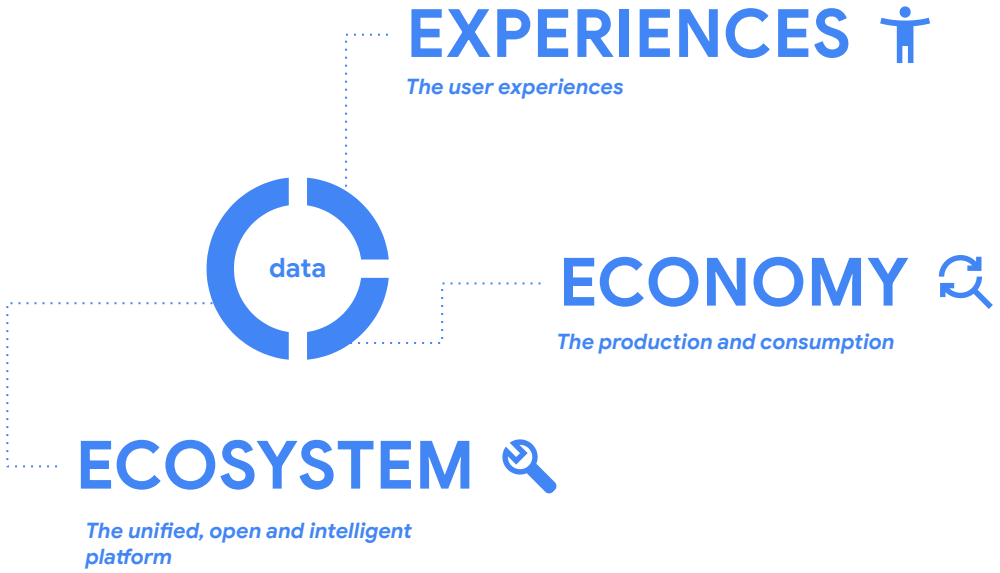
Google Cloud

White paper
February 2023



Three pillars for building a modern data strategy

Thomas de Lazzari, Sina Nek Akhtar, David Montag, Andreas Ribbrock, Zara Wells, Firat Tekiner

Contents

Foreword to the reader

Our intent for this paper is to provide a collection of resources for data and business professionals to inform, implement and execute a value-driven data strategy across your organization.

In our work with customers we have identified several challenges they commonly face:

- Users across the organization lack access to useful data experiences that match their maturity level and business needs.
- Data assets are not leveraged to build incremental value due to a lack of ownership structures and reliability practices.
- Challenges stitching together data tools into an ecosystem that is integrated, secure, easy to operate and govern, compliant with global and local regulations, with simple commercials, and where employees feel their investments in learning will pay off.

Every organization's journey to become more data-driven will be unique. In this paper we discuss strategies for organizations based on their culture. For example, a business with decades of legacy is not expected to perform in the same way as a digital business which has built its processes in the cloud-native era – the expected outcomes and success metrics for the organizations would be different.

Our hope is that this paper will encourage closer collaboration across your business units to determine the most suitable data ecosystem to create accessible data experiences in a data economy.

Introduction

A strategy outlines the initiatives and actions that you believe will drive your desired business outcomes. The purpose of a data strategy is to enable your organization to achieve its mission and objectives using data – giving you a competitive advantage.

This paper covers three key factors to consider when establishing your data strategy (the "3 E's"):

Data Experiences

Productive user experiences enabling all users to access and create value from relevant data

Data Economy

Principles and practices to ensure that data can be published, discovered, built on, and relied on

Data Ecosystem

A unified, open and intelligent platform with end-to-end data capabilities for all users and needs

Organizations guided by a robust data strategy achieve concrete business benefits that include:

- **Accelerated product development** and delivery to market
- **Greater organizational efficiency and agility**, and ability to execute on innovation
- **Increased productivity with talent acquisition**, retention, and development
- **Ability to experiment** and use the next generation of intelligent solutions leading to better decisions
- **Creating differentiated solutions** driven by data and AI

The purpose of a data strategy

Data is a core asset to all organizations. To derive value from data, it must be accessible, actionable and secure wherever and whenever it's needed in your organization.

Every organization is unique and requires a tailored approach – the success metrics for digital natives will differ from a 50 year old brick and mortar business.

A data strategy helps you create the necessary alignment across your organization to create more value from data.

These are examples of activities that your data strategy should drive:

- Principles and processes to guide the organization toward faster decision-making and continued alignment with business goals and objectives
- Continuous review of recommended systems and tools that align with your strategy's vision while avoiding a one-size-fits-all approach
- Clear and consistent policies and procedures for managing data securely throughout its lifecycle, from creation to disposal
- Ensuring that data is used ethically and responsibly, in compliance with relevant laws and regulations
- Creating and enabling a culture of data-driven decision-making through the use of accessible, governed data to drive business value
- The responsible use of AI across your organization

Data Experiences

These are curated environments built for users to access, explore and succeed in finding actionable insights from data. Effective data experiences integrate into a user's workflow and develop in line with increasing data literacy. This enables all employees of varying literacy levels to be immersed in a data-driven culture and to be provided with the opportunity to make data-informed decisions.

Here are some things to consider when building data experiences:

- **The role of culture** – Guiding the desired behaviors of the organization
- **Product-oriented organization** – Bring the business and technical domains together to provide optimal value to the user
- **Data literacy** – Varied literacy levels across the organization will reflect the complexity of the experience
- **Be data-informed** – Applying critical thought to the way you use data
- **Data principles** – Key guidelines the organization sets in relation to the use of data
- **Data privacy** – Managing the lifecycle of your data in compliance with regulations and ethical guidelines

Did you know?

Implementing a data culture can help organizations become more agile, responsive to customer needs, and open to innovation. Read more here.

The role of culture

An organization-wide shift is required to create and foster a data-driven culture, with change management (the 'people' side of change) playing an important role. A common pitfall when adopting new technology and ways of working is ignoring the role of culture and the change management practices required to affect it.

The idea that there is a "one size fits all" or "best" approach to data analytics and data processing is a legacy from a time before today's data-centric, cloud-first world.

Just as in professional sports, successful organizations adopt a strategy that makes sense for the people and skill sets they have. If a sports team has a great defense, they should try to win through defense, not by copying the offensive strategy of a different team with different players.

Similarly, if an organization has a strong bench of data analysts, they should leverage these people instead of trying to transform into an organization full of data engineers.

Think about how you can add to your culture rather than entirely changing it. Take a programmatic approach and use practices like data principles (see "Data principles" section below) to help guide the organization forwards.

Finally, strive to create a psychologically safe work environment where team members are not afraid to collaborate and safely fail. This will enable your teams to experiment and try new practices and tools.

Did you know?

Your data strategy will be influenced by which type of data processing organization you are. Read more here.

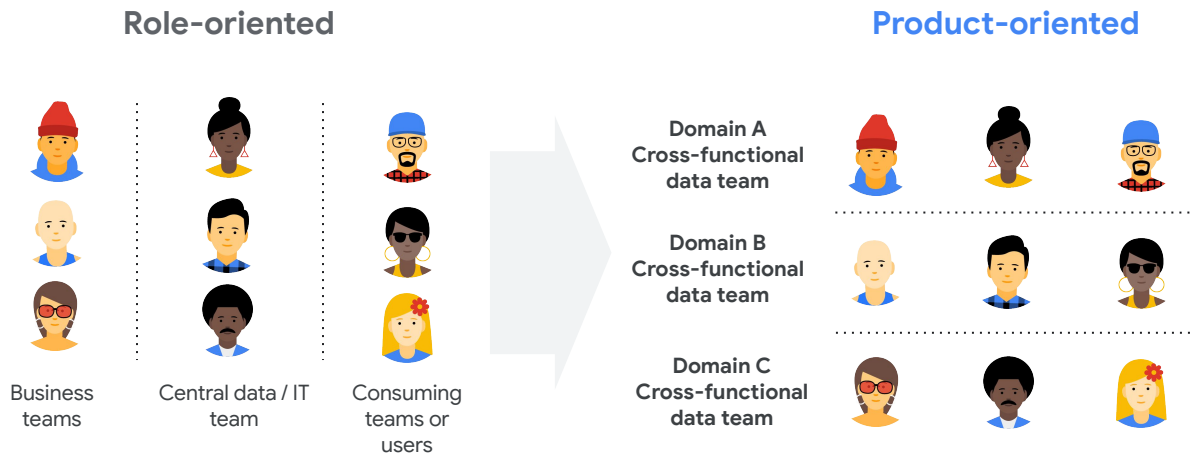


Figure 1: Shifting from a role-oriented to a product-oriented organization

Product-oriented organization

Product thinking helps bring together the business and technical aspects of the delivery in a data-driven organization.

Consider an example — your business teams request particular reports or write requirements that get handed over to a technical team to implement. That team may then work with a data team to implement the request. Finally, something is built and gets into the hands of the consumers. Inevitably, adjustments need to be made (this is normal), and requirements, development, and data need to be adjusted. Each hand-off of information between teams adds delays, reduces the fidelity of the requirements, and adds risk. *See figure 1 (left diagram).*

By shifting to a product-oriented organization you can put the business stakeholders, the data practitioners and the developers in the same team. If the technical team does not understand a requirement, it's usually easily resolved within the team, and the team has a shared knowledge base and vocabulary of business terms. *See figure 1 (right diagram)*

Data literacy

Data literacy is the ability to understand, analyze, argue with and base decisions on data. As data literacy increases, so does the application of critical thought which in turn increases the validity and quality of analysis.

The level of data literacy can often vary widely across an organization, with the highly data literate usually only constituting a small degree of the population, and often limited to a handful of units.

