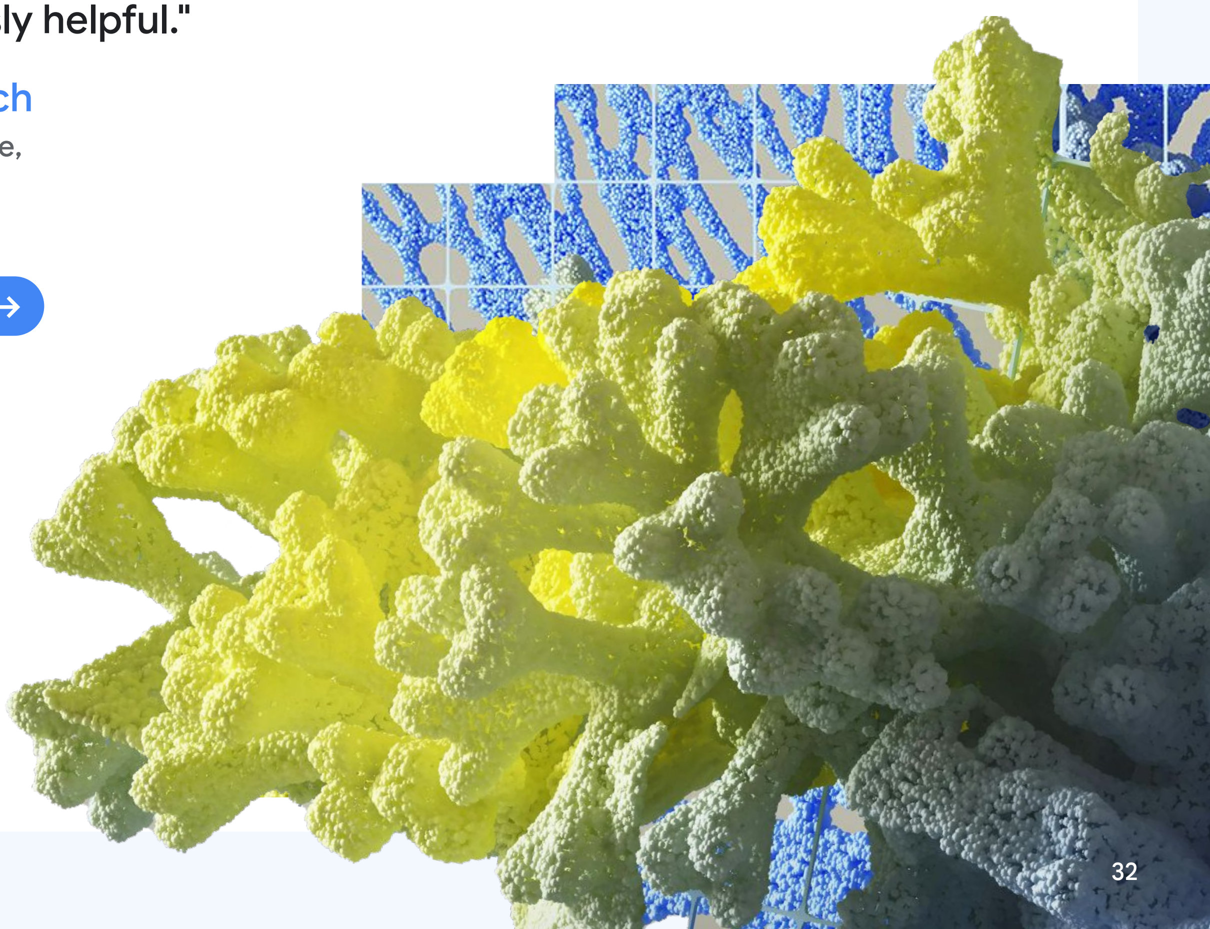
General Mills achieves data lake migration to Google Cloud 30% faster with partner support

Having outgrown its on-premises data and analytics ecosystem and wanting to prepare for AI innovation, General Mills embarked on a data lake migration journey to BigQuery—and turned to Accenture to help bring the vision to life.

The company's transformation programs were originally slated to take nearly three years, but it was able to execute on the plan more than 30% faster in just 21 months with Accenture's support.

