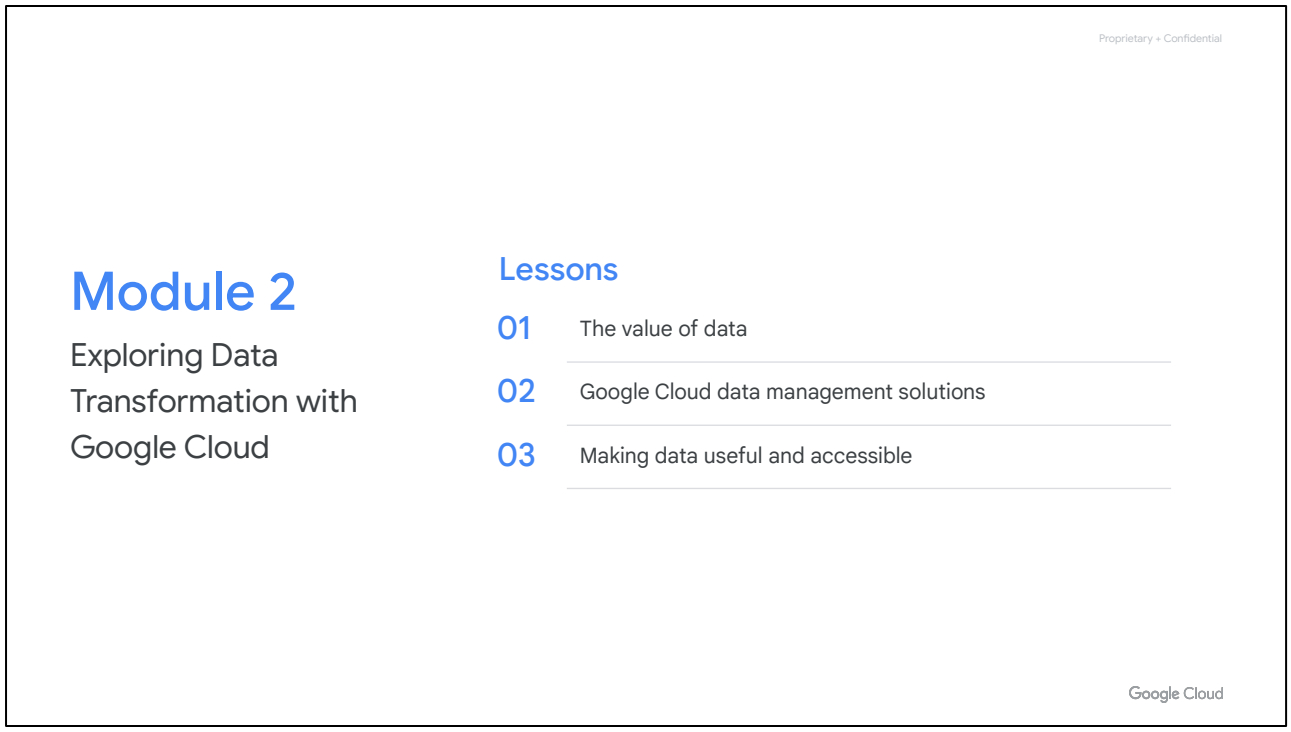


It's now time for module 2, "Exploring Data Transformation with Google Cloud." In this section of the course, we explore how data can now be consumed, analyzed, and used at speed and scale never before possible. We also explore how organizations can benefit from cloud technology to ingest data in real time to train machine learning models and to act in ways that benefit their business. The great news is that you no longer need to be a data scientist or technical expert to perform data analysis.

In this section you learn about:

- The value of data
- Available Google Cloud data management solutions
- How to make data useful and accessible

Module 2

Exploring Data Transformation with Google Cloud

Lessons

The word “data” is used a lot in today’s business world. There’s a good reason for that, because capturing, managing, and using data is central to redefining customer experiences and **creating new value** in almost every industry.

Data is the key to unlocking value from AI

- Powers AI-driven business insights.
- Helps companies make better real-time decisions.
- Is the basis for how companies build and run applications.

And data is an essential ingredient for driving innovation and differentiation, and is the key to unlocking value from **artificial intelligence**.

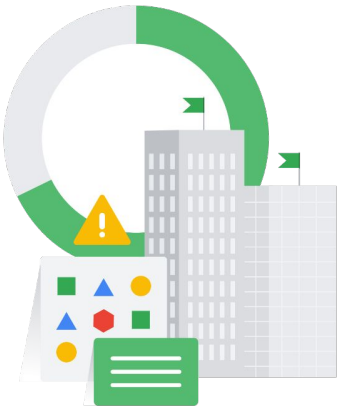
- Data powers AI-driven business insights,
- Helps companies make better real-time decisions, and
- Is the basis for how companies build and run their applications.

We're generating more data every day, and the complexity and speed of data arrival are changing the business environment.

Working with data can still be challenging

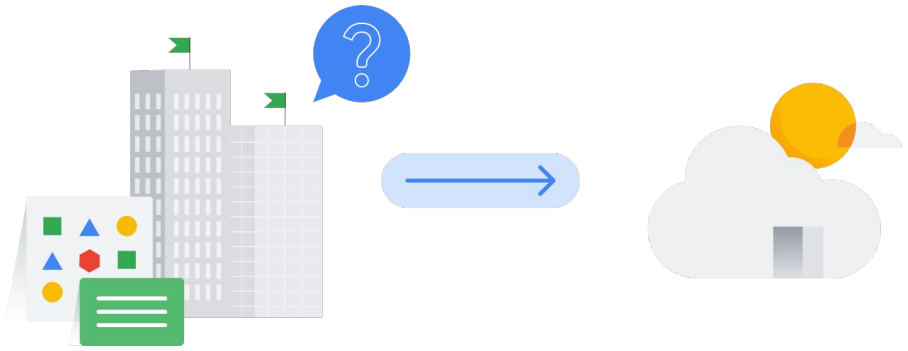
> 68% of organizations are unable to realize tangible and measurable value from data.

Closing the data value gap
Accenture



However, some organizations struggle to remove the barriers that sit between them and their data. According to a report by Accenture titled "Closing the data value gap," 68% of organizations say they are still unable to realize tangible and measurable value from data.

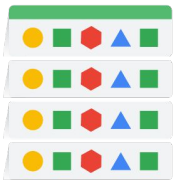
An intelligent data cloud is the key to unlocking more business value



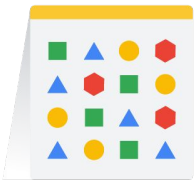
Organizations that want to adapt must determine how to close the gap and support value generation. **An intelligent data cloud** is the key to unlocking more business value.

Unlocking business value from data

Three main types of data



Structured data



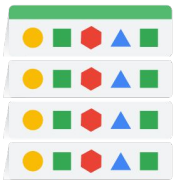
Semi-structured data

Unstructured data

Unlocking the value of data is central to digital transformation. To generate insights, you might need to combine different types of data. However, not all data is created and organized the same way.

Data can be categorized into three main types: **structured**, **semi-structured**, and **unstructured**.

Structured data



Structured data

- Is highly organized and well-defined.
- Is typically stored in a table.
- Includes spreadsheets and databases.
- Is easy to analyze.

Structured data is highly organized and well-defined. It's typically stored in a table with relationships between the different rows and columns, like in a spreadsheet or database.

Because structured data is organized in this way, it is easy to analyze. For example, it's common for organizations to use structured data in customer relationship management tools, or CRMs, as they follow customer behavior patterns and trends.

Unstructured data

Unstructured data

- Doesn't have a predefined data model.
- Isn't organized in a predefined manner.

Categories

- Text, like documents and presentations
- Data files, like images, audio, and video
- Infrastructure activity and performance data

The opposite of structured data is **unstructured data**. Unstructured data is information that either doesn't have a predefined data model or isn't organized in a predefined manner.

Categories include:

- Text, which is the most common, and is often generated and collected from sources like documents, presentations, or even social media posts.
- Data files, like images, audio files, and videos.
- And infrastructure activity and performance data, like log files from servers, networks, and applications or output data from Internet of Things (IoT) sensors.

Semi-structured data

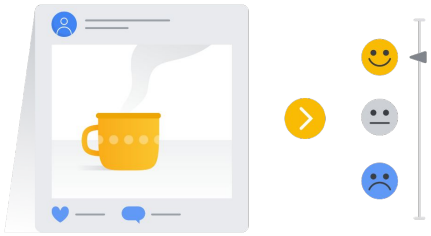
Semi-structured data

- Is organized into a hierarchy.
- Lacks full differentiation or order.
- Includes examples like emails, HTML, JSON, XML.
- Doesn't have a formal structure.
- Contains tags for easier analysis.

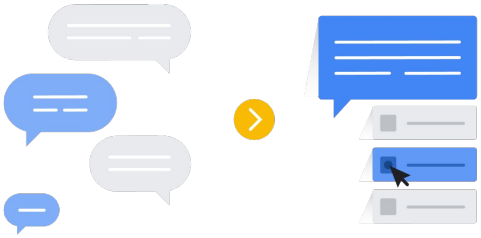
Semi-structured data falls somewhere in between structured and unstructured data. It's organized into a hierarchy, but without full differentiation or any particular ordering. Examples include emails, HTML, JSON, and XML files.

Although this data type doesn't have a formal structure, it contains tags or other markers that make it easier to analyze than unstructured data.

Organizations can use unstructured data in many ways



Analyze social media posts to identify sentiment toward a brand.

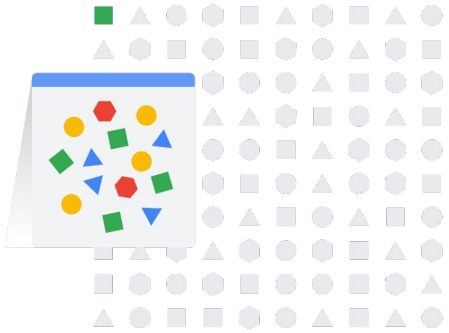


Analyze customer communications to provide responses for chatbots.

Organizations can use unstructured data in many ways. For example, a marketing team might analyze social media posts to identify sentiment toward a brand.

Or customer service teams might train automated chatbots to augment support staff by analyzing language in customer communications and providing interactive responses.