There is large untapped potential across less data literate departments. By providing access to relevant data experiences and improving data literacy levels, organizations can benefit from wins such as reduction of operational overhead and diversity of thought.

It's helpful to incorporate new data experiences into existing tools and processes. This includes providing access to relevant data, but also ensuring that new insights and AI is being put to use within the tools and processes that are being used on a day to day basis by the organization.

Analytical maturity

Everyone benefits from learning basic data skills such as understanding that the quality of the data they enter in systems ultimately impacts how well the organization can derive value from that data.

In the data realm, we consider three levels of analytical maturity in an organization:

- **Hindsight** – Looking at reports to see *what* happened. Assumptions, gut feel, and experience determine what action is taken.
- **Insight** – Exploring the data to understand *why* something happened. Applying critical thought to test and validate their findings such as questioning timestamps and ruling out correlation.
- **Foresight** – *How* the outcome was influenced. Users understand that data can be used for more than looking backward. It can be used to inform their strategy, influence desired outcomes, and create new business value.

Organizations should empower all employees to operate on the **insight** level as a minimum.

Example. Imagine giving three different people a potted plant with browning leaves. Notice the difference in approach according to their analytical capability, the actions they take, and the outcomes they influenced (if any).

Person A: (What). Observes browning leaves and removes them. They operated with **hindsight** and acted on what they could see.

Person B: (Why). Seeing the brown leaves, they explored other data points from the environment and discovered the soil was highly acidic. They operated with **insight**, seeking to understand and diagnose why something was happening by collecting and analyzing data with critical thought.

Person C: (How). Not only did Person C diagnose the issue, but they also analyzed the events that lead to the incorrect soil quality to prevent this from occurring in the future. To confirm their findings, they compared the same data points from environments of healthy plants of the same species. Based on this analysis, they were able to recommend regular monitoring of specific data points and automate the addition of feed. They operated with **foresight**, questioning various data points critically to determine preventative measures.

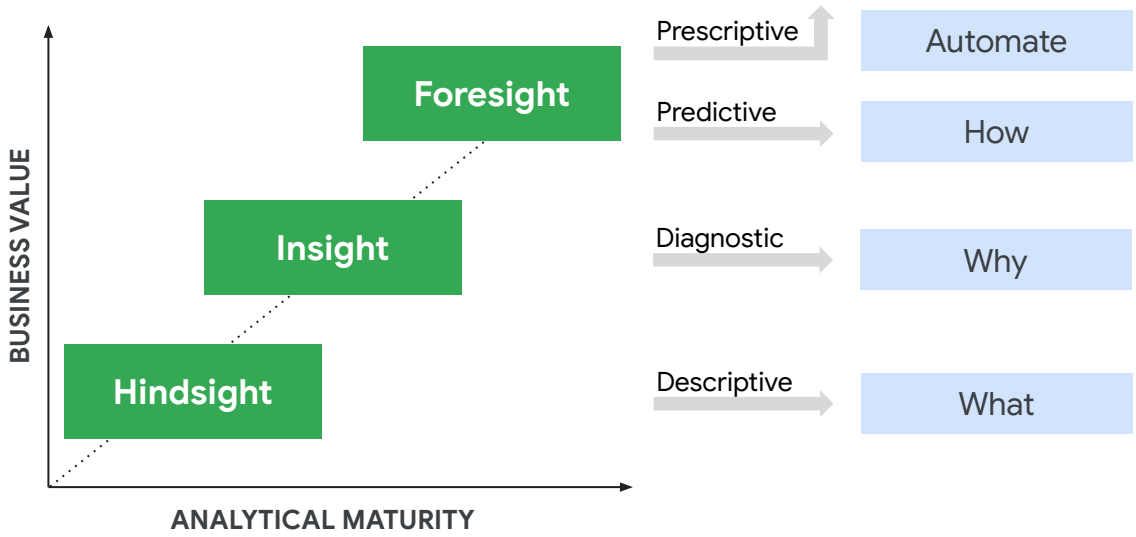


Figure 2: Differing levels of analytical maturity in action

Data and AI academy with pathways for everyone

Create pathways for everyone with a data university program. Data beginners are often over-represented in the business. If you can improve data literacy for everyone, your average data literacy will be higher than if you only invest in the already highly skilled groups.

Business users should develop a high-level understanding of data analysis, and technical team members should develop a good understanding of the business context. This could be done with an internal training class where employees learn about the basics of data analysis, how to discover data, how to process data, and how to present it.

The tools, processes, and enablement need to be relevant to the learner. Especially for new learners, it is important to manage anxiety by pairing concepts and terminology with some simple and safe exercises where they can apply their knowledge.

Make sure to communicate and show that everyone has the potential to become more impactful at work through increased data literacy. Data is everyone's business — today there is no reason why the CMO, COO, or CFO shouldn't be as data literate as the CDO.

Did you know?

Airbnb democratizes data science with Data Academy, a data education program for anyone at Airbnb that scales by role and team. Read more.

Business glossary

The business glossary is the single source of truth for all business metrics. It defines your terms and guidelines for how you interpret and visualize data in your organization. This allows teams to communicate their findings to others in a clear and concise manner, which is necessary for achieving a shared understanding of what the data is saying and whether it can be used for making a certain decision.

Having a well-curated business glossary helps users from various backgrounds to match their business problem to the right data, and helps to communicate more effectively. For example, gross profit, operating profit, and net profit all mean different things, and if you don't know which profit you are using your calculations could be incorrect.

Did you know?

Looker's semantic model enables complex data to be simplified for end users with a curated catalog, pre-defined business metrics, and built-in transformation. Read more.

Be data-informed

One of the practices that a strong, data-driven culture can unlock is growth hacking. It is the practice of rapidly conducting experiments using methods like A/B testing and quickly adjusting based on the data. A complementary approach to growth hacking is to have a data-informed human in the loop. This puts more emphasis on the use of critical thought in experiments versus being blindly guided by the numbers.

A data-informed approach can help you consider what data you don't have today, such as data on potential new users you want to attract to your business.

Data insights should raise the questions that will lead you to formulate hypotheses based on your product intuition and critical thinking. Define ways to test those hypotheses using the data you collect. This will help you avoid growth hacking problems such as selection bias, confirmation bias, and reporting bias.

Did you know?

One way to reveal biases in your data: Google created the What-If tool to give people a simple, intuitive, and powerful way to experiment with a trained ML model on a set of data through a visual interface. Read more.

Data consumers must interpret data and communicate around it accurately. To do this, they must be able to self-serve information about the domain the data belongs to, and what the data means in the context of the domain. This can allow easy interpretation of an experiment's results, for example.

Having a business glossary and ensuring that data owners and stewards keep metadata accurate and fresh is a cornerstone of data interpretability and explainability.

Data principles

Principles help the organization move in the same direction, simplifying the small decisions that need to be made and encoding the organization's values and culture. In essence, principles are a license to behave in a certain way.

Here are some example principles (these may or may not be right for you):

- Use cloud managed or serverless services whenever possible
- All teams have the skills needed to work with data
- Default to storing all data (when permitted), then decide on retention
- Treat data like a product, even internally
- Every data product has clearly defined ownership and business goal

A useful method for helping uncover your principles is known as "positive provocation". Put the word "should" or "what if" in front of any statement. This will generate some response, perhaps as simple as yes/no. Then you ask "why?". The discussion that follows will help surface the principles behind the reasoning.

Figure 3 below includes some example questions to get you started.

These are example questions you can use in a positive provocation exercise to help you derive your organization's data principles.

Should all data be captured?
Should all raw data be saved?
Should all data be processed the same way?
Should it be possible to destructively alter collected data?
Should adding new data sources be easy?

Should all data have classifications?
Should all data be versioned?
Should all data conform to strict data models/schemas?
Should all data be tagged with informational metadata?
Should metadata be crowdsourced in the organization?
Should we know why a piece of data is being collected?

Should all data be owned by somebody?
Should all data be owned by the team that created it?
Should all data have a managed lifecycle?
Should all data be discoverable or cataloged?

Should data speed through the system matter?
Should the quality of data be assessed and made visible?
Should all data be kept equally secure?
Should privacy principles or guidelines always be followed?

Should data producers have to care about who consumes?
Should data consumers use the same system to access the data?
Should all data be discoverable and usable within the organization?

Should all teams have data competency?
Should eng resources be spent on building data infrastructure?
Should teams rely on repeatable patterns and code?
Should we track the cost of data?
Should training be mandatory?
Should machine learning be applied?
Should cloud-native services always be preferred?
Should managed services always be preferred?

Figure 3: Example questions to get you started on uncovering your principles

Data privacy

Organizations are required to maintain compliance with regulations set by governing bodies. This requires the development of privacy policies that govern what data you store and how you treat it.

In your data privacy documentation you should outline the standards your organization upholds in relation to data being collected, stored, accessed, audited, used, retained, and deleted.

Teams benefit from guidance on data privacy. For example, if a team cannot justify the collection of certain data from users, then perhaps it should not be collected at all. Similarly for sensitive data that you have collected, if teams cannot demonstrate needing that data for a justifiable business purpose, then you should limit access to it.

The level of consent a user may have given at a certain point in time needs to be managed, and data management processes should adhere to changes in user consent automatically. Also, having an approach to data deletion in your data strategy is important since users may have the right to be forgotten. Simply bolting this onto existing processes can be challenging.

Data privacy needs to be built together with the business functions of your organization, including data privacy officers, internal audit team, risk managers, senior management, engineering, legal, and ethics boards.

Data Economy

> *Economics is the social science that studies the production, distribution, and consumption of goods and services.* Wikipedia

The more teams can build on each others' work without reinventing the wheel, the more efficiently and quickly the organization can move. Cultivating rich flows of data across teams is key to creating a data economy across the organization. Let's take a look at some aspects that make a data economy work.