> “Accenture was a key part of helping us execute our migration strategy. They already had experience working in Google Cloud and with the broader architecture that we were moving toward. That kind of insight was tremendously helpful."
>
> **Jason Staloch**
> VP of Digital Core,
> General Mills

Read the full story →



A checklist to help you plan

This checklist provides a comprehensive overview of the key considerations for migrating a data lake to the cloud. Don’t forget to adapt this checklist to your specific needs and requirements.

Download the checklist →

Chapter 4

How Google can help

Google Cloud offers a comprehensive suite of services and resources to support your data lake migration project and accelerate data-driven innovation across your organization.

Your Google Cloud account team

Your Google Cloud account team provides personalized guidance and support throughout the migration journey. They can help you assess your current environment, define your cloud strategy, and develop a tailored migration plan. They also act as a single point of contact, coordinating various Google Cloud resources and ensuring your project stays on track.

Google automated migration tooling

Google Cloud provides automated migration tooling to streamline the process of moving your data and workloads to the cloud. For example, BigQuery Migration Services simplifies the discovery of various sources, including on-premises data lakes and data warehouses, and transfer of data from them to Google Cloud.

Google Consulting Services

For complex migration scenarios, Google Consulting Services can provide expert guidance and hands-on support. With a proven track record, Google Cloud consultants have collaborated with customers across industries, successfully completing many complex and large-scale data lake migration projects.

Bringing a deep understanding of best practices, plus the challenges and nuances involved in migrations, they can help assess your needs, design your cloud architecture, migrate your data and workloads, and optimize your data lake for performance and cost-efficiency. At the same time, they can help reduce the risks associated with complex migrations, accelerate timelines, and minimize overall migration costs. Ultimately, Google Consulting Services helps drive better business outcomes by ensuring a successful migration that aligns with your business objectives.

Migration Black Belt Program

To address technical challenges during migration and accelerate the process, Google Cloud offers technical assistance from experienced engineers, including architectural reviews to optimize your cloud data lake design for performance, scalability, and cost-efficiency. Google Cloud's Migration Black Belt engineering team can also provide guidance on best practices for data and workload migration, security, and governance in the cloud.

Google Cloud Partner Program

The Google Cloud Partner Program provides access to a vast network of certified partners with specialized expertise in data lake migration and cloud technologies. These partners offer a range of services, from migration planning and implementation to ongoing support and managed services.

Financial incentives

To help offset the upfront costs of migration, Google Cloud offers various financial incentive programs—which either provide credits towards your cloud usage to help offset the costs of running two systems in parallel, or cover some of the upfront migration costs.

Are you eligible for migration incentives?

Find out here →

Why migrate your data lake to Google Cloud?

Migrating your data lake to Google Cloud offers a compelling combination of advantages over other cloud providers.

Superior data analytics capabilities

Google Cloud is renowned for its leadership in data analytics and AI, offering powerful tools like BigQuery, an AI-ready, unified data platform; and Dataproc, a fully managed and cost-effective Spark service. These tools enable faster and more efficient data processing and analysis across all data types, empowering you to extract valuable insights from your data.

Open and flexible ecosystem

Google Cloud embraces open-source technologies and provides a flexible ecosystem that supports various data lake architectures and tools. This means you can get the most from your existing investments and choose the best solutions for your specific needs.

Innovation and AI/ML

Google Cloud is at the forefront of innovation in AI and machine learning, offering pre-trained models, custom model development tools, and specialized AI/ML infrastructure which you can use to build intelligent applications.

Cost-effective

Google Cloud offers competitive pricing and a variety of cost optimization tools and strategies, helping you manage your cloud expenses and maximize the value of your investment.

Security and compliance

Google Cloud has a strong commitment to security and compliance, offering a comprehensive set of security features and certifications to protect your data and meet regulatory requirements.

Global infrastructure

Google Cloud's global infrastructure provides high availability, low latency, and scalability for your data lake, ensuring reliable performance and access to your data from anywhere in the world.

Ready to take your next steps?

To get started on your data lake migration, reach out to your Google Cloud team. Don't miss out on our limited-time incentives program—designed to accelerate your migration journey.

Fill out this form today and unlock the full potential of your data with Google Cloud.

Contributors:

Angela Soares
Product Marketing Manager,
Google Cloud

Adnan Hasan
GTM Strategist, Analytics Data Platform for AI,
Google Cloud

Sajal Agarwal
Senior Outbound Product Manager,
Google Cloud

Jill Hardy
Product Marketing Manager,
Google Cloud