Unstructured data has historically been difficult to analyze



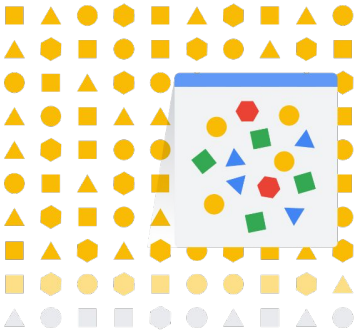
> Less than 1% of data is analyzed or used.
>
> *Harvard Business Review*

But in general, unstructured data has historically been difficult to analyze. According to the Harvard Business Review, on average less than 1% of an organization's unstructured data is analyzed or used at all. Until recently, tools to tap the potential of unstructured data were either unavailable or prohibitively expensive and complex.

Proprietary + Confidential

But unstructured data represents 80% to 90% of all new enterprise data

> Unstructured data represents 80% to 90% of all new enterprise data
>
> *Gartner research*

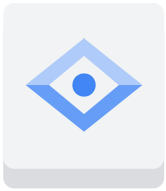


Google Cloud

What makes this statistic even more concerning is that, according to Gartner research, unstructured data represents 80% to 90% of all new enterprise data.

This reveals a staggering gap between the data being generated and the value that it's providing. But, cloud technology has changed that.

APIs can help extract value from unstructured data



Cloud
Vision API

- Uses machine learning.
- Detects products within a picture.
- Labels the picture to describe its contents.

With the right cloud tools, businesses can extract value from unstructured data by using machine learning to discover trends, or even using application programming interfaces, or APIs, to extract structure from the data. An example of an API is **Google Cloud's Vision API**, which uses machine learning to detect products within a picture and can then even label the picture to describe its contents.

Understanding the different types of data available can help organizations define what's possible with the data solutions they have. One of the transformative powers of the cloud is how it can unlock value from structured and the previously untapped, unstructured data.

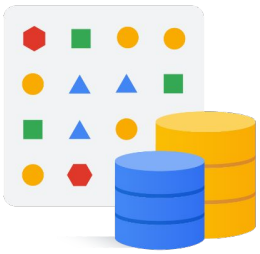
Discussion

How data-driven is your organization? Do you have unstructured data that's not being used to its full potential?

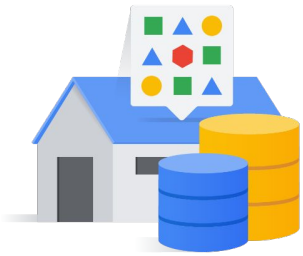
- How would you describe your company's data culture?
- How do you feel about using the full potential of your unstructured data?

Database, data warehouses, and data lakes

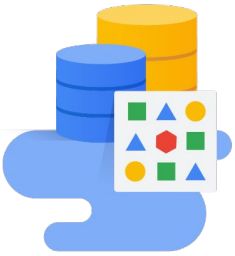
Organizations need a modern approach to managing vast volumes of data



Databases



Data warehouses



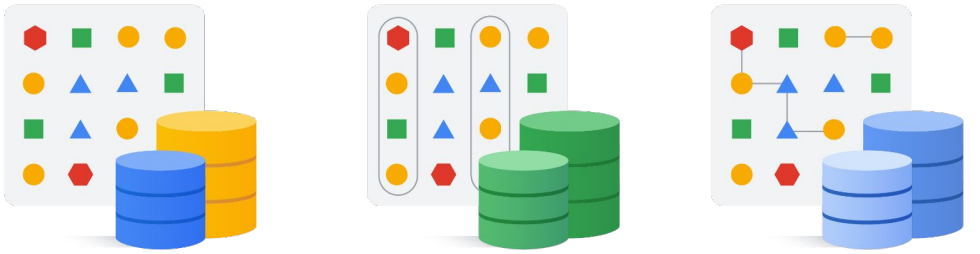
Data lakes

Organizations need a modern approach to enterprise data to manage the vast volumes that are produced. The list of options often includes **databases**, **data warehouses**, and **data lakes**.

Let's explore each of these options; starting with databases.

Database

An organized collection of data **stored in tables** and accessed electronically from a computer system.

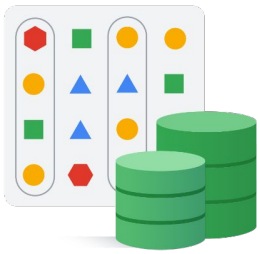


Databases → Relational databases | Non-relational databases

A **database** is an organized collection of data stored in tables and accessed electronically from a computer system.

Let's examine two types of databases: relational and non-relational.

Relational database



Relational databases

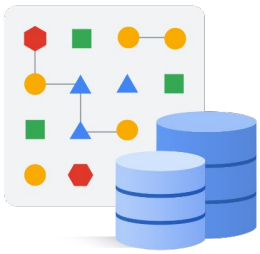
Stores data points in tables, rows, and columns that have a clearly defined schema.

- Is highly consistent and reliable
- Suited for large amounts of structured data.
- Is designed for business data processing.
- Is designed for storing online transactional data.

A **relational database** stores and provides access to data points that are related to one another. This means storing information in tables, rows, and columns that have a clearly defined schema that represents the structure or logical configuration of the database.

A relational database can establish links—or relationships–between information by joining tables, and structured query language, or SQL, can be used to query and manipulate data. Relational databases are highly consistent, reliable, and best suited for dealing with large amounts of structured data. They're designed for business data processing and storing the online transactional data needed to support the daily operations of a company.

Non-relational database



Non-relational databases (NoSQL)

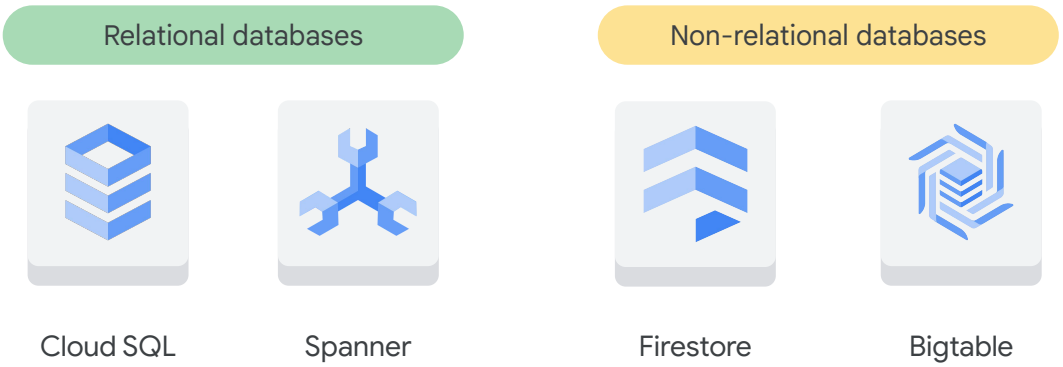
- Doesn't use a tabular format.
- Follows a flexible data model.
- Ideal for data with changing organization.
- Ideal for applications with diverse data types.

A **non-relational database**, sometimes known as a NoSQL database, is less structured in format and doesn't use a tabular format of rows and columns like relational databases.

Instead, non-relational databases follow a flexible data model, which makes them ideal for storing data that changes its organization frequently or for applications that handle diverse types of data. This includes when large quantities of complex and diverse data need to be organized, or when the data regularly evolves to meet new business requirements.

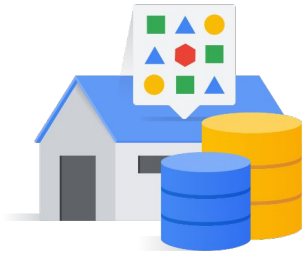
Choosing the right database depends on the use case.

Google Cloud database products



Google Cloud relational database products include Cloud SQL and Spanner, while Firestore and Bigtable are non-relational database products. We'll look at these products in more detail later.

Data warehouse



Data warehouse

An enterprise system used for the analysis and reporting of structured and semi-structured data from multiple sources.

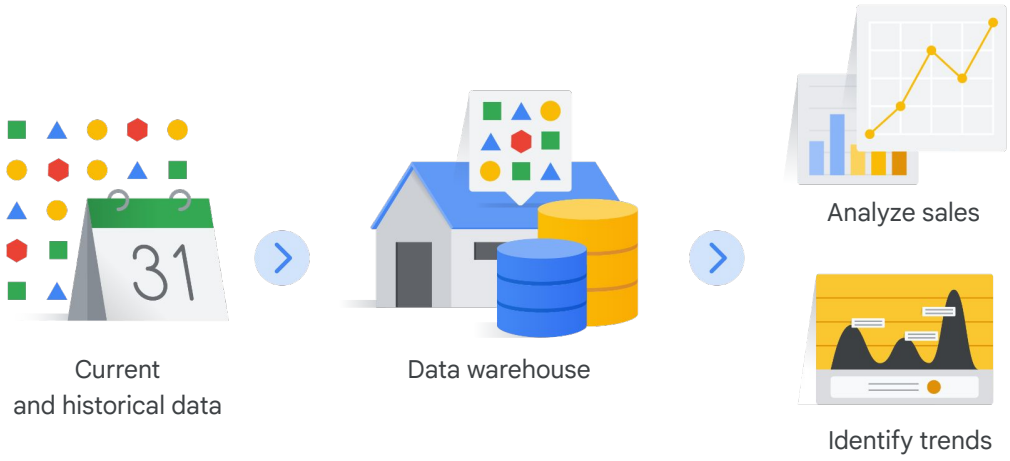
Business data:

- Point of sale
- Marketing automation
- CRM data

Let's explore another data management concept, the **data warehouse**. Like a database, a data warehouse is a place to store data. However, while a database is designed to *capture* data for storage, retrieval, and use, a data warehouse is designed to *analyze* data.

A data warehouse is an enterprise system used for the analysis and reporting of structured and semi-structured data from multiple sources. Think of the data warehouse as the central hub for all business data. Business data might include point-of-sale transactions, marketing automation, or even customer relationship management data.

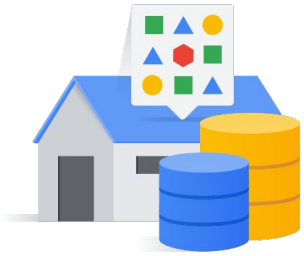
A data warehouse provides a long-range view of data over time



Suited for both ad hoc analysis and custom reporting, a data warehouse can help analyze sales and identify trends, because it can store both current and historical data in one place.