In this chapter, we give you a vocabulary to start thinking about the data economy:

- **The data economy** – Unlock deep collaboration across teams using data
- **Compounding benefits** – Teams that can rely on each others' data are able to innovate faster
- **Distributed responsibilities** – When teams rely on each other for data, you need to think about the impact that changes can have
- **Decoupling the platform** – Managing the platform separately from the teams using it with self-service
- **Reinventing the data platform team** – The team that owns and obsesses about how everyone else will get their work done with data
- **Data products and domains** – Clear data ownership across a distributed organization
- **Distributed data governance** – Practices that help keep the data economy healthy

The data economy

Your data strategy can unlock the data economy, which brings a focus on data accessibility, discovery, usability, and reliability.

In a data economy, you improve on and enrich data that is already available from other teams, be it through training a machine learning model on diverse datasets or combining data from across the organization to find new insights and build better products. You also publish your own valuable data for others in the organization to consume.

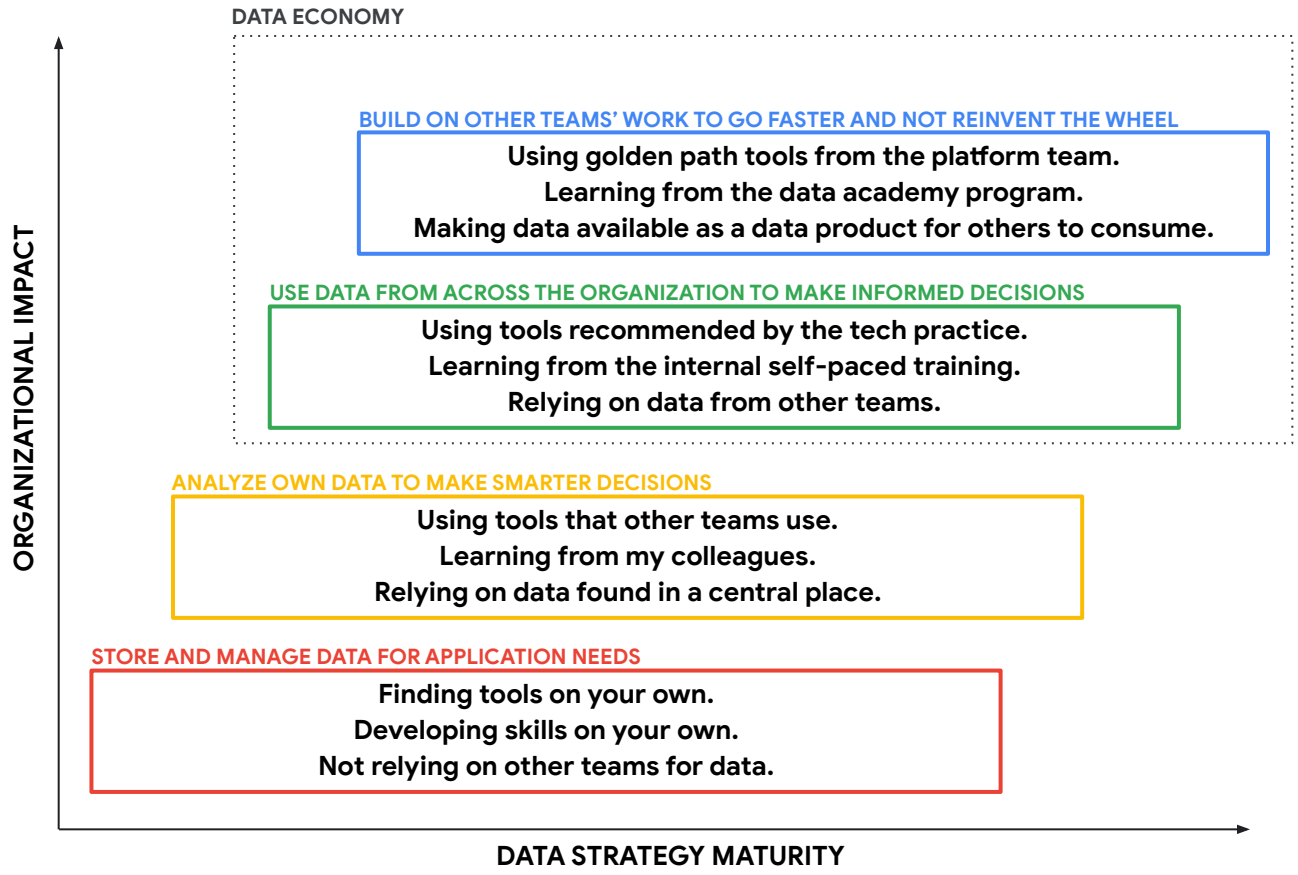


Figure 4: Categorization of data usage maturity and its corresponding organizational impact

To achieve a data economy you must consider changing your current ways of working. Organizations that achieve this are regarded as some of the most data-driven organizations within their industries.

Did you know?

The data mesh design pattern can be used as a tool to help create a data economy. Read more about building a data mesh.

Compounding benefits

In a data economy, teams build on and rely on each others' work to quickly create value from data, with compounding benefits for the organization. These teams can be virtual teams and made up of experts across the organization and different reporting lines.

Let's illustrate this with an example.

Your sales ops team is building a customer churn KPI dashboard for their own needs. They select tools for their own needs, ingest data, process data, figure out how to monitor the pipelines, monitor data quality, scale the pipelines, and much more. Meanwhile, your e-commerce team is building a recommendation engine for their own needs. They go through the same steps of selecting tools, ingesting data, and so on.

What's happening is that silos are being built in several dimensions — across teams, across the tools used, across formats, across how data is processed and interpreted, across how quality is monitored. This prevents collaboration and may increase the organization's technical debt.

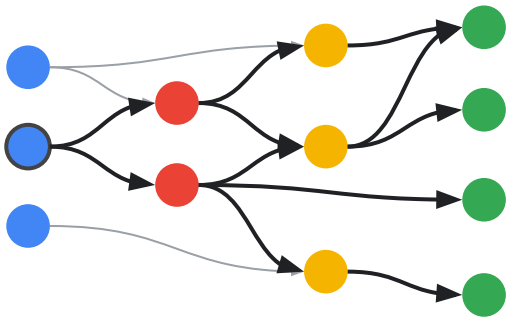
What if your sales ops team, in addition to creating their own dashboards, also published their churn KPIs for other teams to consume? The e-commerce team could perhaps make use of this data in their recommendation engine!

However, if the e-commerce team is going to rely on data from the sales ops team, it needs to be accessible and usable, meaning it can be relied on. This could mean keeping it fresh, complete, monitored, and maintained so that it stays usable and reliable for the teams relying on it.

Now you have the beginnings of a data economy: teams building on each others' work to move faster and create new value without reinventing the wheel.

Distributed responsibilities

Teams that publish data need to think about the impact that any change can have. This includes changes — intentional or unintentional — to schema, data volume, data freshness, PII status/sensitivity content, data quality, and more. It takes work to keep a data economy healthy.



A distribution of responsibilities takes place in the data economy. In the past, teams have usually relied on a central team for access to data, who hardly ever had capacity for implementing new requirements or delivered on short notice.

Having teams instead depend on each other creates new challenges around reliability. Upstream changes or failures have the potential to negatively impact many downstream data consumers – the responsibility to keep the data economy stable is distributed across many participating teams. The single point of failure of a central team has been replaced by many points of failure.

There are many parallels to software engineering and microservices architecture where learnings and best practices such as API contracts, versioning, impact analysis, fail fast, CI/CD, and site reliability engineering (SRE) can be applied to the data economy.

Decoupling the platform

The decentralization of work in the data economy requires decoupling the ownership and processing of data from the platform that provides these capabilities. Teams must eventually be able to self-serve everything they need to accomplish their goals.

Teams working with data have historically worked very closely with the platform or database – sometimes it has even been the same people forming the centralized team discussed above. You had to develop skills around optimization, capacity planning, and using proprietary features in the system — all while access to the system may have been limited.

Figure 5 illustrates a way of thinking about the differences between what happens across product teams in the business and what happens behind the scenes to make all that possible.

Traditionally these two sides have been tightly coupled. In a data economy, teams can operate independently because of the central support they receive across tooling, templates, supporting systems, governance practices, and more.

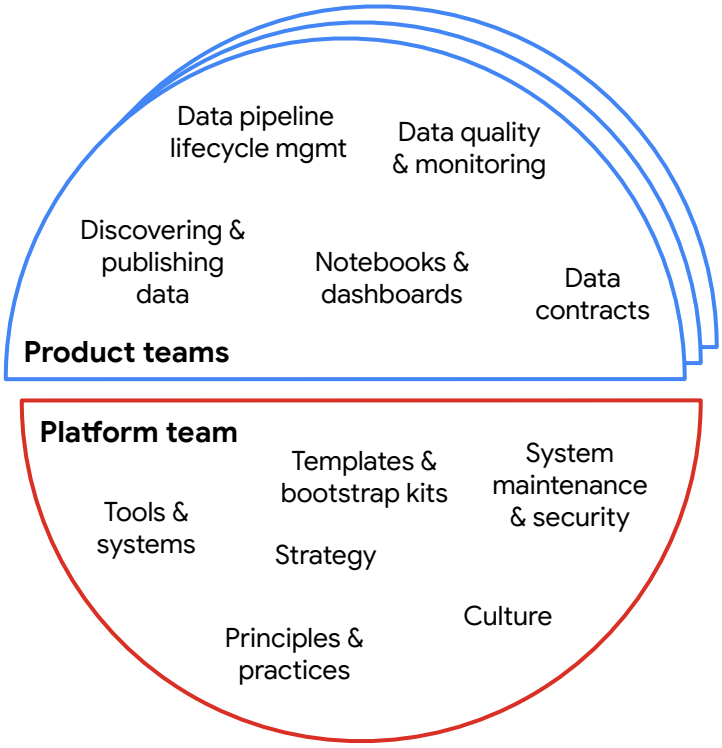


Figure 5: The value-driving work happening across the organization is supported by centrally driven initiatives focused on making work easier.

