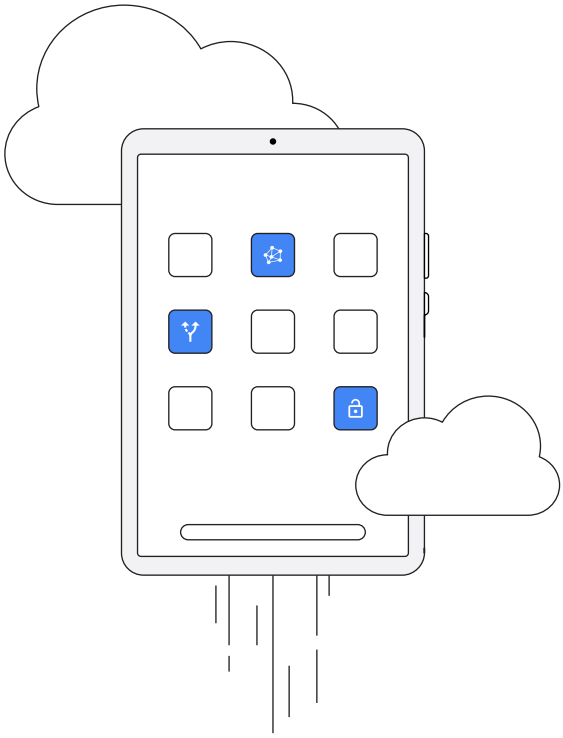
Google Cloud

The future of data will be unified, flexible, and accessible

Tech companies and startups are learning that to be successful:

- Data must be *unified* across their entire company, and even across suppliers and partners. This involves unlocking unstructured data and breaking down organization and technology silos.
- Their technology stack must be *flexible* enough to support use cases ranging from offline data analysis to real-time machine learning.
- The stack must also be *accessible* from anywhere and everywhere. It must support different platforms, programming languages, tools, and open standards.



Google Cloud

Chapter 1

Why making the most of your data can be a competitive advantage

Everyone recognizes that data is important, but very few companies are able to extract innovative business and customer insights from their data. What does it mean to get the most from your data? Why is it a challenge?

If you are making the most of your data, it means that you can make product and operations decisions using data. So, ask yourself a few questions. Do you know how your customers' expectations are changing? Are you using data to improve the customer experience? In terms of the challenge, ask yourself where your data engineers and scientists are spending their time today?

Data is crucial to driving innovative product direction and user experiences along with broad go-to-market decisions. Harnessing your data successfully can give you a significant competitive advantage. That's why most tech companies and startups are under tremendous pressure to do more — to modernize and operate at larger and larger scales, to justify current and future data costs, and to elevate their organizational maturity and decision-making.

However, there are challenges surrounding access, storage, inconsistent tools, compliance, and security that make it hard to go below the surface and unlock real value from your data.

Maybe you've inherited legacy systems that you're trying to marry with new ones. Should all your data be in one cloud? Or should it be distributed across multiple clouds? How do you modernize analytics stacks (that historically have been vertically integrated) to work with platforms that can scale horizontally?

Or maybe you're batching or micro-batching your data today instead of processing it in real time. The resulting orchestration system and scheduling add complexity to your architecture and require maintenance around contention and resilience. The operations overhead from managing and maintaining the batch architecture is expensive, and you're still compromising on data latency.

Lacking easy access to all your data and missing out on the ability to process and analyze it as it comes in puts you at a disadvantage. The modern tech stack needs to be a streaming stack that keeps up with your data's scale, uses the most recent available data, and incorporates and understands unstructured data. And the most advanced analytics teams have shifted their focus from operation to action by using AI/ML to experiment and operationalize processes.

Google Cloud

Chapter 2

How to get your data to work for you so you can focus on innovation

What does it mean for your data to work for you? It means improving the customer experience, reaching new customers, and growing your topline. Primarily, it's about being able to innovate. We recommend two principles for choosing a data platform that will help you achieve these outcomes.

Principle 1: Simplicity and scalability

It's likely that you have a lot of data at your disposal right now. Perhaps it's growing exponentially and you want to maintain or increase your ROI while keeping up with volume. Maybe you're anticipating how much data you'll have in the future (e.g. a terabyte) and designing your systems to process that amount while knowing that if growth exceeds those expectations you'll be looking at a wholescale system migration. Or maybe you've chosen a data warehouse that can scale to your expected growth, but increasing processing needs are making it complex to manage.