This capability can provide a long-range view of data over time, which makes a data warehouse a primary component of business intelligence.

BigQuery is Google Cloud's data warehouse offering

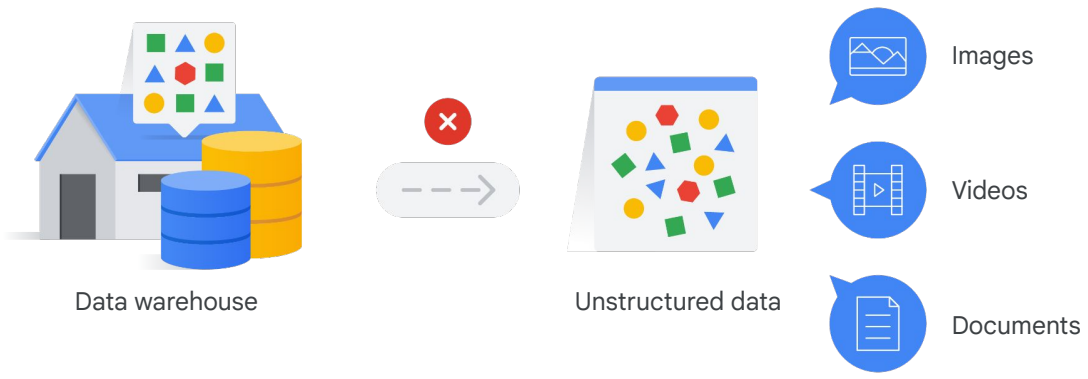


Data warehouse

BigQuery

BigQuery is Google Cloud's data warehouse offering. We'll explore BigQuery in more detail later.

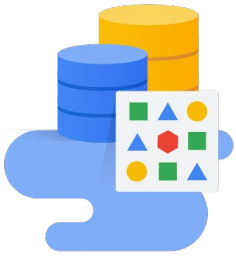
A data warehouse is not the answer for unstructured data



Although data warehouses handle structured and semi-structured data, they're not typically the answer for how to handle large amounts of available unstructured data, like images, videos, and documents.

Unstructured data, which doesn't conform to a well-defined schema, is often disregarded in traditional analytics.

Data lake



Data lake

A repository designed to ingest, store, explore, process, and analyze any type or volume of raw data

- Operational systems
- Web sources
- Social media
- IoT

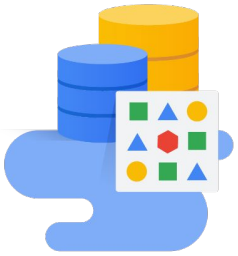
It can store **different types of data**:

- In its original format
- By ignoring size limits
- Without much preprocessing
- Without adding structure

A **data lake** is a repository designed to ingest, store, explore, process, and analyze any type or volume of raw data, regardless of the source, like operational systems, web sources, social media, or Internet of Things, or IoT.

It can store different types of data in its original format; ignoring size limits, and without much pre-processing or adding structure.

Data lake



Data lake

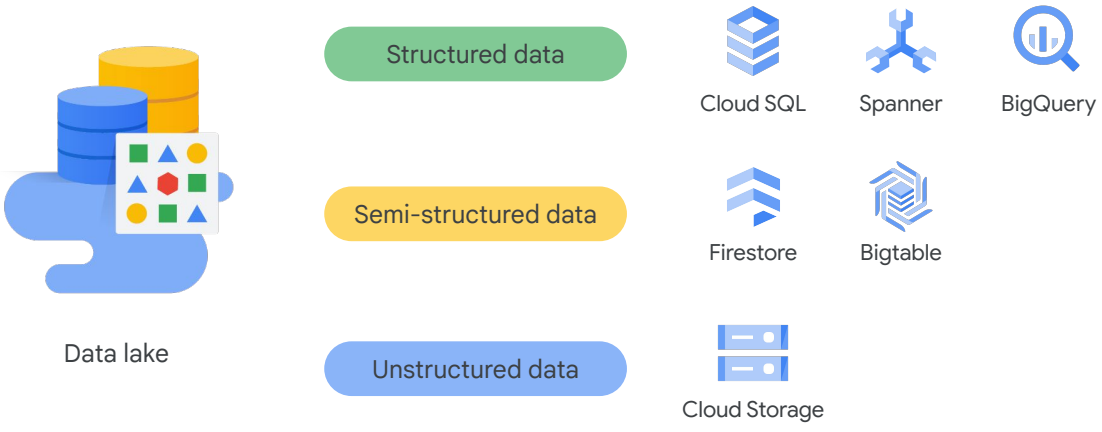
Raw data

- Prevents data contamination.
- Prevents adding bias.
- Can be enriched with other data.

Having this unprocessed, raw data available for analysis prevents unintentionally contaminating the data or adding bias. It also means that the raw data can be enriched by merging it with other data at the same time.

This differs from a data warehouse that contains structured data that has been cleaned and processed, ready for strategic analysis based on predefined business needs.

Data lakes often consist of many different products



Data lakes often consist of many different products, depending on the nature of the data that is ingested. For example:

- The best Google Cloud products for storing structured data are Cloud SQL, Spanner, or BigQuery.
- For semi-structured data, the options include Firestore and Bigtable.
- And for storing unstructured data, Cloud Storage is an option.

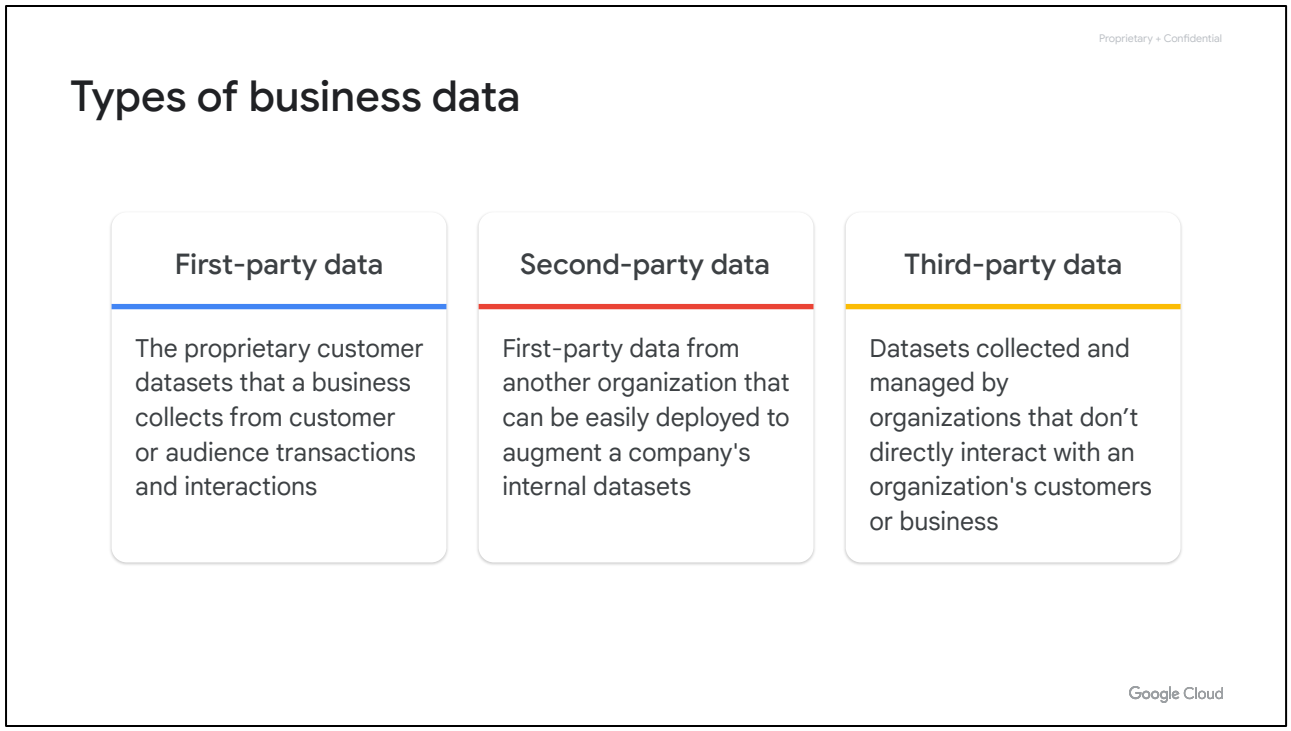
Data warehouses and data lakes are complementary tools

Data warehouses and data lakes should be considered complementary instead of competing tools. Although both store data in some capacity, each is optimized for different uses.

- Traditional data warehouse users are business intelligence analysts who are closer to the business and focus on driving insights from data. These users traditionally use the data to answer questions.
- Data lake users, and also analysts, include data engineers and data scientists. They're closer to the raw data with the tools and capabilities to explore, mine, and experiment with the data. These users find answers in the data, but they also find questions.

As enterprises are increasingly focused on data-driven decision making, data warehouses and data lakes play a critical role in an organization's digital transformation journey. Democratization of data lets users gain a deeper understanding of business situations because they have more context than ever before. Today, organizations need a 360-degree real-time view of their businesses to gain a competitive edge.

The role of data in digital transformation



As organizations have digitized their operations, many types of business data have become available, including information about their customers. This includes both internal information, called **first-party data**, and external information, which is usually data about customers and industry, often called **second** or **third-party data**.

- **First-party data** is the proprietary customer datasets that a business collects from customer or audience transactions and interactions. These datasets might include information about digital interactions, like the length of time a user spends on a web page.
- **Second-party data** often describes first-party data from *another* organization, such as a partner or other business in their supply chain, that can be easily deployed to augment a company's internal datasets. The organization does not directly own this data, but it's relevant to their business.
- Finally, there's **third-party data**, which are datasets collected and managed by organizations that don't directly interact with an organization's customers or business.