Did you know?

The traditional approach of a centralized team owning all of the data analytics for an organization, has several points of failure. For instance, this team may not have enough resources to deliver on all requests for data in a timely manner, resulting in other teams building "shadow IT" infrastructure. These "skunkworks" projects rarely incorporate good data governance, resulting in data quality issues and a lack of collaboration.

Reinventing the data platform team

Data platform teams focus on understanding the business requirements and user journeys to support the data strategy and product-oriented thinking. This allows the data platform team to create an environment to support the data product teams. They provide a portfolio of tools, reusable assets and educate users to increase efficiency and effectiveness.

The data platform team simplifies working with data for other teams by building organization-specific frameworks and templates that make it easy for teams to get their work done. They own the setup and integration of tools used for CI/CD, data governance, data access, privacy, security, orchestration, scheduling, event delivery, cost control, and making all of that consumable in a self-service way.

The mission of the data platform team (sometimes called a data ops team) is to make other teams successful on their data journey. The data platform team does not own or process data for other teams. They exist to make working with data as easy as possible for other teams.

In smaller organizations, and in the early stages of implementing your data strategy, you may find it efficient for the central team to also own some data, especially your largest data pipelines functioning as a lighthouse/blueprint implementation of a data product on the data platform. Over time, these data products can graduate into the teams managing those source systems.

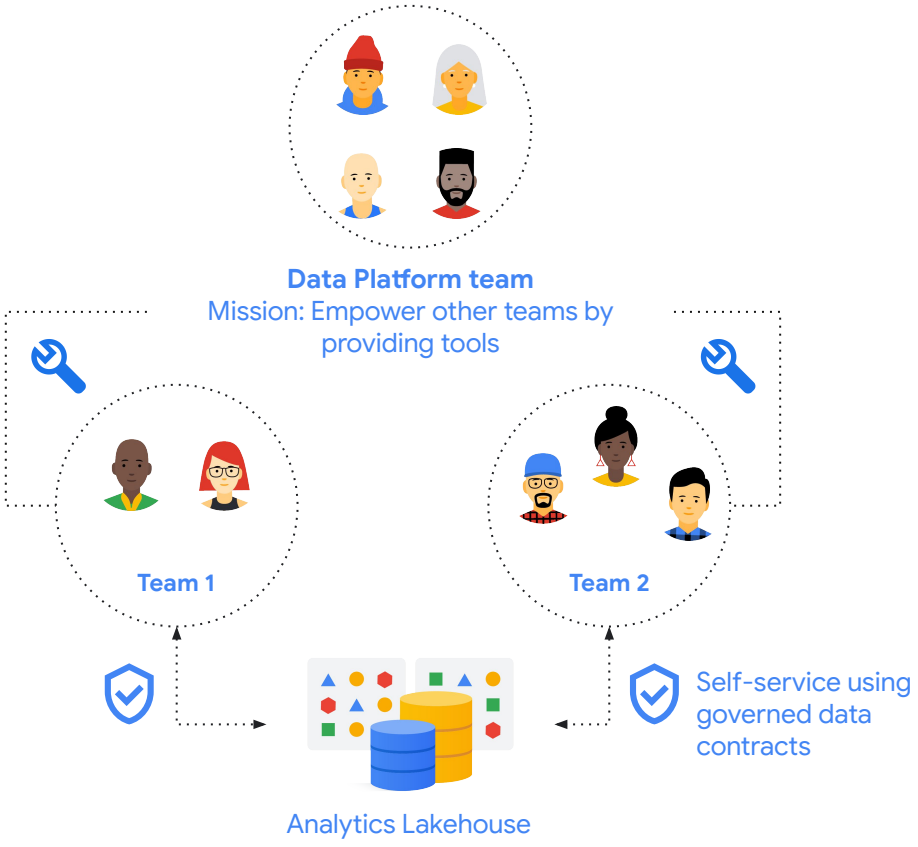


Figure 6: The data platform team supports teams across the organization to get great work done.

Data products and domains

Since data is being consumed by others, you can benefit from treating your datasets like any other product. This means having a product owner who knows your customers/users, what problems you are solving for them, deciding how to market your product, and making it reliable and useful.

When data is being discovered, consumed, and published systematically, you need to organize the ownership of it on an organizational level. This is often done by categorizing data into data domains. For example, customer, finance, human resources, and sales are commonly used domains.

Data products will look somewhat different depending on the implementation. It could be as simple as having a table tagged and documented in the data catalog, or accessible through an API. In the next chapter, we explore some key aspects of data product governance.

Note: You may wonder about the distinction between datasets and data products. The simple answer is that data products are something that other teams can rely on and build on. Datasets are collections of data that can still be shared and made available but should not be relied on or built on. Datasets normally do not provide any guarantees around SLOs such as quality, freshness, and correctness.

Distributed data governance

When ownership of data becomes distributed, clear data governance practices become necessary for the data economy and organization to function.

Having good data governance is core to solving these challenges. Pay attention to the following five areas when building your foundation: stewardship, discovery, reliability & observability, data quality, and security & privacy.

Did you know?

Dataplex provides a way to centrally manage, monitor, explore and govern your data across data lakes, data warehouses, and data marts, making this data securely accessible to a variety of analytics and data science tools.

Data stewardship

The practice of data stewardship is about ensuring your users — i.e. the consumers of your data products — are happy while upholding the policies of the organization. It is the practice of knowing what systems hold your data, knowing who has access to it, and ensuring that your data products conform to your principles, practices, policies, and standards.

Data stewardship is a broad field of practice. It can include tasks such as classifying your data, having accurate metadata for your data products, ensuring PII is tagged, ensuring schemas are available, and making sure that quality is checked regularly.

It can also span broader tasks such as understanding the root source of the data, tracking the upstream and downstream lineage of the data, tracking KPIs for the data product, identifying the costs that your data products are generating, and keeping track of other data products that may be duplicating your work.

Data stewardship practices all rely on having clearly defined and documented authority over data domains. The data stewards/trustees/delegates can make authoritative decisions about the data they manage.

Did you know?

Easier risk management is one of the benefits of data governance. Read more

Discovery and cataloging

A data economy is made possible when data can be discovered and published by anyone in the organization. It is good practice to tag the source of truth of certain data so that everyone knows which data products to rely on for their work. Sometimes these are known as "golden data products."

Your users should be able to search for datasets and data products and identify their purpose and schema without necessarily having access to them (see principle of least privilege). This can be facilitated by auto-registering all datasets in the catalog — even those not offered as data products. If a particular dataset is frequently requested, that is a good signal that it's providing some value to users and should perhaps be turned into a golden data product.

A data exchange lets teams discover and share data in an automated way without human interaction (though team communication is still important and something a great data exchange can facilitate).

We include a small example to illustrate this:

One of your teams collects raw data from a range of devices. Other teams could use this data, but it hasn't been combined into a unified schema and it hasn't been checked for quality. Instead of publishing this raw dataset directly, your team combines and processes this data into a data product that other teams can easily understand and rely on. You tag this as a golden data product, meaning that it is the authoritative source of the data it describes.

Did you know?

A data marketplace or exchange, for example Analytics Hub, can let you share, consume, and monetize data across or outside your organization. Read more

Reliability and observability

When teams depend on each other for business-critical data, the reliability of those data products matters like it does for any other product used to run the business. In a data economy, making a breaking change to a schema, or not making data available on time, or missing some data becomes a reliability concern that must be managed systematically.

You can use your data CI/CD pipelines to prevent making schema changes that could break downstream consumers. You can monitor SLIs and track their performance toward set SLO targets.

You can create alerts for abnormal conditions that would cause significant problems. For example, you might set an SLO target that your data is no older than one hour 99% of the time, or you may alert by comparing against past calculations with some margin of error.

To track the performance of your data products, your pipeline and quality metrics need to be brought together in a unified monitoring and logging system. This will allow you to define alerting policies and do troubleshooting in one place, creating a solution that scales to all teams.

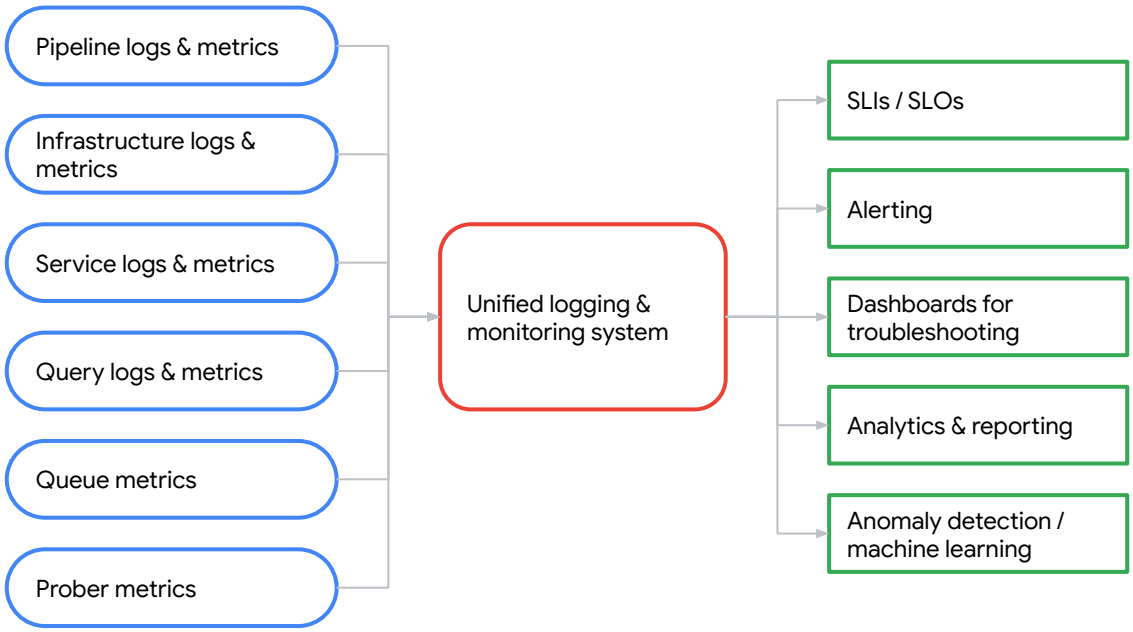


Figure 7: Reliability engineering is underpinned by having great monitoring and observability capabilities.