Smaller systems have generally been simpler. However, you no longer have to choose between a system that's easy to use and a system that's highly scalable. Using a serverless architecture eliminates the need for cluster management and gives you the ability to handle massive scale for both compute and storage, so you never have to worry about data size exceeding your technical capacity again.

For both simplicity and scalability, we recommend a serverless data platform. We suggest you discard any option that requires you to install software, manage clusters, or tune queries.

Principle 2: Agility and keeping costs down

Any data management system that combines compute and storage will force you to scale up compute to deal with increasing data volume, even if you don't need it. This can be expensive, and you might find yourself making compromises such as only storing the last twelve months' worth of data in your analytics warehouse. You might also choose not to include data because you don't have an immediate use case for it, only to find down the road that you can't test a hypothesis because the data isn't there and would require a new pipeline to get started.

Other systems get you halfway there, letting you scale and pay for compute and storage independently but still making you manually set up, scale, and optimize clusters. To reduce infrastructure management as much as possible, consider a serverless, multicloud data warehouse with enhanced reliability, performance, and built-in data protection (such as BigQuery).

Beyond cost and management, you also want to think about agility. When your data changes, how long does it take for you to notice and react? When there's a new version of some software or a tool that you use, how long does it take you to embrace its new capabilities? The path to greater agility is to choose flexible tools that require less handholding and are applicable to a wide variety of workloads.

Queries on systems such as Redshift have to be optimized to be efficient. This limits the amount of experimentation you can do, so you might only extract and pull in data when you suspect there might be a problem. The compromises you make due to the lack of compute/storage separation and the need to optimize your data warehouse tie one hand behind your back.

With something like BigQuery, you don't need to plan queries in advance or index your data sets. Decoupled storage and compute let you land data without worrying that it's going to drive up your querying costs, and your data scientists can experiment without having to worry about clusters or sizing their data warehouses to try new ideas through ad hoc queries.

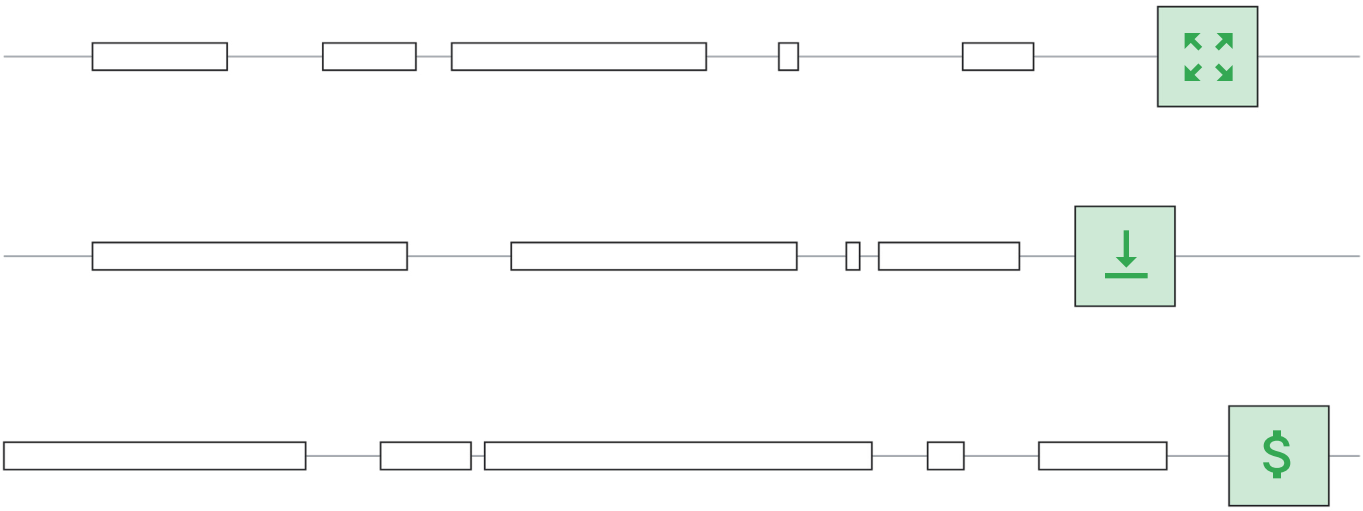
We've looked at how a simple, scalable, flexible, cost-effective platform puts you in a position to innovate. Now we'll discuss how your data can help make it happen.