Third-party data

Government

Nonprofit

Academic

Analyst reports

Third-party data

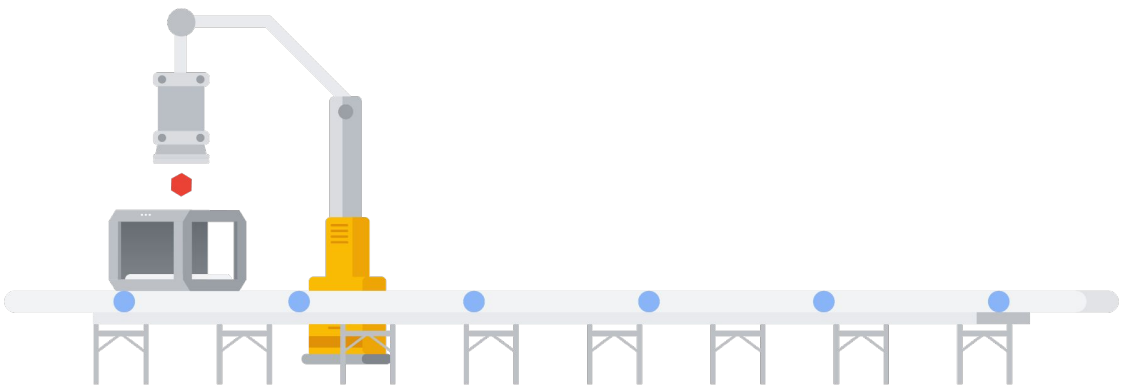
Datasets collected and managed by organizations that don't directly interact with an organization's customers or business

Third-party datasets might come from governmental, nonprofit, or academic sources, like weather or public demographic data, or from industry-specific sources like analyst reports or industry benchmarking.

Third-party data is often shared or purchased on data marketplaces or exchanges, such as the Google Cloud Marketplace. Using external data can greatly increase the value of data by providing new context and insights.

The data value chain

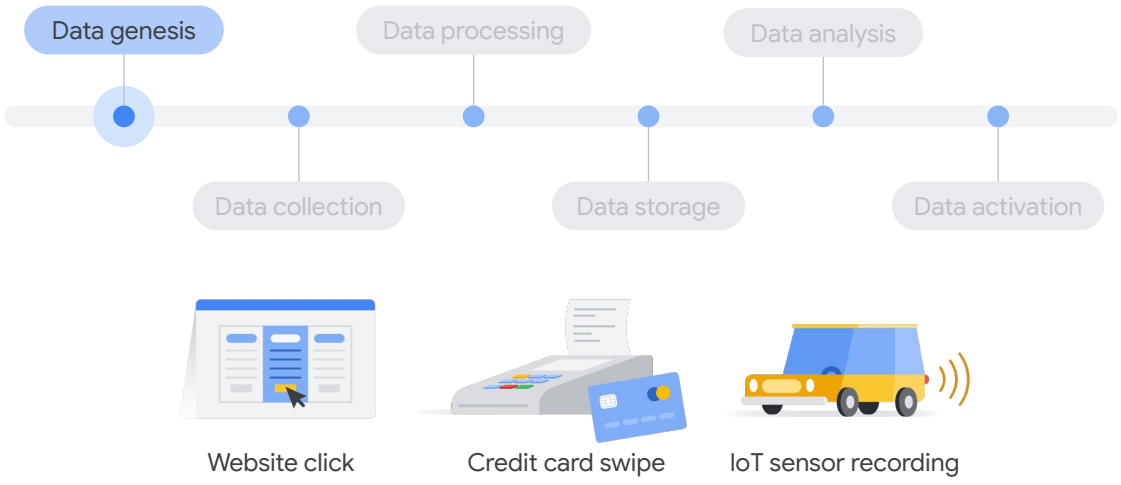
Data value chain



When you think about data processing, it's important to place it within the broader context of the data value chain.

Imagine data traveling along an assembly line, like a car in a factory. The assembly line progressively adds parts and value to an object that moves along it. Raw data at the beginning of the line is eventually transformed into actions that humans or machines take.

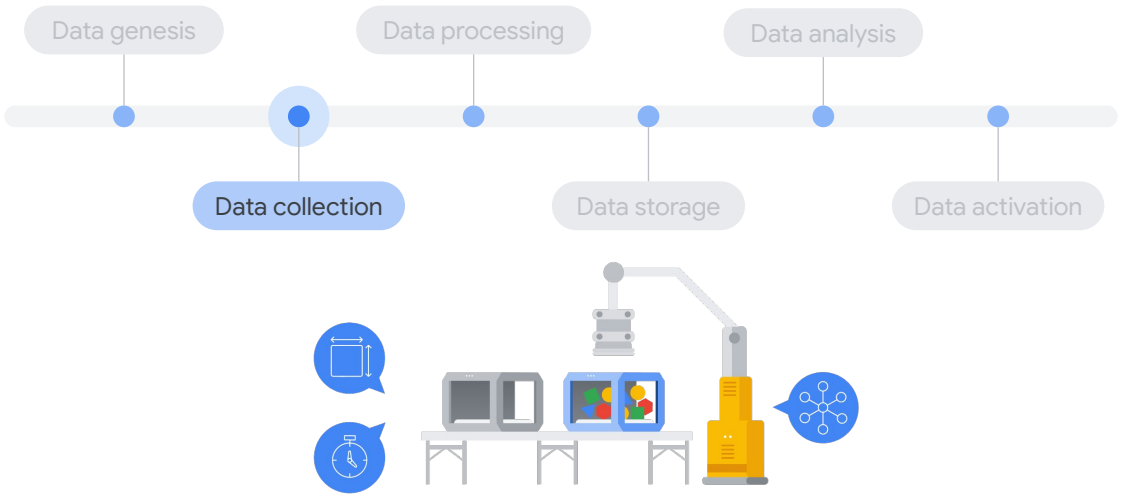
Step 1: Data genesis



Let's examine the steps in this data value chain, starting with data genesis.

Data genesis is the initial creation of a unit of data; this could be a click on a website, the swipe of a card, a sensor recording from an IoT device, or countless other examples. It's the raw material that will eventually be turned into an insight ready for action.

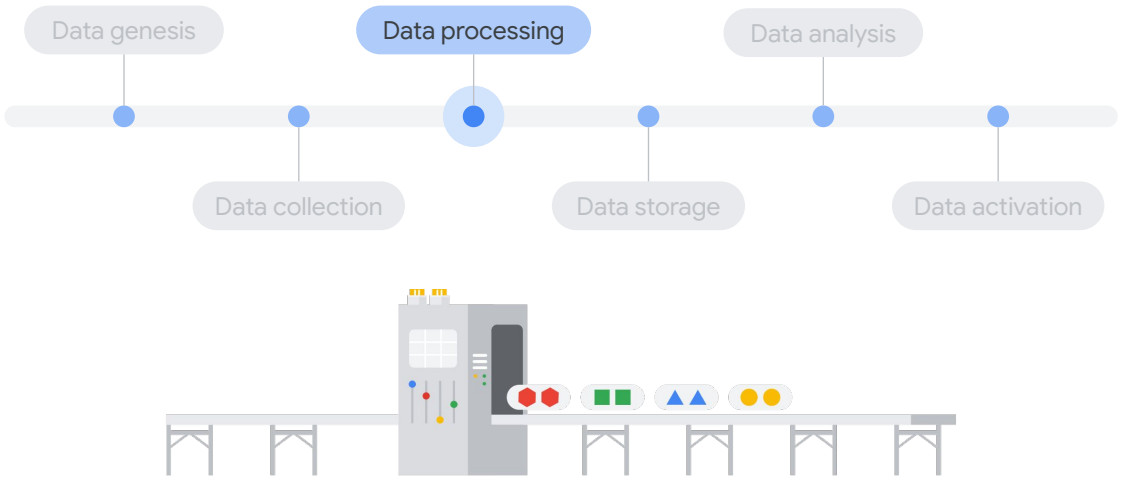
Step 2: Data collection



Data collection brings that initial unit of data to the assembly line through **ingestion**. The basic function of ingestion is to extract data from the system in which it's hosted and bring it to a new system.

It can have dramatically different requirements based on the volume, velocity, and variety of the raw data that's required for a given analysis, and how fast the data needs to be analyzed.

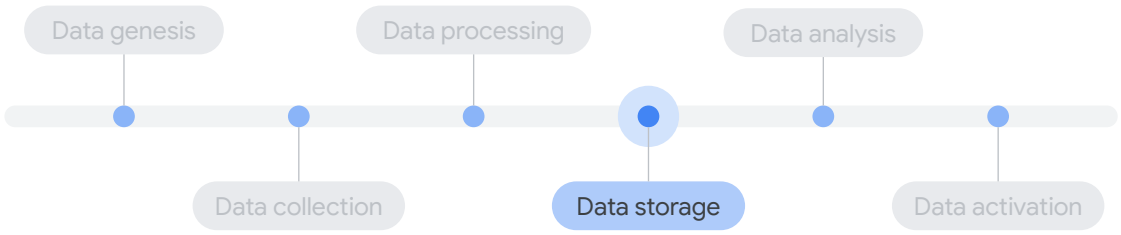
Step 3: Data processing



Data processing is where the collected raw data is transformed into a form that's ready to derive insights from. The data will likely need to be adjusted, for example, by merging different datasets together. It can be a single-stage operation, or it can be a complex tree of cascading procedures.

In our manufacturing process analogy, this phase is where raw materials take the shape of the pre-assembly parts of a manufactured product.

Step 4: Data storage



NoSQL

Fast reads and writes

Data warehousing

Fast access to analysis

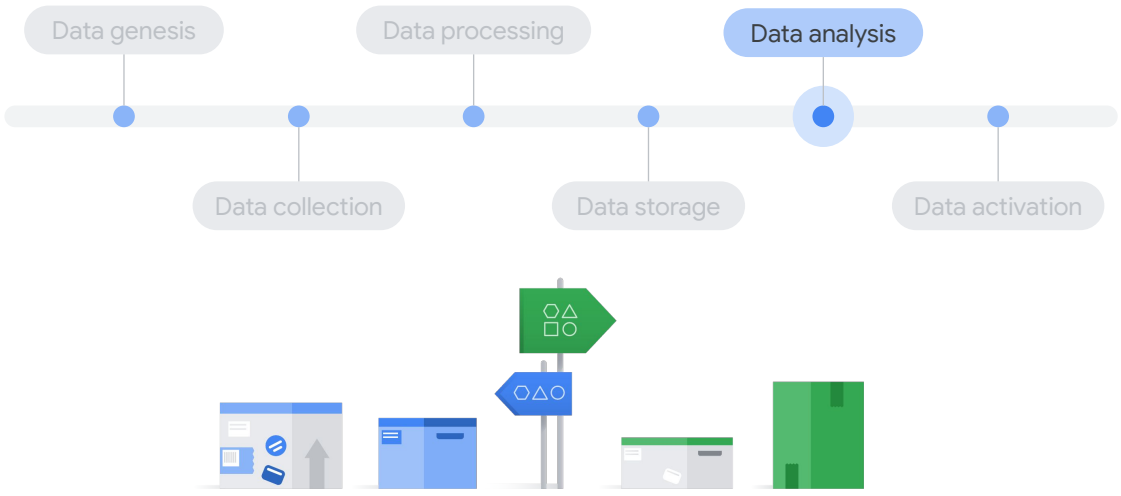
Object storage

Unstructured data

Data storage is where the data lands, can be found, and is ready for analysis and action.