Data quality

The value of data increases with its quality, and data may not be useful at all if it doesn't meet a minimum bar for quality. Data offered in a data exchange must satisfy defined quality guidelines. Many organizations have therefore implemented certifications for their data pipelines and products. Data quality checks can be strict and stop pipelines, or more relaxed and only report the quality metrics into metadata and monitoring.

How can you report on financial data that is missing an hour of transactions or has duplicates? Even worse, if someone is depending on your data, a duplicate value could break their application without them even finding out before it is too late. Streaming systems are particularly vulnerable to this as errors propagate downstream quickly, sometimes without the ability to reprocess the data correctly because it has already been consumed.

On the other hand, some applications tolerate data quality issues better. For example, if you are processing sensor readings and miss a few values, it may not pose a massive problem.

The central platform team can play an important role in helping the organization develop its data quality muscle. A low-hanging fruit is to develop data quality templates for teams to utilize, and incorporating automated data quality checks into the data exchange. The platform team can also help teams set up monitoring and alerting, helping them be a reliable source of data in the growing data economy.

Did you know?

"Midas Certified" data represents the gold standard for data quality at Airbnb. Read more here.

Security and privacy

Modern data platforms have become quite good at maintaining a baseline of data security. For example, encryption at rest and in transit is possible with most platforms. The platform team can simplify a lot of security controls for the many teams they serve across the organization.

Security is a broad field that requires careful and dedicated attention. Beyond the baseline on most platforms there are many questions that may narrow down your options:

- Do you need the encryption keys to be stored in your own key manager on your premises?
- Do you need confidential computing to keep data encrypted within your VMs?
- Do you need compliance with strict government regulations?
- Do you need comprehensive audit logging across your entire data lifecycle?

Depending on which tools you choose, there is the possibility to simplify many of these controls for the platform team. Enabling audit logs across the tools can also help you gain traceability in case of an incident.

Did you know?

Data Loss Prevention (DLP) in Google Cloud can be used to scan the organization to find all sensitive data and tag it in the data catalog. DLP can also be used to determine different sensitivity levels for applying data masking.

Data Ecosystem

Taking a holistic perspective on data strategy ensures that all critical components are covered. Deliver value end-to-end from source data to insights and apps by choosing a unified, open and intelligent platform.

Even if you have a pressing need for a data warehouse, machine learning platform, or data processing tool, you must consider the data ecosystem beyond your immediate needs to ensure you are setting your organization up for fast adoption without friction.

We present a minimum set of capabilities that should be present in the data ecosystem you choose. Across these capabilities, you must also plan to manage several horizontal concerns. See *figure 13* at the end of the chapter for a visualization of this.

In this chapter, we give you a vocabulary to start thinking about your data ecosystem:

- **Golden paths** – Simplify the data user journeys in your organization
- **Open ecosystem** – Choose an ecosystem that covers all vital capabilities, but also lets you plug in and replace components using open standards
- **Unified platform** – Data is best managed and used when stored in a unified consolidated platform as part of an ecosystem
- **The shift to serverless** – Use tools that take away as much operational management burden as possible
- **Data integration and exchange** – Bring your enterprise data into the analytics lakehouse, so that it can be used by the entire organization in the data economy
- **Time for real-time** – Leverage real-time data to build faster and more reactive systems
- **Enabling AI across the ecosystem** – Easily incorporate AI & ML whenever possible to improve your competitive advantage and make it usable by all users in your organization
- **Sustainability impact** – Consider how your data strategy supports your sustainability strategy and initiatives
- **Data ecosystem baseline** – Detailed description of the data ecosystem capabilities and horizontal concerns

Golden paths

Create golden paths for different personas and user journeys in your organization. Playbooks, blueprints, and automation are some examples of ways that golden paths can be created. A good objective is to have a large majority of your developers using self-service golden path tools supporting common user journeys when developing their solutions.

Golden paths provide recommended and well-supported tools and methods for performing common data tasks, across collection, processing, presentation, governance, monitoring, and more.

They radically simplify tasks for people in your organization, such as:

- A data engineer setting up a new data pipeline
- An application engineer ingesting streaming data from an app
- An ML engineer building a new ML model
- A developer making a dataset available to others
- A data scientist testing a hypothesis on collected data
- A business analyst creating a dashboard for business stakeholders

The first step to creating a golden path is writing guiding documents or code templates that teams can pick up and start using. Once a golden path gains traction, it can be upgraded to leverage automation. This could come in the form of a web-based tool that automates the steps.

Doing periodic engineering reviews of new needs is necessary to continue evolving and adjusting the golden path.

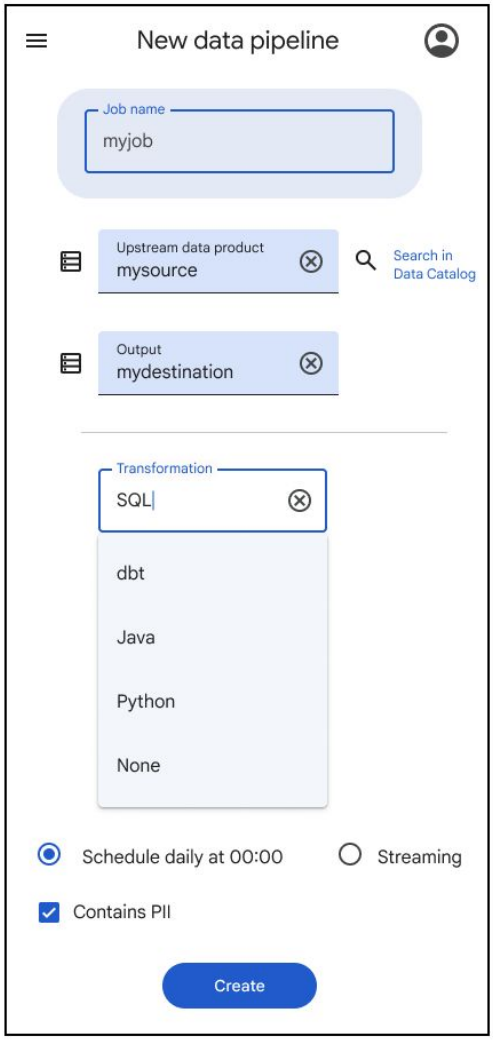


Figure 8: Example workflow for creating a new data pipeline using golden path tools.

Did you know?

The integrated and no-ops turnkey nature of Google BigQuery makes it possible to build great user experiences for any persona.

The unified control plane

The entire ecosystem of capabilities can be overwhelming for users to manage. A unified control plane can provide abstractions for managing and securing access to tools and data.

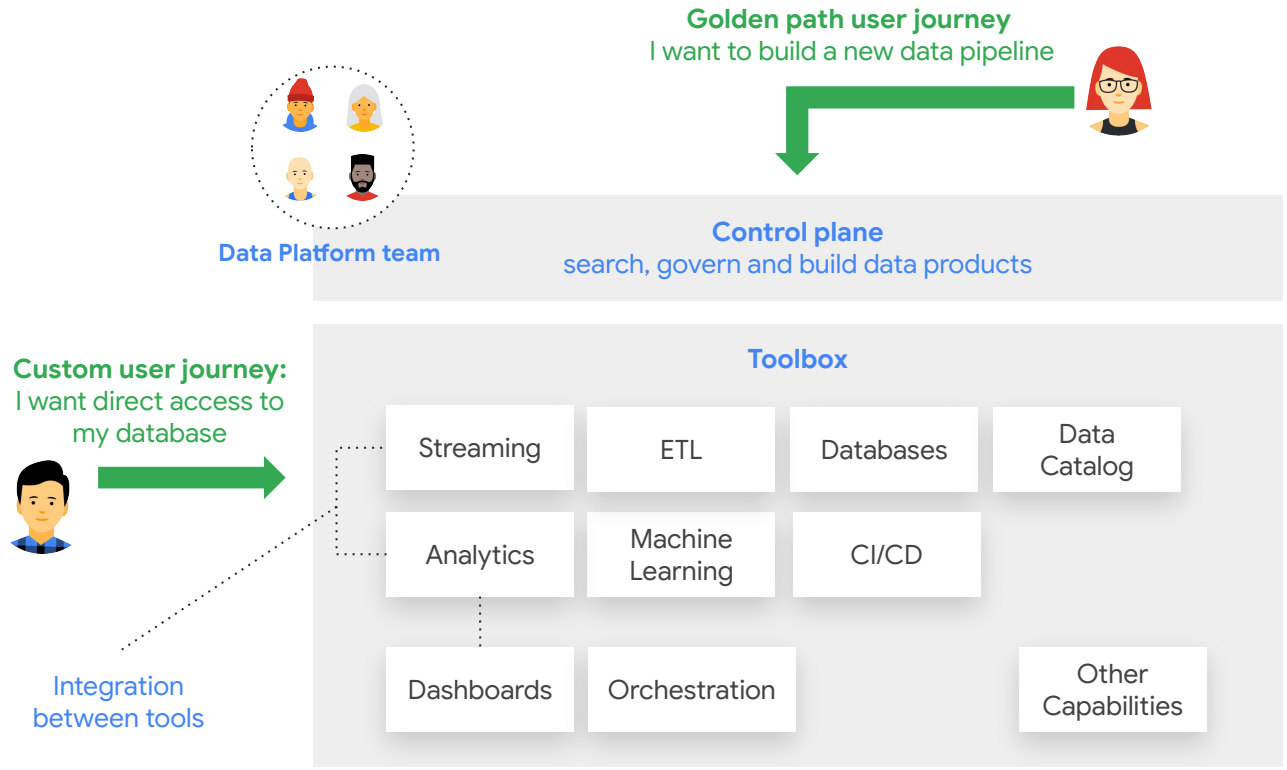


Figure 9: Adding a unified control plane can simplify work for teams.