Google Cloud

Make data-driven decisions in real time

The pace at which business operates keeps accelerating. Customer expectations have also changed. Where once you could reconcile a transaction or approve a return in three days, you now have to provide answers immediately. Faster, more timely decision-making is leading to an increased need for streaming.

You want to be able to capture data in realtime and make that data available for low-latency querying by your business teams. You also want to make sure your streaming pipelines are scalable, resilient, and have low management overhead. This is the only way your team can react in real time at the speed of your business. It won't surprise you to learn that BigQuery has native support for ingesting streaming data and makes that data immediately available for analysis using SQL. Along with BigQuery's easy-to-use Streaming API, Dataflow gives you the ability to manage your seasonal and spiky workloads without overspending.



Break down data silos

Many organizations end up creating silos because they store data separately across departments and business units, with each team owning their own data. This means that whenever you want to do analysis that spans departments, you have to figure out how to break down those silos, probably by running extraction (ETL) pipelines to get data and land it in your data warehouse. But departments that own the data often have little incentive to maintain the pipelines; over time these go out of date and the landed data becomes more obsolete and less useful.

Beyond organizational silos, many companies today have adopted a multicloud strategy based on departmental preference, capability alignment, and regulatory pressure. These companies often also deal with the reality of legacy data lakes and data warehouse investments that live on-prem. Today's multicloud, hybrid-cloud reality requires another level of sophistication in managing and accessing siloed data.

Moving to a distributed warehouse with a common control pane — sometimes called a data fabric or data mesh — increases your ability to access high-quality data across departments, clouds, and on-prem systems. This can solve business problems such as product performance or customer behavior, and empowers you to query data on the fly.

BigQuery provides the technological underpinnings of such a data mesh — users across your organization can manage, secure, access, and share data assets and insights regardless of who in the organization owns the data. For example, you can land all your data in BigQuery and provide reusable functions, materialized views, and even the ability to train ML models without any data movement. This means that even non-technical domain experts (and partners and suppliers who have permission) can easily access and use SQL to query the data using familiar tools such as spreadsheets and dashboards.

The 'hub and spoke' analogy is appropriate here. BigQuery is the hub that contains your data. The spokes are reporting tools, dashboards, ML models, web applications, recommendation systems, and more — all of which read data live from BigQuery without having to copy it. Looker, for example, helps you visualize your data and integrate it into users' daily workflows. This approach lets you improve your data's usability, security, and quality at the same time.

Simplify access to all your data

Historically, unstructured and semi-structured data was best served by data lakes, while structured data fit best in data warehouses. This separation created technological silos that made crossing the format divide difficult; you'd store all your data in a data lake because it's cheaper and easier to manage, then move the data to a warehouse so you could use analytics tools to extract insights.

The increasingly popular 'lake house' merges these two worlds into a unified environment for all types of data; you can use BigQuery as both your data warehouse and your data lake. BigQuery's Storage API allows you to access storage directly to power workloads usually associated with data lakes. Because data can be stored in BigQuery as a single source of truth, fewer copies need to be created and maintained. Instead, you can carry out downstream processing via SQL transformations that are stored in logical views without having to move data around.

Ease of use matters — if you can get results from queries in 30 seconds rather than 30 minutes or 3 hours, you'll likely make more use of data in your decision-making.

Use AI/ML to experiment faster and operationalize workloads

How quickly are your data scientists able to experiment? Chances are that they need to stop development and operationalize their models in order to evaluate their experiments with real users. They develop and iterate on a model using historical data before handing the model off to the engineers, who often completely rewrite it to incorporate it into the production system and carry out A/B testing. Then they wait, iterate on their model, and productionize again. This cycle involves a lot of stop-and-go and rewriting of code, with all the necessary coordination between teams introducing errors along the way. Your data scientists aren't experimenting as much as they could because it can take a long time to experiment this way. This makes it hard to predict how long a project will take and whether it will be successful, let alone how long it will take to get into routine use. To go beyond this, you need to provide your data scientists with powerful but familiar tooling. Vertex AI Workbench allows data scientists to work effectively in Jupyter notebooks, but get accelerated training, quick experimentation, and rapid deployment.