As with real-world manufacturing, where storage options vary depending on the type of product that is processed, different types of data can be stored in different ways. For example, NoSQL is available for fast reads and writes, data warehousing for fast access to analysis, and object storage for unstructured data. There are also customized options of these standard stores.

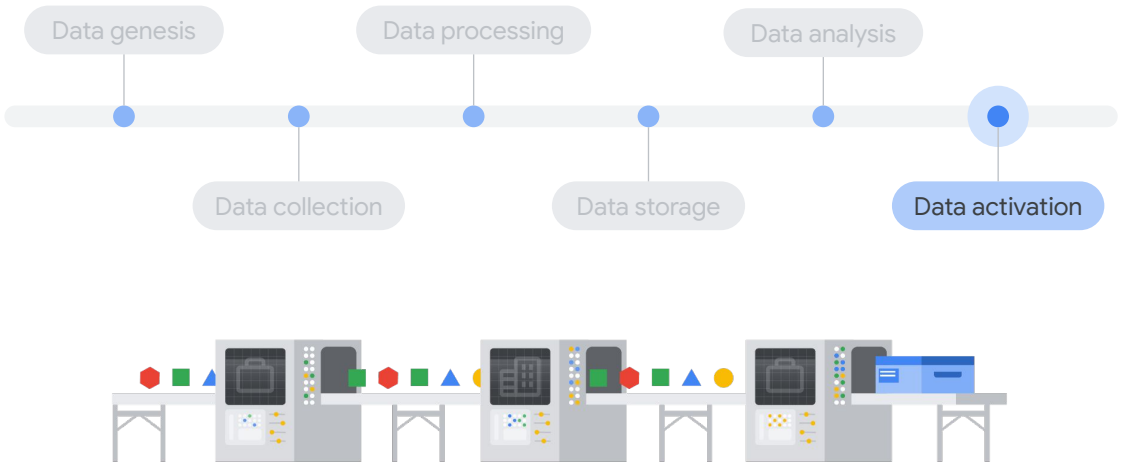
Step 5: Data analysis



Data analysis provides direction for business-oriented action.

To continue with our manufacturing line analogy, in this stage, inputs from the data processing stage are assembled into a final product.

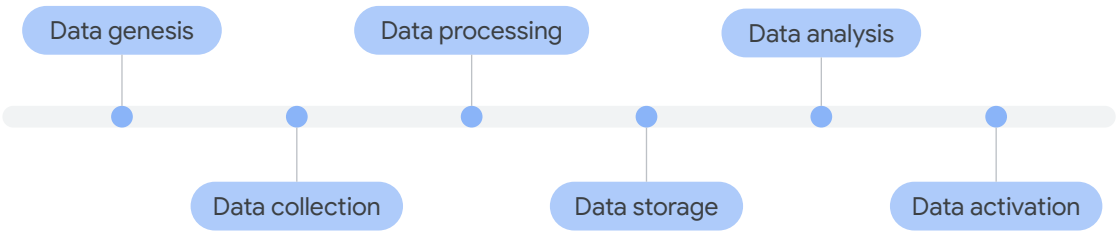
Step 6: Data activation



And finally, the last step in the data value chain is **data activation**. When an analysis is produced, it needs to be pushed to the relevant business procedures and decision makers so that action can be taken and the value chain completed.

The most common points of activation are applications that make automated decisions and business intelligence dashboards that guide humans toward better, more informed decisions. In our manufacturing line example, this is the step where a fully produced product is put to its intended use.

There is no one way to assemble a data value chain



There is no one way to assemble a data value chain, as there's no one way to create a real-world manufacturing line. Similarly, as technologies progress, new inputs become available, your workforce evolves, or the desired output changes, the optimal value chain will also change.

However, at its core, the value chain principles hold. We want to use raw data to perform actions that benefit the business.

Quiz

Question

Which of these best defines the **data value chain**?

A. The steps involved in collecting, storing, and processing data.

B. A framework for evaluating the efficiency of data processing systems.

C. A set of best practices for data security and compliance.

D. The entire process of creating and using data to generate business value.

Which of these best defines the data value chain?

A. The steps involved in collecting, storing, and processing data.
B. A framework for evaluating the efficiency of data processing systems.
C. A set of best practices for data security and compliance.
D. The entire process of creating and using data to generate business value.

Quiz

Answer

Which of these best defines the **data value chain**?

A. The steps involved in collecting, storing, and processing data.

B. A framework for evaluating the efficiency of data processing systems.

C. A set of best practices for data security and compliance.

D. The entire process of creating and using data to generate business value.

The correct answer is D.

A. The steps involved in collecting, storing, and processing data.
   - Why this is the **incorrect** answer: While these are essential steps *within* the data value chain, they don't represent its entirety. Missing are the crucial phases of utilizing the data, measuring its impact, and potentially reusing it to generate further value.

B. A framework for evaluating the efficiency of data processing systems.
   - Why this is the **incorrect** answer: This represents a potential component of the data value chain, focusing on optimization, but doesn't encapsulate the entire process of turning raw data into actionable insights.

C. A set of best practices for data security and compliance.
   - Why this is the **incorrect** answer: Data security and compliance are fundamental considerations woven throughout the data value chain, but they don't define its full scope.

D. The entire process of creating and using data to generate business value.
   - Why this is the **correct** answer: The data value chain encompasses all the stages data goes through – from identifying a need for data, to its collection, processing, analysis, usage, potential reuse, and the ultimate impact it has on decision-making or business outcomes.

Quiz

Question

Which step in the data value chain describes when collected raw data is transformed into a form that's ready to derive insights from?

A. Data genesis
B. Data processing
C. Data storage
D. Data analysis

Which step in the data value chain describes when collected raw data is transformed into a form that's ready to derive insights from?

A. Data genesis
B. Data processing
C. Data storage
D. Data analysis

Quiz

Answer

Which step in the data value chain describes when collected raw data is transformed into a form that's ready to derive insights from?

A. Data genesis

B. Data processing 

C. Data storage

D. Data analysis

The correct answer is B.

A. Data genesis
   - Why this is the **incorrect** answer: This isn't an established term in the data value chain. It might loosely refer to the initial creation of data, but not its transformation stage.

B. Data processing
   - Why this is the **correct** answer: This step is all about transforming raw data into a usable, clean, and integrated format suitable for analysis. Activities here include:
     i. Cleaning: Identifying and correcting inaccurate, incomplete, or corrupt data.
     ii. Formatting: Structuring data for easy analysis by tools and models.
     iii. Integrating: Combining data from different sources.

C. Data storage
   - Why this is the **incorrect** answer: This focuses on securely storing data both before and after processing. It's essential, but not where the transformation for analysis happens.

D. Data analysis
   - Why this is the **incorrect** answer: This focuses on extracting insights and information from the processed data, using techniques like visualization, statistical analysis, or machine learning. It relies on the output of the processing stage.

Data governance

Data governance

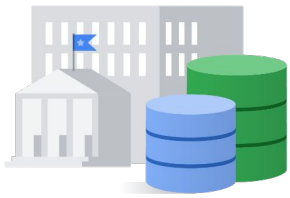
- Setting internal standards for data
- Granting access permissions
- Complying with external standards

In the last decade, the amount of data produced has increased exponentially, and the cloud has made it easier to collect, store, and analyze it at a lower cost. Organizations are now challenged to democratize and embed data in every decision, while they also ensure that it's secure and protected from unauthorized use.

An effective **data governance** program can help implement data directives to achieve this. But what exactly is data governance?

- Data governance means setting internal standards—data policies—that apply to how data is gathered, stored, processed, and disposed of.
- It governs who can access certain data and what data is under governance.
- It also involves complying with external standards set by industry associations, government agencies, and other stakeholders.

Data governance focuses on making data available to all stakeholders



Data governance

Lifecycle

Access

Insights

Standards

Data governance focuses on making the data available to all stakeholders:

- Across the full lifecycle of the data
- In a form that they can readily access and use
- In a manner that generates the desired business outcomes through insights and analysis
- And if relevant, in a way that conforms to regulatory standards and compliance needs

Data governance benefits

- It makes data more valuable.
- It helps users make better, more timely decisions.
- It improves cost controls.
- It enhances regulatory compliance.
- It helps earn greater trust from customers and suppliers.
- It helps manage risk.
- It allows more personnel access to more data.

Data governance

Data governance brings several benefits:

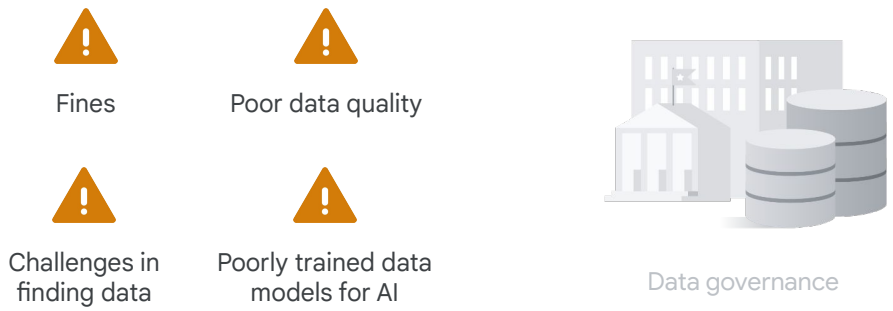
- **It makes data more valuable.** Data governance implements processes to ensure high quality data, and provides a platform that makes it easier to share data securely with stakeholders across the organization.
- **It helps users make better, more timely decisions.** Through data governance, users throughout an organization get the data they need to reach and service customers, design and improve products and services, and seize opportunities for new revenues. By democratizing data, organizations can embed data in all decision making.
- **It improves cost controls.** Data helps organizations manage resources and operate more effectively. Because they can eliminate data duplication caused by information silos, they don't overbuy—and have to maintain—expensive hardware.
- **It enhances regulatory compliance.** An increasingly complex regulatory climate has made it even more important for organizations to establish rigorous data governance practices. They avoid risks associated with noncompliance and proactively anticipate new regulations.
- **It helps earn greater trust from customers and suppliers.** By being in auditable compliance with both internal and external data policies,

- organizations gain the trust of customers and partners.