Having a unified control plane toward your data infrastructure simplifies the building of golden paths while also allowing advanced users to drop down to the lower layers for customization, adaptation, and evolution of the golden path.

The unified control plane provides users with simple, governed, and secure access to the core capabilities of the data ecosystem. You can use a combination of open-source solutions like Backstage, Terraform, and Crossplane to help your central platform team build the capabilities to get this work done easily.

Open ecosystem

The data ecosystem you choose should be modular but also feature-complete. At first, these objectives may sound contradictory. How can you choose an ecosystem that provides you with everything out of the box but is also modular, letting you replace components where necessary? This is possible by choosing tools that use open standards.

It is common for organizations to build several data platforms, sometimes even across clouds. While this approach may meet certain requirements such as managing data everywhere or having cross-cloud solutions, it is a more costly, slower, and complex compromise compared to a consolidated approach.

Did you know?

McKinsey advocates using an architecture that supports modularity, which helps you serve the different needs of your organization as not all use cases can be solved with a specific technology.

Avoiding lock-in with open standards

The notion of lock-in is sometimes misunderstood. Any time you are making a choice, you are locking yourself in, in some way. A good middle ground is to build on open standards when possible. Within data tooling, we need to differentiate between the interface and the tool itself. A set of common interfaces and protocols are emerging and each tool supporting the interface will do it in its own way.

Open standards give you a degree of freedom to move between implementations — most importantly, freedom to move upward to more managed or scalable implementations.

Did you know?

Within databases, we are seeing PostgreSQL emerging as a common interface across a range of products, from plain PostgreSQL databases to massively scalable systems like Spanner. Google Cloud makes this simple through compatibility with and contribution to open-source ecosystems.

Once you've chosen the interface to use, you can choose the implementation that gives you the most competitive advantage. For example, by choosing a fully managed cloud service, your teams can focus on the business problem, innovate faster, and keep costs under control easier. When you apply this thinking across all the components in the data ecosystem, it can become a strong competitive advantage.

It is important to not confuse open standards with multi-cloud tooling. You must evaluate the compatibility of your data platform with open standards and make sure paths are available for teams who want to try alternative solutions.

Did you know?

Kubernetes and dbt are technologies that can help you abstract the peculiarities of each cloud.

Unified platform

Your goal should be to build a consolidated data platform where all the necessary capabilities are available in a unified way. This usually includes moving data from all your source systems into one platform where you can build the best data experiences and a robust data economy.

You cannot avoid transferring data between systems. You either move it once or you move it every time you access it. If you are using Google Cloud as your data platform but running a source system on-premises or on another cloud platform, this transfer will happen across the internet or a private connection.

Network-related costs are often viewed as a blocker for these setups. However, when running the numbers, we usually find that the network costs are small compared to the costs of alternate solutions and the upside opportunity of having a consolidated data platform.

Resilience

If you are storing data in one cloud region, you should have a business continuity plan including replication of critical data to other regions where you could continue operations in the case of a disaster in the region you are operating.

Managed and serverless systems make it easier to get back up to speed from backups, as the cloud capacity is often more flexible. Regular failover exercises are recommended.

Read more about the multiple disaster recovery scenarios for data.

Multi-cloud data vs applications

Multi-cloud data is the notion that data can be managed in different locations. In practice, data is very different from applications. Different datasets often need to be combined, and repeatedly moving data between locations is not efficient. We recommend choosing a consolidated data platform with the capability to analyze data in other locations when needed.

Google Cloud is a strong advocate of the multi-cloud advantage within computing and best-of-breed flexibility. These differ from multi-cloud data.

Multi-cloud computing is the notion that you may need to serve clients in countries where only a particular vendor has a cloud region or that you want portability across clouds in case you need to change providers.

Multi-cloud best-of-breed refers to the flexibility of being able to choose the best provider for different tasks. For example, you may run your website on one provider but store and process your analytical data using another provider.

Read more about how to do multi-cloud right.

The shift to serverless

Serverless systems let your teams focus on getting work done with data while requiring fewer people to look after the systems. They often also let you index your costs to your consumption, ensuring you spend your budget where it makes a difference. Serverless is quickly becoming the gold standard for cloud-native organizations, perhaps for the simple reason that it makes it easy to get work done with data across the organization.

Serverless computing refers to systems that require no operational work, meaning systems that require no patching, no upgrades, no maintenance, and no service windows.

Serverless data systems tend to separate the storage layer from the computation layer. This opens up for a wide range of users to use the stored data simultaneously, e.g. a reporting tool querying the data with a SQL engine, a data pipeline reading the same data with a Spark-based engine, and a machine learning model trained on that data using an AutoML engine — all of these can also run in a serverless fashion.

Using serverless systems across the organization can reduce or remove the need for specialized teams tasked with keeping the infrastructure healthy.

Did you know?

Some serverless systems are also instanceless. BigQuery is an example of such a system. This means that you don't need to create an instance/cluster/warehouse before you can start using it. Serverless Spark is another example.

Intelligent maintenance and optimization

Modern platforms are increasingly using machine learning to provide users with proactive recommendations and remove operational decisions. This reduces the management cost considerably and active recommendations can help you rightsize your resource consumption to further reduce infrastructure costs.

In your selection of tools, consider platforms and products that perform automatic optimization and require little-to-no maintenance, ideally at no additional cost. Traditionally, data systems have required careful tuning and expertise around execution plans, reorganizing storage, optimizing indexes, data compression, scaling decisions, data vacuuming, etc. This no longer has to be the case in a cloud-based data ecosystem.

Cost distribution and FinOps

Cloud-based systems, and in particular serverless systems, create mechanisms where each data consumer can carry their own costs and capacity completely separate from the owner of the data. This is a game-changer when it comes to scaling a tool across an organization.

Building on a unified and integrated ecosystem brings the ability to apply FinOps practices for cost control and cost transparency. This will allow you to not only understand where your cost centers are, but more importantly, it will enable you to put cost transparency in the hands of all teams. When teams can self-serve their operating expenses, they can make informed decisions about how to optimize that spend.

Culture is key here, and developing a blameless culture around cloud spend is important when it comes to encouraging experimentation and learning.

Did you know?

There are three stages in the basic FinOps lifecycle: inform, optimize, and operate. Read more.

Data integration and exchange

Your data strategy should incorporate data from all systems and make it available for consumption within the data economy. However, your enterprise systems like ERP and CRM are not participating in the data economy — the data originates from within each system.

To unlock this data, each source system's owner should ultimately be responsible for ensuring that their data is replicated into the common platform and made available as data products. They would also be accountable for the reliability of that replication.

Share a data product, not your database

Applications store data in operational databases. These databases are not normally shared with other teams. The way you make this data available to others is by bringing it into the analytical realm and making it available as a dataset or data product. Choosing the right ecosystem can significantly simplify the integration and replication between these systems.

Bringing transactional data into the analytical system needs to be effortless. This type of replication and sharing is easily done with the tools of a modern data ecosystem. Replication, ETL, ELT, and real-time messaging components can be used to bring data into the analytical systems. Increasingly, modern analytical tools also provide automatic replication from select operational sources.

Ultimately this will allow more data to be processed, shared, and analyzed by teams across the organization.

Did you know?

Datastream on Google Cloud enables easy replication of data from OLTP systems to BigQuery among other targets.

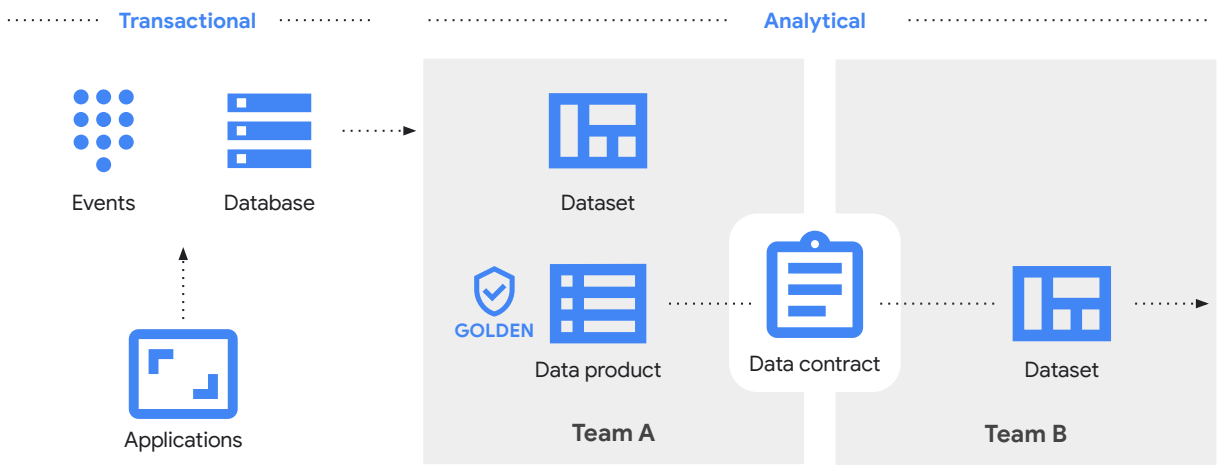


Figure 10: Convergence of Analytical and Transactional

Rise of the analytics lakehouse

Modern cloud data systems combine the characteristics of structured data warehouses with the flexible and open nature of data lakes in a class of systems called analytics lakehouses.