If you're serious about differentiating based on data, you want to extract the highest value you can from the data you're collecting. To do that, you want your data science teams to be as productive as possible and not miss opportunities to build a model because even simple things take too long or are too hard.

The quality of your pre-built and low-code models is crucial. AutoML on Vertex AI makes best-of-class AI models available in a no-code environment, which allows for fast benchmarking and prioritization. Having pre-built models such as Entity Extraction or Vertex AI Matching Engine on your own data significantly speeds up value creation from data; you're no longer limited to just classification or regression.

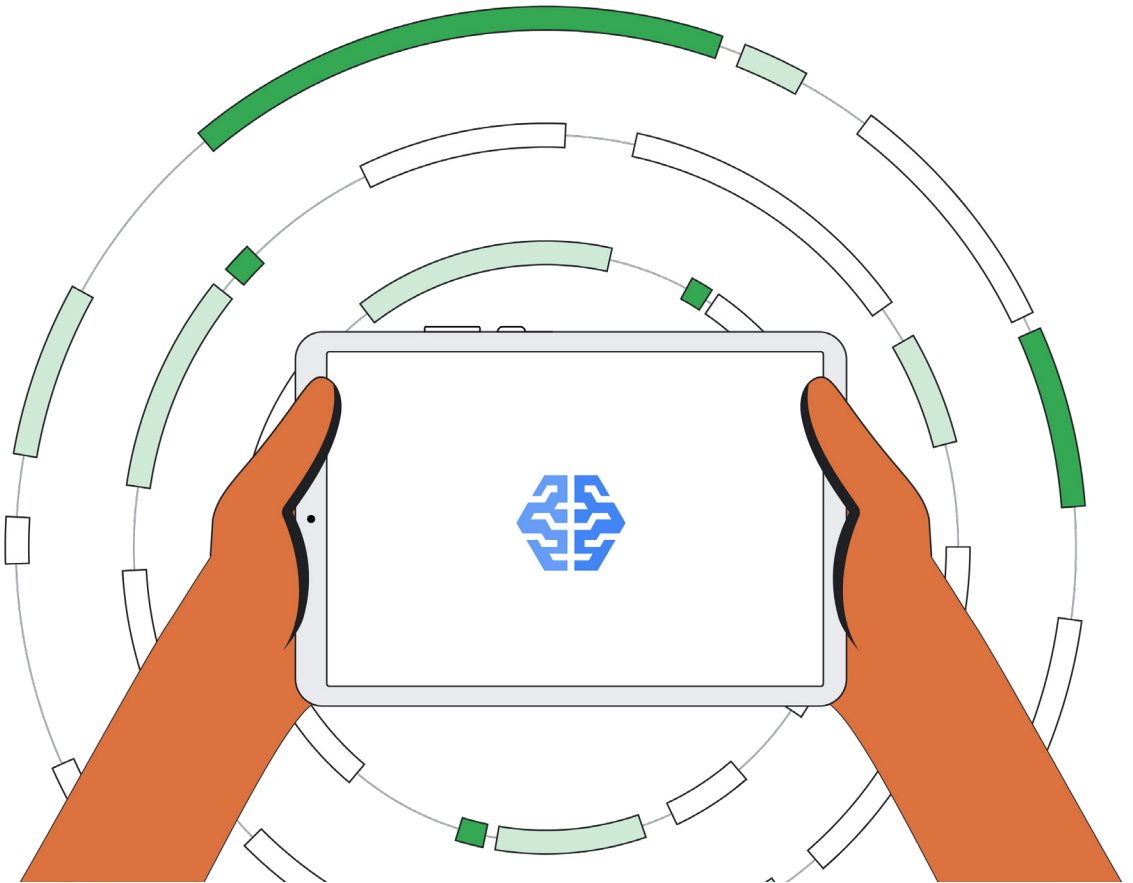
The key to maintaining your data agility is to do end-to-end experiments early and often. Vertex AI Pipelines give you a history of experiments that let you look back, compare against benchmarks and endpoints, and A/B test with shadow models. Because the code is containerized, the same code can be used between development and production systems. Data scientists work in Python and production engineers get containers that are fully encapsulated. Both teams can standardize by operationalizing the models with Vertex AI Prediction and you can move quickly.