- **It helps manage risk.** With robust data governance, organizations can reduce concerns about exposure of sensitive data to individuals or systems who lack proper authorization, security breaches from malicious outsiders, or even insiders who access data they don't have the right to see.

- **It allows more personnel access to more data.** Strong data governance provides confidence that the right personnel get access to the right data, and that this democratization of data does not negatively impact the organization.

Organizations without an effective data governance program might suffer from compliance violations



It's possible that organizations without an effective data governance program will suffer from compliance violations. This can lead to:

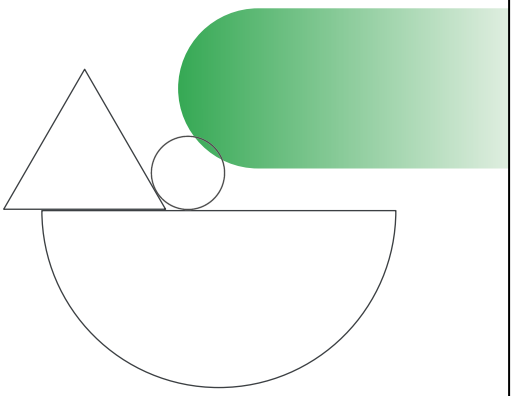
- Fines
- Poor data quality, which generates lower quality insights that impact business decisions
- Challenges in finding data, which results in delayed analysis and missed business opportunities
- And poorly trained data models for AI, which reduces the model accuracy and benefits of using AI

Every organization needs data governance. As businesses throughout all industries progress on their digital transformation journeys, data has quickly become the most valuable asset they possess.

Activity

5 min | Class | Page 13

Practice identifying the best data management solution (databases, data warehouses, and data lakes) for various use cases.

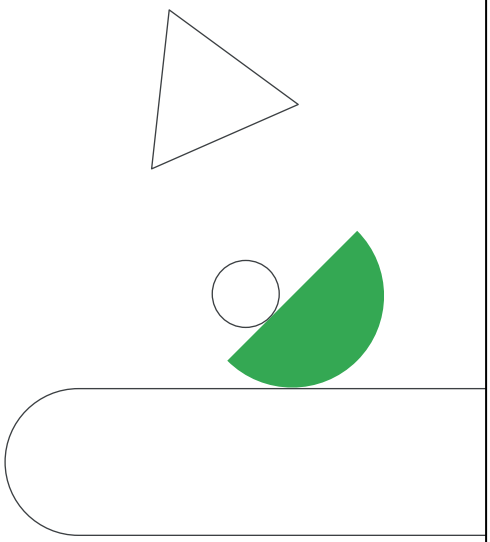


Now that you have reviewed the three main types of data management—databases, data warehouses, and data lakes— let's practice identifying the best data management solution for different use cases.

Example 1

A coworking office rental business uses an **online tool to record** daily desk, room, and meeting **bookings**. If a client books a desk for the day, that data is captured and desk availability is **updated in real time** on all customer channels. The rental business now wants to do even more with their data. They **want to use multiple types and sources of data to gain insights about facility quality** and, ultimately, to **improve their service** to customers.

Which data management solution best fits the coworking business' needs?



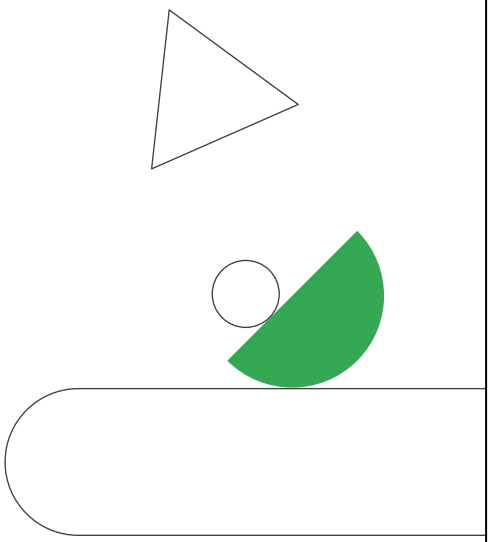
Example 1:
A coworking office rental business uses an online tool to record daily desk, room, and meeting bookings. If a client books a desk for the day, that data is captured, and desk availability is updated in real time on all customer channels. The rental business now wants to do even more with their data. They want to use multiple types and sources of data to gain insights about facility quality and, ultimately, to improve their service to customers. Which data management solution best fits the coworking business' needs?

Answer: Data warehouse

Example 1

A coworking office rental business uses an **online tool to record** daily desk, room, and meeting **bookings**. If a client books a desk for the day, that data is captured and desk availability is **updated in real time** on all customer channels. The rental business now wants to do even more with their data. They **want to use multiple types and sources of data to gain insights about facility quality** and, ultimately, to **improve their service** to customers.

Answer: Data warehouse



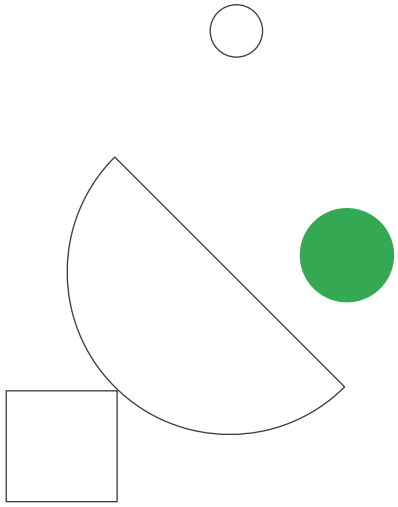
Example 1:
A coworking office rental business uses an online tool to record daily desk, room, and meeting bookings. If a client books a desk for the day, that data is captured and desk availability is updated in real time on all customer channels. The rental business now wants to do even more with their data. They want to use multiple types and sources of data to gain insights about facility quality and, ultimately, to improve their service to customers. Which data management solution best fits the coworking business' needs?

Answer: Data warehouse

Example 2

A bank is launching a **mobile banking app**, and wants to **track money transfers** from one account to another. They want to ensure that the transferred figure is **updated** in the bank's records in **real time** and the user is able to see the **most up-to-date account balance**.

Which data management solution best fits this bank's needs?



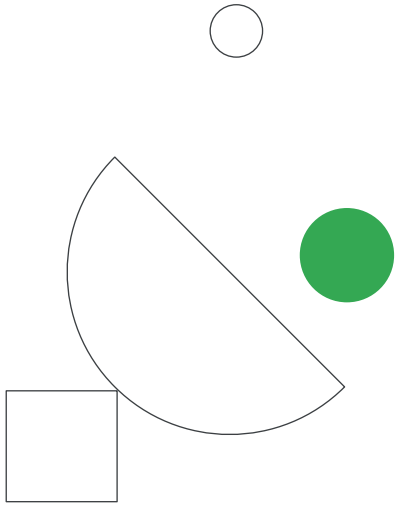
Example 2:
A bank is launching a mobile banking app, and wants to track money transfers from one account to another. They want to ensure that the transferred figure is updated in the bank's records in real time and the user is able to see the most up-to-date account balance. Which data management solution best fits this bank's needs?

Answer: Database

Example 2

A bank is launching a **mobile banking app**, and wants to **track money transfers** from one account to another. They want to ensure that the transferred figure is **updated** in the bank's records in **real time** and the user is able to see the **most up-to-date account balance**.

Answer: Database



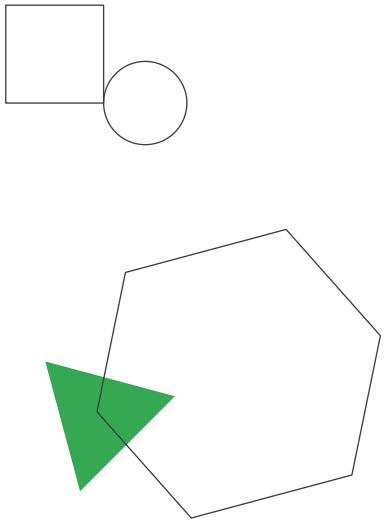
Example 2:
A bank is launching a mobile banking app, and wants to track money transfers from one account to another. They want to ensure the transferred figure is updated in the bank's records in real time and the user is able to see the most up-to-date account balance. Which data management solution best fits this bank's needs?

Answer: Database

Example 3

An online music streaming company **stores raw music data** that is accessed by users worldwide and constantly analyzed by their systems. They want to geographically disperse **backup copies** of their **raw data in very large volumes**. This data comes in a **variety of formats**, must retain **full fidelity**, and be **accessible for processing and analysis at any time, at short notice**.

Which data management solution best fits this streaming service's needs?



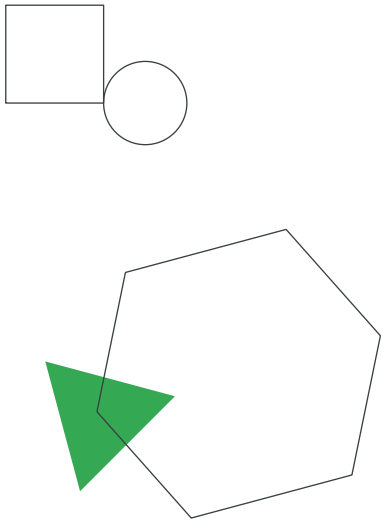
Example 3:
An online music streaming company stores raw music data that is accessed by users worldwide and constantly analyzed by their systems. They want to geographically disperse backup copies of their raw data in very large volumes. This data comes in a variety of formats, must retain full fidelity, and be accessible for processing and analysis at any time, at short notice. Which data management solution best fits this streaming service's needs?