Applications rely on many different types of databases — relational, NoSQL, NewSQL, graph, document, key-value, wide-column, in-memory, etc. The choice of database architecture can and should depend on the requirements of each use case. However, for analytical purposes, all this data needs to come together.

Data lakes have provided limitless compute and storage to process petabytes of data stored in various file formats. This has made them best suited for data science and machine learning needs.

Data warehouses have provided a governed, integrated and structured view of enterprise data, which has made them best suited for business intelligence and SQL-based exploration and analysis.

The analytics lakehouse is an evolved architecture that combines the best of both worlds, supporting both business intelligence and AI/ML use cases alike, while breaking data silos. It also enables data engineers, data scientists, and other data users to share data, collaborate more effectively, and deliver more business value faster, with unified governance across all data.

This enables the analytics lakehouse approach to scale to an entire organization.

Closing the data loop

Bringing your analytical results back into the application realm is an important part of serving insights back to customers.

To take advantage of the insights you create, data may need to be fed to various systems. It could be to an API endpoint or into an operational database. The data ecosystem you choose should have capabilities to push analytical data back to these systems. This process is sometimes known as reverse ETL.

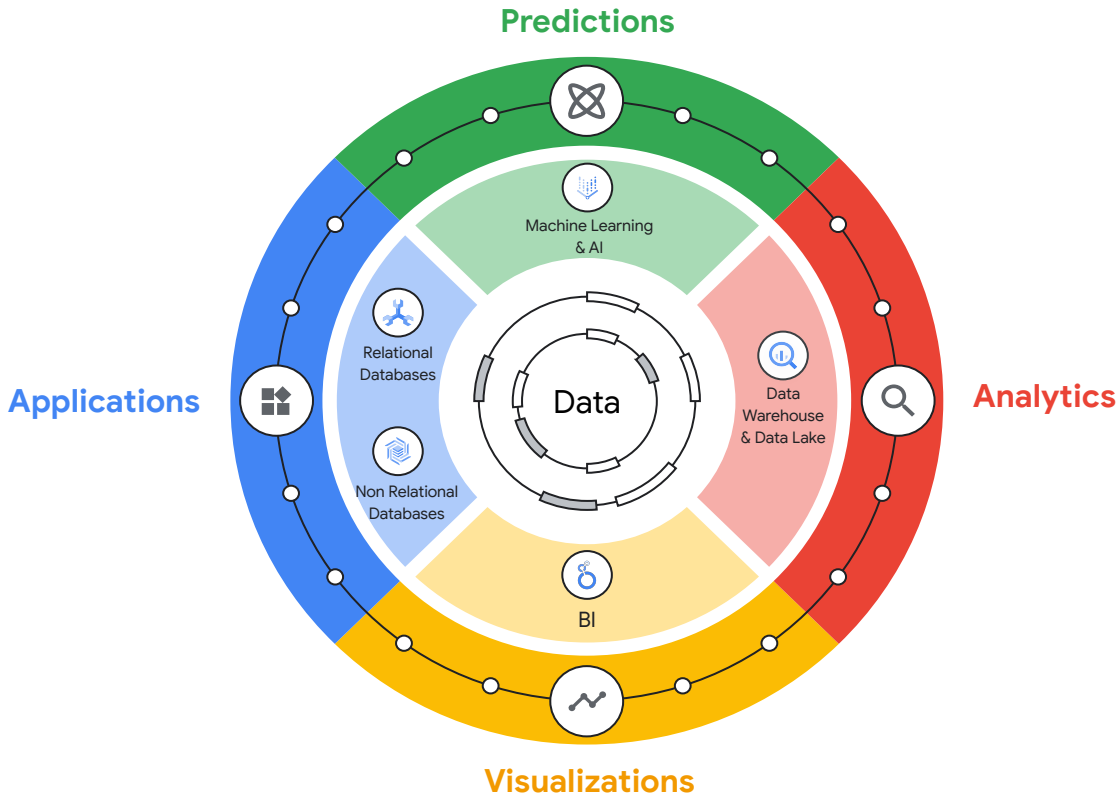


Figure 11: A unified, open and intelligent data ecosystem

Time for real-time

Near real-time decision-making requires access to the latest information. We refer to this as the "peak of now". The peak has steep drop-off to where the data is no longer relevant to take immediate action on. By connecting streaming data to the lakehouse analytics capabilities, organizations can quickly analyze and activate data-derived insights as they happen instead of waiting for a batch process to complete.

The faster an insight can be derived from data and action taken, the higher the value of that insight. Conversely, outdated insights can quickly become worthless. However, over time as events are collected, a "mountain of knowledge" is built which opens up for new value to be created through analytics or machine learning.

Real-time streaming analytics use cases include fraud detection, inventory and fleet management, dynamic recommendations, predictive maintenance, and capacity planning — just to name a few.

Streaming architecture

Architectures like Lambda and Kappa have now been around for a while to help organizations support both batch and streaming data processing. However, these architectures have some drawbacks. Back in 2015, Google published the Dataflow model paper which is a unified programming model for both batch and streaming. Apache Beam is the open-source implementation of this model.

With Lambda, the batch and streaming processes may require different code bases. With Kappa everything is considered a stream of data, even large files which have to be run through the stream processing system, sometimes impacting performance.

Apache Beam provides a smooth transition from batch to streaming by allowing developers to reuse the same code for their data pipelines and switch from a bounded input/output (batch) to an unbounded input/output. This allows for using the same programming paradigm and tooling regardless of whether data is coming in via a stream or a batch file transfer.

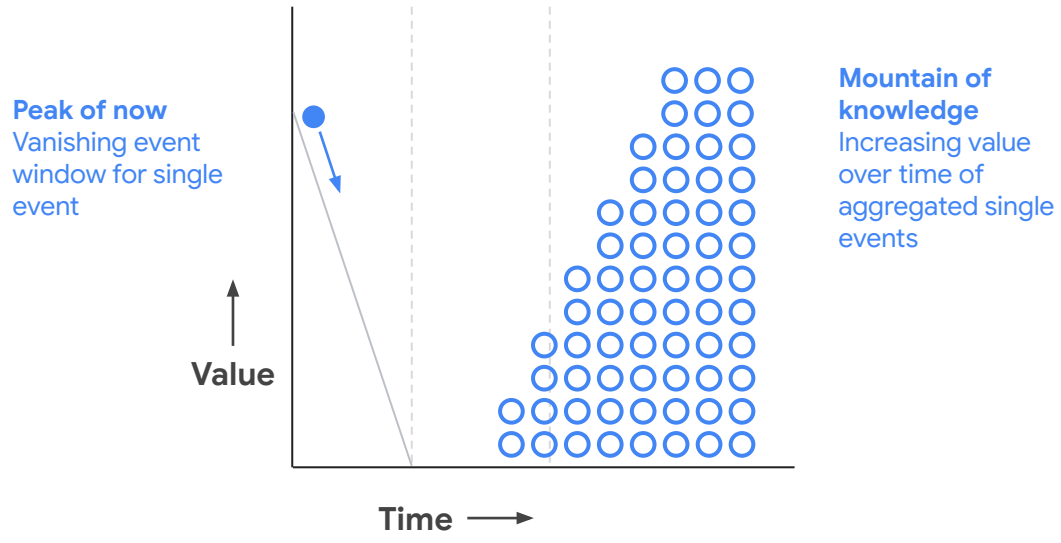


Figure 12: The value of a captured data point varies over time.

Enabling AI across the ecosystem

A modern data ecosystem must include native AI/ML capabilities that remove the need for data transfer between the data storage and the AI/ML tools. This is an important factor for translating data products into AI solutions faster.

Adding AI/ML capabilities to business processes is one of the ultimate goals when building your data ecosystem. In many cases, AI/ML provides additional information to humans for decision-making. However, there are more and more business processes where AI/ML models are embedded to drive automated decision-making.

Removing barriers

Choose an ecosystem that removes as many barriers as possible for getting value from machine learning. The right integrated ecosystem has the potential to significantly increase your rate of experimentation with AI/ML so that teams can test and create new solutions.

Machine learning is truly a journey of trying things, failing, and learning. Make the tasks in this cycle as fast, easy, and cheap as possible — data discovery, data movement, model training, model deployment, and model monitoring.

AI/ML is also becoming available to an increasing number of users who don't possess the typical technical skills. AutoML capabilities allow practically everyone to be able to create high-quality models. This lets analysts incorporate machine learning without having to learn new programming frameworks, nor having to think about how to scale the solutions.

The citizen data scientist

When everyone can create high-quality models easily and become citizen data scientists, a new range of challenges need to be addressed, like avoiding biases in models and explainability of predictions. It is strongly encouraged for everyone to obtain a high-level understanding of AI (see Data University) to better grasp the implications of using AI/ML in business processes.

It is a best practice to include AI/ML principles and guidelines in your data strategy and to create checkpoints early in the design process of AI/ML-driven solutions.