Domain experts can often use BigQuery ML to test an idea's feasibility by training custom models using only SQL without needing additional experience with traditional data science tools. This means you can experiment in a production-like system and conduct feasibility studies in a matter of days instead of months. The BigQuery ML model can be deployed into Vertex AI for all the advantages we just discussed. You can use Looker to create consistent data models on top of all your data and use LookML to query data, meaning that everyone in the organization can create easy-to-read reports and dashboards to explore patterns of data.

To drive real value in production, systems must be able to ingest, process, and serve data, and machine learning must drive personalized services in real time based on the customer's context. However, a continuously running production application demands that models be constantly retrained, deployed, and checked for security. Incoming data requires preprocessing and validation to make sure there are no quality issues, followed by feature engineering and model training with hyperparameter tuning.

Integrated data science and machine learning are essential for easily orchestrating and managing these multiphase ML workflows and for running them reliably and repeatedly. MLOps tooling and automated workflows enable rapid continuous delivery and simplify management of models to production. There's a single workflow and vocabulary for all our AI products regardless of the layer of abstraction, and you can easily interchange custom and AutoML models since they leverage the same format and technical foundation.

For example, what if you want to apply anomaly detection to live, unbounded data streams to combat fraud? With the right approach, you'd generate a sample data stream to simulate common network traffic and ingest it to Pub/Sub, then create and train an anomaly detection model in BigQuery using BigQuery ML K-means clustering after masking personally identifiable information (PII) using DLP. You'd then apply the model to live data for real-time detection using Dataflow and would use Looker to create a dashboard, alerts, and actions to handle the identified events.



Google Cloud

Chapter 3

Why choosing a well-rounded data warehouse option matters

We've talked about BigQuery and Redshift, but these are not the only data warehouse options available. There are other data analytics products (such as Snowflake and Databricks) that work across all three major clouds. So if you pick BigQuery, is cloud lock-in an issue?

The first thing to note is that with BigQuery you're not limited to analyzing only the data you have stored in Google Cloud. BigQuery Omni gives you the ability to seamlessly query your data in Amazon S3 and Azure Blob Storage from the Google Cloud console.

The reality, though, is that if you use Snowflake or Databricks, the switching costs of moving from AWS to Google Cloud or vice versa are lower. But what about the cost of moving to another data warehouse? What if you want to move from Snowflake to BigQuery, or from Databricks to EMR? There's still a switching cost; it's just a different scenario.

Because there's going to be a switching cost in any scenario, you ultimately want to choose the tool or platform that's going to work for you in the long term. You're making a choice based on a given platform's differentiating features, cost today, and the rate at which it will add innovation in the future. When you choose Snowflake, you're betting that a company focused on data warehousing will give you faster innovation in that space. When you pick BigQuery, you're counting on a company known for inventing many data and AI technologies to continue to innovate across the platform.

We believe that an innovative, well-integrated platform better powers the flywheel effect of innovation. When a managed service offering such as Google Kubernetes Engine (GKE) makes container images load faster, that helps Serverless Spark work better, and because Serverless Spark can operate on data in BigQuery, it makes BigQuery more valuable to you. The flywheel spins faster when you bet on a platform rather than on individual products.

Google Cloud

Chapter 4

How to approach your data migration journey with confidence

How long will your data migration take? Six months? Two years? How much effort does that represent, and is it all worth it?