Answer: Data lake

Example 3

An online music streaming company **stores raw music data** that is accessed by users worldwide and constantly analyzed by their systems. They want to geographically disperse **backup copies** of their **raw data in very large volumes**. This data comes in a **variety of formats**, must retain **full fidelity**, and be **accessible for processing and analysis at any time, at short notice**.

Answer: Data lake



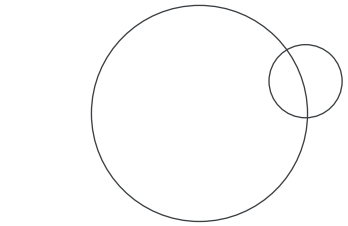
Example 3:
An online music streaming company stores raw music data that is accessed by users worldwide and constantly analyzed by their systems. They want to geographically disperse backup copies of their raw data in very large volumes. This data comes in a variety of formats, must retain full fidelity, and be accessible for processing and analysis at any time, at short notice. Which data management solution best fits this streaming service's needs?

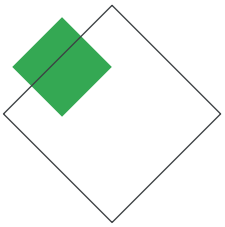
Answer: Data lake

Example 4

A lifestyle company is launching a casual dating **mobile app**. By signing onto the app through social media, **users provide details** such as gender, location, and interests, and headshot images. The lifestyle company wants to **display this information** to other app users **through an algorithm**, which **depends on compatibility**, and needs a **cost-effective data management** solution that can hold **large volumes of data**. They also **can't afford downtime** that would drive users away.

Which data management solution best fits this company's needs?





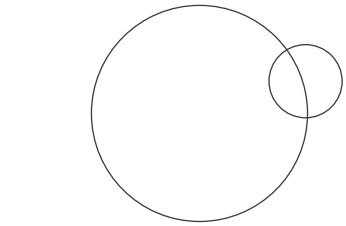
Example 4:
A lifestyle company is launching a casual dating mobile app. By signing onto the app through social media, users provide details such as gender, location, and interests, and headshot images. The lifestyle company wants to display this information to other app users through an algorithm, which depends on compatibility, and needs a cost-effective data management solution that can hold large volumes of data. They also can't afford downtime that would drive users away. Which data management solution best fits this company's needs?

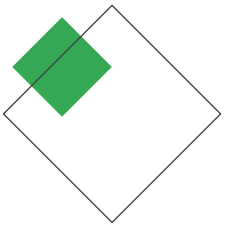
Answer: Database

Example 4

A lifestyle company is launching a casual dating **mobile app**. By signing onto the app through social media, **users provide details** such as gender, location, and interests, and headshot images. The lifestyle company wants to **display this information** to other app users **through an algorithm**, which **depends on compatibility**, and needs a **cost-effective data management** solution that can hold **large volumes of data**. They also **can't afford downtime** that would drive users away.

Answer: Database





Example 4:
A lifestyle company is launching a casual dating mobile app. By signing onto the app through social media, users provide details such as gender, location, and interests, and headshot images. The lifestyle company wants to display this information to other app users through an algorithm, which depends on compatibility, and needs a cost-effective data management solution that can hold large volumes of data. They also can't afford downtime that would drive users away. Which data management solution best fits this company's needs?

Answer: Database

Module 2

Exploring Data Transformation with Google Cloud

Lessons

01 The value of data

02 Google Cloud data management solutions

03 Making data useful and accessible

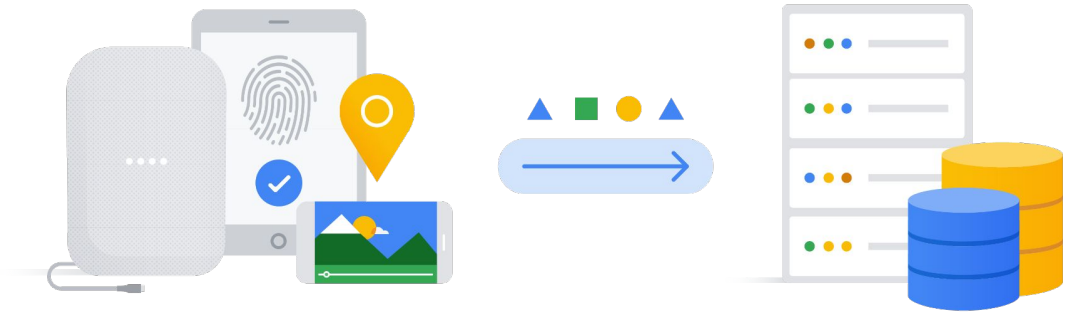
Google Cloud offers a wide range of data management products and solutions, each applicable to different business use cases.

In this section of the course, you'll explore:

- Google Cloud data management options and the differences between them.
- The different storage classes available with Cloud Storage.
- How to choose the right storage product to meet the needs of your organization.
- And ways an organization can migrate and/or modernize their current database in the cloud.

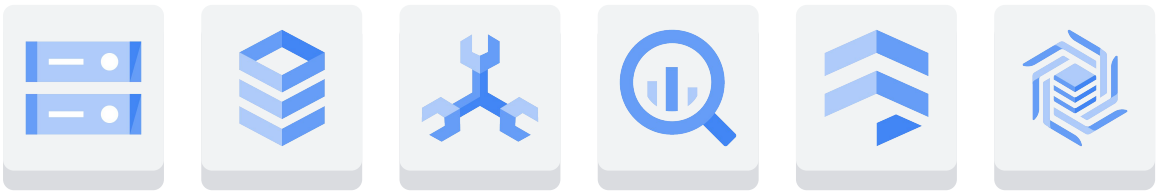
Unstructured data storage

Every application needs to store data



Every application needs to store data, like media to be streamed or even sensor data from devices, and different applications and workloads require different storage solutions.

Google Cloud's core storage products

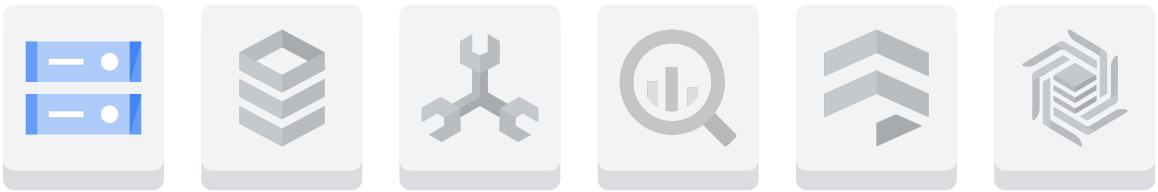


Google Cloud offers several core storage products. This list includes:

- Cloud Storage
- Cloud SQL
- Spanner
- BigQuery
- Firestore
- And Bigtable

Depending on your use case, you might use one or several of these services to do the job.

Google Cloud's core storage products



Cloud Storage

Let's begin with **Cloud Storage**, which is a service that offers developers and IT organizations durable and highly available **object storage**.

What is object storage?

Object storage
is a computer data storage architecture that manages data as *objects*.

File storage
stores data in a file and folder hierarchy.

Block storage
stores data as fixed-sized chunks on a disk.

But what is object storage?

Object storage is a computer data storage architecture that manages data as “objects” instead of as file storage, which is a file and folder hierarchy, or as block storage, which is chunks of a disk.

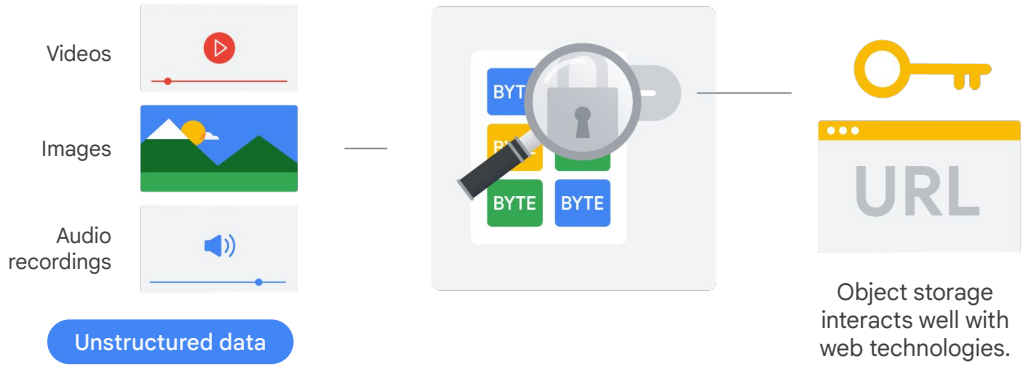
Object storage

Object packaged format contains:

- Binary form of the actual data itself.
- Relevant associated metadata.
- Globally unique identifier.

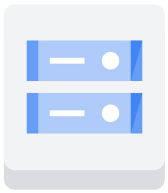
These objects are stored in a packaged format that contains the binary form of the actual data, and relevant associated metadata–such as creation date, author, resource type, and permissions–and a globally unique identifier.

Unique keys are URLs



These unique keys are in the form of URLs, which means object storage interacts well with web technologies. Data commonly stored as objects include video, pictures, and audio recordings.

This type of data is referred to as **unstructured**, which means that it doesn't have a predefined data model or isn't organized in a predefined manner, as you might find in a structured database format.



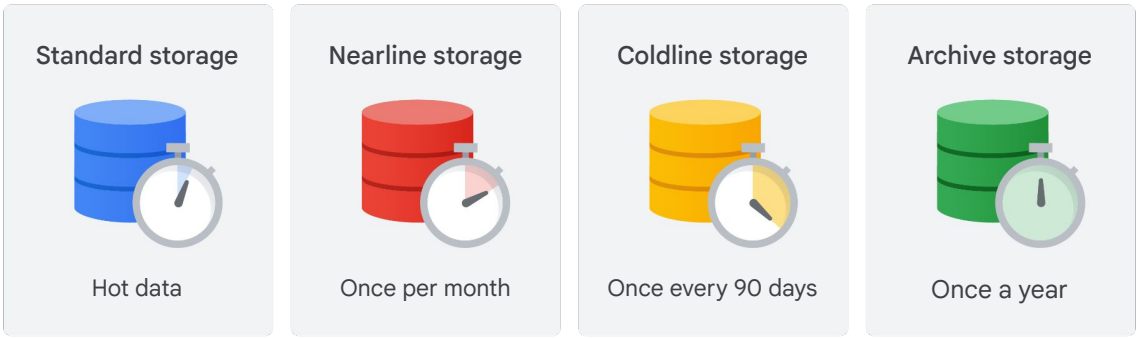
Cloud Storage

- Allows customers to store any amount of data, and to retrieve it as often as needed.
- Fully managed scalable service that has a wide variety of uses.