The last mile of ML

Consider ecosystems that allow you to easily deploy models for batch, streaming, or transactional processing, seamlessly integrating with the data required to train and serve the models. The easier it is to get the model serving in production, the easier it becomes to integrate it with the business process.

It is common for AI/ML projects to never make it into production, even if they have produced a high-quality model. One common cause is that the business process that was supposed to be improved did not have the technical capability to integrate with the model, the model didn't incorporate all the business requirements, or the business process needed to be changed as such. It is important to address and define these requirements upfront.

Did you know?

BigQuery allows you to create, train and serve high-quality ML models right from the analytics lakehouse using nothing but SQL. Read more here.

Operational ML

Machine learning models need close monitoring as they usually require periodic re-training. Changes in real-world conditions may require a model to be retrained or adjusted if its performance degrades outside of defined SLO boundaries. With the growing number of AI artifacts created and used in business processes, the data ecosystem must offer robust automation capabilities to keep these models performant.

ML Ops is a concept that was coined based on a 2015 paper by Google researchers called "Hidden Technical Debt in Machine Learning Systems". At that point, we had learned a lot about what it takes to use machine learning in production. Today ML Ops is considered a core practice for engineers working with machine learning systems.

Did you know?

Vertex AI provides AI/ML capabilities for all types of users, allowing you to easily manage and deploy models for batch and real-time inference. Vertex AI is closely integrated with BigQuery, enabling SQL-based ML development. Vertex AI also integrates with transactional databases like Spanner and AlloyDB for real-time predictions in your transactional workloads.

Sustainability impact

Your data strategy should support your organization's sustainability strategy and initiatives.

Here are some examples:

- Choose to work with infrastructure providers that are driving the strongest sustainability programs
- Take advantage of the capabilities in serverless systems to optimize your workloads
- Use data to develop better optimization strategies for efficient operations and reduce carbon emissions
- Measure and report your cloud carbon emissions to encourage awareness around the choices data engineers make

Did you know?

Lufthansa Group uses data to reduce carbon emissions of airline travel. Read more.

Data ecosystem baseline

The data ecosystem is the sum of the data platform capabilities you use, plus the horizontal concerns you manage across those capabilities and the users you are serving.

We present a baseline of capabilities that your chosen data ecosystem should support on day 1, such as data quality measurement, data cataloging, and ML tooling. The ecosystem must also support horizontal concerns, such as managing data security and governance across all capabilities. *Figure 13* illustrates this.

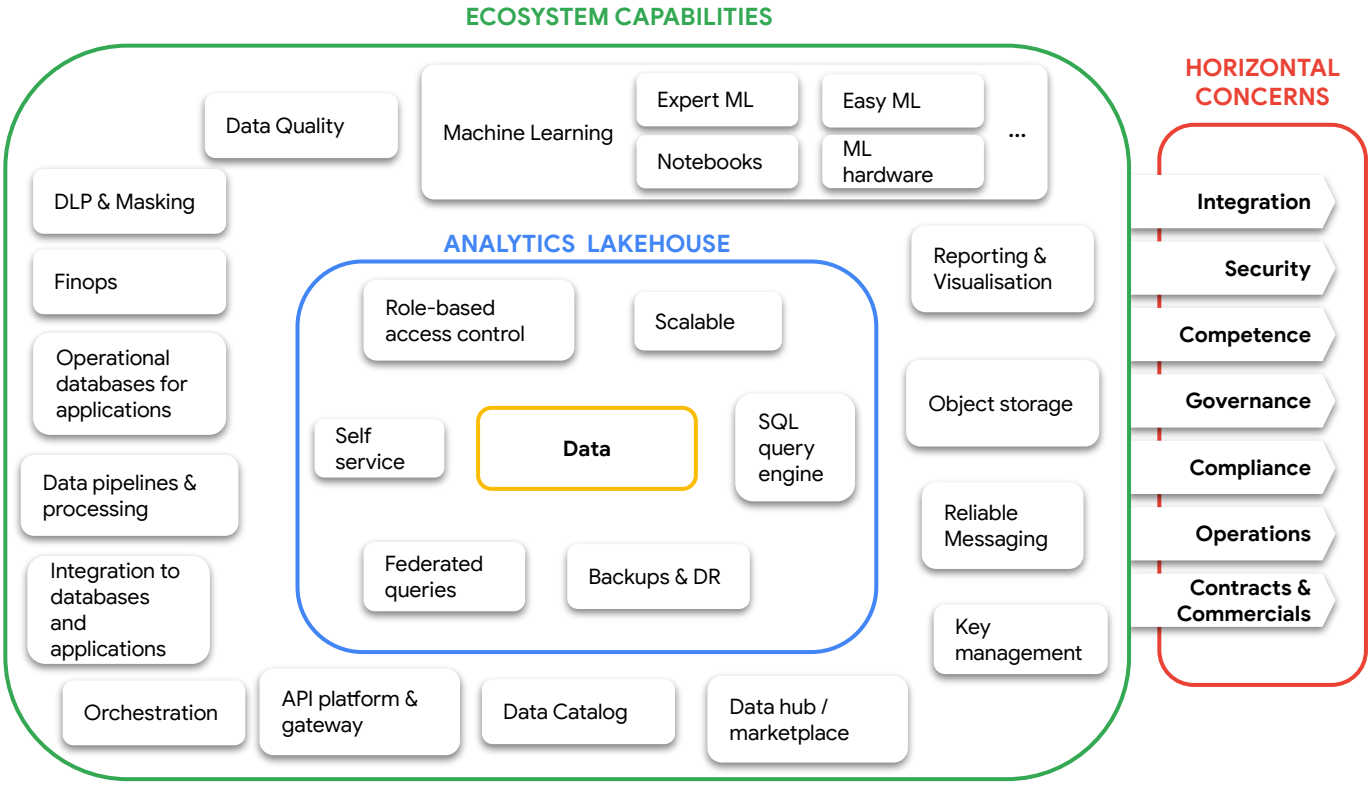


Figure 13: The Data Ecosystem

Capabilities

Data quality	Including both measuring quality and acting on the results.
Masking and data loss prevention (DLP) tools	Helping prevent privacy incidents and ensure sensitive data is not exposed, internally or externally.
FinOps	The ability to analyze, predict and pinpoint costs across the value chain.
Operational databases	Data in your analytical systems will be going to and from operational databases to serve applications.
Data pipelines & processing	Simple and powerful tools to process and ingest data using SQL or a programming language.
Integrations	To bring external data into your analytics lakehouse you'll need integrations to your enterprise systems and databases (ETL and ELT capabilities).
Orchestration	Ensure that your jobs and pipelines execute as intended.
API platform	Expose API-based functionality to other internal and external systems.
Data catalog	Facilitate metadata management, data discovery, and lineage.
Data hub/marketplace	Making data available to others for discovery and consumption, both for internal and external users.
Key management	Manage encryption keys for your data end-to-end.
Reliable messaging	Decouple systems from each other, e.g. separating data capture from data processing, which is also necessary to support real-time events.
Object storage	Storing unstructured data and objects (including files with structured data) is a core part of a analytics lakehouse, so that the data can be made actionable.
Reporting & visualization	Ability to create reports, dashboards, and draw insights from data.
Machine learning	Empower every person in the organization to use machine learning, whether it's directly with SQL in the lakehouse, expert users building custom models, or developers that want to add intelligence to their applications.

Horizontal concerns

Integration	Each capability will potentially need to interact with all other capabilities. The more integrated the capabilities are out of the box, the faster and cheaper you get results. For example, if the machine learning platform can read directly from the analytics lakehouse without any additional setup, integration, data copying, or extraction, adoption of machine learning will be smoother.
Security	Managing cross-cutting security across all these capabilities, including identities, permissions, and audit logging without extra integration or setup. This can also include data exfiltration protection measures.
Competence	The more you can apply skills across capabilities, the faster and easier it will be for teams to learn the skills needed to be effective with the tools.
Governance	Managing standards and policies across the capabilities and data, end-to-end.
Compliance	Ensure that the entire ecosystem is compliant with the wide range of compliance that you need to manage. Each capability may need to conform to specific regulatory requirements around e.g. encryption, data residency, or access control.
Operations	Operations and lifecycle management for each capability. Prefer serverless or managed services here to reduce operational burden.
Contracts & commercials	Procurement and commercial contracts for each vendor.

Getting started

Below you will find a few recommendations from each section. We encourage you to jump around in the paper to the sections that interest you the most.

Experiences

Productive user experiences enabling all users to access and create value from relevant data

- Establish a data university to advance data literacy for everyone in your organization
- Pivot your organization from role-oriented to product-oriented with cross-functional teams
- Define data principles for your organization that align with priorities and provide clarity in decision making

Economy

Principles and practices to ensure that data can be published, discovered, built on, and relied on

- Staff a data platform team that is obsessed with the developer and analyst experience
- Make sure you get external/enterprise source data into the data economy early, such as customers, transactions, and product data
- Find teams that depend on each other for data and help them share their data as a data product
- Implement basic data governance practices — stewardship, discovery, quality, reliability, and privacy

Ecosystem

A unified, open and intelligent platform with end-to-end data capabilities for all users and needs

- Choose a data ecosystem that provides your organization with all the capabilities you need out of the box, with open standards to plug in other components where needed
- Make a plan to enable everyone in the organization to apply AI in their work using the capabilities in the ecosystem
- Keep your operational work to a minimum by choosing serverless tools for the capabilities in your data ecosystem

Measure progress using KPIs that track the number of learners, the number of data products, and the amount of data queried – to name a few. Make sure the data matches the stories from the business — anecdotes rarely lie!

Finally, make sure to have fun along the way!

Three pillars for building a modern data strategy

February 2023

Interested in getting started? Contact us to learn more.

Google Cloud