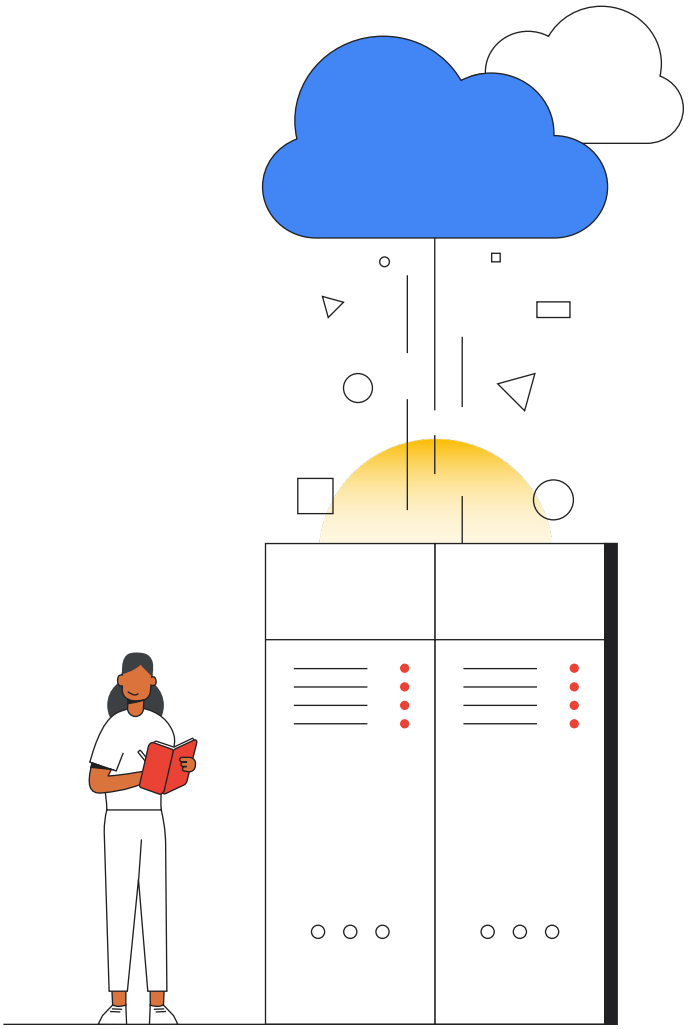
If you're migrating from one cloud to another, that's likely to be easier than migrating from on-prem to cloud, simply because you'll usually have a lot more technology depth on-prem. Regardless, focus on your goal, which is usually something like 'How fast can I innovate?'

Think about all the innovative things you want to be doing that you aren't doing today, then set up new projects and transfer the data that you need to carry them out. We can help you build these new use cases and mirror the data sources that you'll need. For a while you'll be in a hybrid environment in which many use cases run on premises, but are driven by data that's mirrored in real time or in batch from your on-prem environment or your other cloud provider.

Your second consideration is around cost. Look at the really expensive Teradata instances that you're running. We see customers cutting their costs in half by switching to BigQuery, and these migrations are much easier than they used to be due to automated assessment tools and automated SQL transpilers that convert the vast majority of your scripts. We have ways to virtualize things so that your clients think they're talking to Teradata when they're actually talking to BigQuery. There are lots of ways we can help you migrate without having to shut everything down; you can use those migration tools to move away from your expensive Teradata and Hadoop workloads.

The third consideration is to look at your ERP systems, such as SAP, Salesforce systems, and Oracle. If you want to optimize your supply chain, carry out lead scoring, or detect fraud, it's important to be able to connect your analytics workloads to your ERP systems. There are third-party connectors that we can use to get data from those systems, which we can then utilize to build modern AI-driven use cases on that data in the cloud.

The order in which you do these things depends on your situation. If you're a startup, you might begin with innovation, move on to cost optimization, and finally take advantage of existing pipelines and connectors. If your business has a significant dependence on supply chains, you might start with the ERP connectors. Regardless of the order in which you do all three, you'll find that you've moved a considerable amount of your valuable data estate into the cloud. Now look at what's left over and consider whether it's worth moving at all. Often we find the answer is no — once you've moved the 70-80% of workloads that are truly necessary, you need to start making hard decisions. Is the remaining 20-30% worth migrating, or should you consider rewriting or doing the task differently? You don't want to get into the mode of moving everything to the cloud as-is, or you'll find yourself replicating all the technology debt that you had on-prem in your new cloud environment rather than focusing on data value.



Ready to take your next steps?

We've talked a lot about harnessing your data and what that really means, along with some considerations you might face while migrating to a data warehouse in the cloud.

To learn more about how Google Cloud can help you use insights to gain a significant advantage, help your company drive down costs, and increase productivity by optimizing your use of data and AI, please get in touch.

Talk to an expert

Further reading

- Learn what type of data processing unit you are
- To learn more about the elements of how to build an analytics data platform depending on what type of organization you are, read our paper here.