Cloud Storage lets customers store any amount of data and retrieve it as often as needed.

It's a fully managed, scalable service that has a wide variety of uses, such as serving website content, storing data for archival and disaster recovery, and distributing large data objects to end users through direct download.

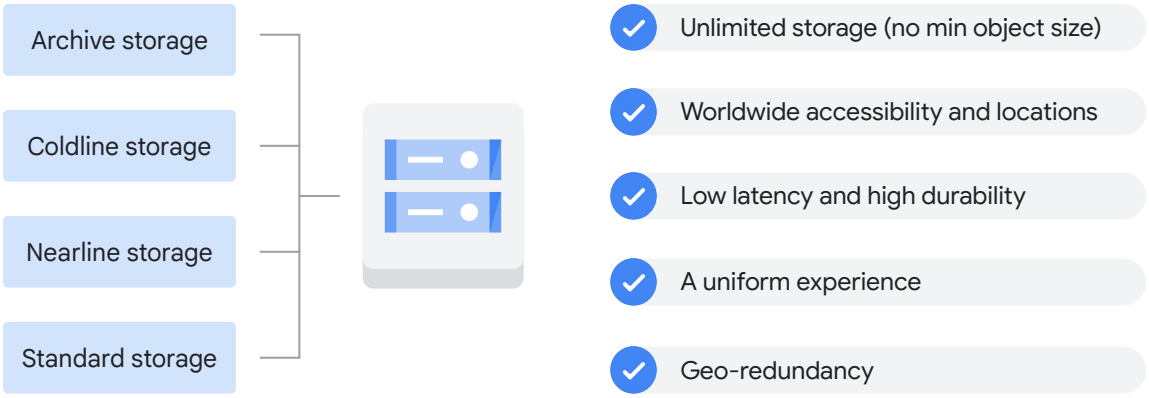
Cloud Storage's four primary storage classes



There are four primary storage classes in Cloud Storage.

- The first is **Standard storage**. Standard Storage is considered best for frequently accessed, or "hot," data. It's also great for data that's stored for only brief periods of time.
- The second storage class is **Nearline storage.** This option is best for storing infrequently accessed data, like reading or modifying data on average once a month or less. Examples might include data backups, long-tail multimedia content, or data archiving.
- The third storage class is **Coldline storage**. This is also a low-cost option for storing infrequently accessed data. However, as compared to Nearline storage, Coldline storage is meant for reading or modifying data, at most, once every 90 days.
- And the fourth storage class is **Archive storage**. This is the lowest-cost option, used ideally for data archiving, online backup, and disaster recovery. It's the best choice for data that you plan to access less than once a year, because it has higher costs for data access and operations and a 365-day minimum storage duration.

Common characteristics among storage classes



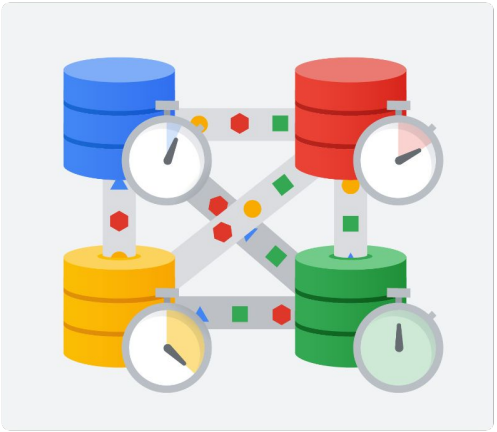
Although each of these four classes have differences, it's worth noting there are several characteristics that apply across all of these storage classes, which include:

- Unlimited storage with no minimum object size requirement
- Worldwide accessibility and locations
- Low latency and high durability
- A uniform experience, which extends to security, tools, and APIs
- And geo-redundancy if data is stored in a multi-region or dual-region (this means placing physical servers in geographically diverse data centers to protect against catastrophic events and natural disasters, and load-balancing traffic for optimal performance)

Autoclass automatically transitions objects to appropriate storage classes

Autoclass

- Moves data that is not accessed to colder storage classes to reduce storage cost.
- Moves data that is accessed to standard storage to optimize future accesses.



Cloud Storage also provides a feature called **Autoclass**, which automatically transitions objects to appropriate storage classes based on each object's access pattern.

The feature moves data that is not accessed to colder storage classes to reduce storage cost and moves data that is accessed to Standard storage to optimize future accesses.

Autoclass simplifies and automates cost saving for your Cloud Storage data.

Quiz

Question

A data analyst for an online retailer must produce a sales report at the end of each quarter. Which Cloud Storage class should the retailer use for data accessed every 90 days?

A. Coldline

B. Standard

C. Nearline

D. Archive

A data analyst for an online retailer must produce a sales report at the end of each quarter. Which Cloud Storage class should the retailer use for data accessed every 90 days?

A. Coldline
B. Standard
C. Nearline
D. Archive

Quiz

Answer

A data analyst for an online retailer must produce a sales report at the end of each quarter. Which Cloud Storage class should the retailer use for data accessed every 90 days?

A. Coldline 

B. Standard

C. Nearline

D. Archive

The correct answer is A.

A. Coldline
- Why this is the **correct** answer: Coldline storage is designed for data that is infrequently accessed, with a minimum storage duration of 90 days. Since the sales report is produced quarterly, this aligns perfectly with Coldline's access patterns.

B. Standard
- Why this is the **incorrect** answer: Standard storage is meant for frequently accessed data, with no minimum storage duration. This class would be more expensive without additional benefit.

C. Nearline
- Why this is the **incorrect** answer: Nearline has a minimum storage duration of 30 days. If the reports were monthly, this would be the right choice, but quarterly access makes it less ideal.

D. Archive
- Why this is the **incorrect** answer: Archive storage is for long-term retention of data not frequently accessed, with a minimum storage duration of 365 days. This level of infrequency makes it less suitable for the quarterly report use case.

Quiz

Question

Which characteristic **applies to all** Cloud Storage classes?

A. Maximum storage limits

B. Accessibility only within one region

C. Geo-redundancy if data is stored in a multi-region or dual-region

D. High latency and low durability

Which characteristic is true for all Cloud Storage classes?

A. Maximum storage limits
B. Accessibility only within one region
C. Geo-redundancy if data is stored in a multi-region or dual-region
D. High latency and low durability

Quiz

Answer

Which characteristic **applies to all** Cloud Storage classes?

A. Maximum storage limits

B. Accessibility only within one region

C. Geo-redundancy if data is stored in a multi-region or dual-region 

D. High latency and low durability

The correct answer is C.

A. Maximum storage limits
   - Why this is the **incorrect** answer: Cloud Storage has virtually no maximum storage limits for any of its classes. Users can store as much data as they need.

B. Accessibility only within one region
   - Why this is the **incorrect** answer: This depends on the specific storage class and how data is stored. Cloud Storage classes, except for Archive, are accessible from anywhere with internet connectivity.

C. Geo-redundancy if data is stored in a multi-region or dual-region
   - Why this is the **correct** answer: All Cloud Storage classes offer geo-redundancy when data is placed in multi-regional or dual-regional locations. This means multiple copies of the data are stored across geographically distinct locations, providing enhanced data protection and high availability.

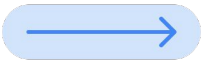
D. High latency and low durability
   - Why this is the **incorrect** answer: All Cloud Storage classes are designed to provide low latency (quick response times) and incredibly high durability (often expressed as 11 nines: 99.999999999%). Low durability would make them unreliable for long-term data storage.

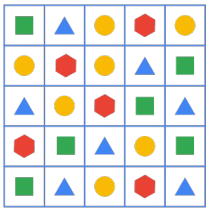
Structured data storage

Now let's explore some Google Cloud data storage products that are suited for storing *structured* data.

Structured data

Relational databases



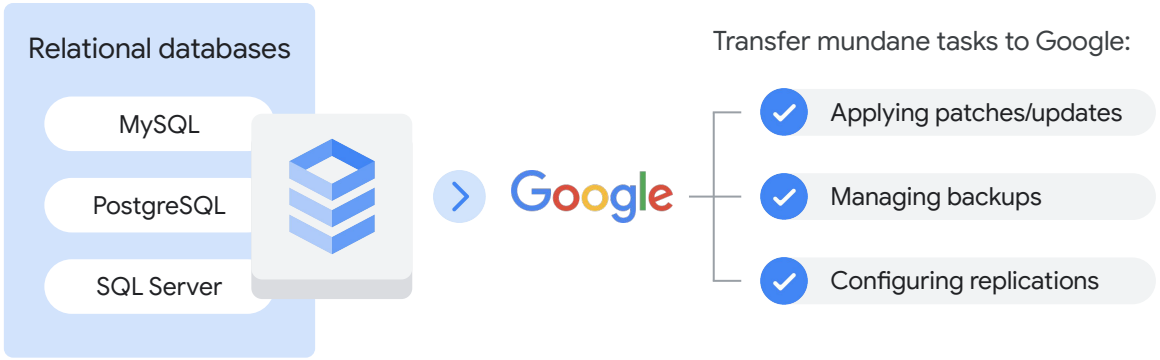


Stores information in tables, rows, and columns that have a clearly defined schema that represents the structure or logical configuration of the database.

Structured data consists of numbers and values that are organized in a predefined format that's easily searchable in a relational database.

Earlier in the course, we mentioned that a relational database stores information in tables, rows, and columns that have a clearly defined schema that represents the structure or logical configuration of the database.

Cloud SQL



Cloud SQL offers fully managed relational databases, including MySQL, PostgreSQL, and SQL Server as a service.

It's designed to transfer mundane—but necessary and often time-consuming—tasks to Google, like applying patches and updates, managing backups, and configuring replications, so you can focus on building great applications.

Cloud SQL benefits

Cloud SQL

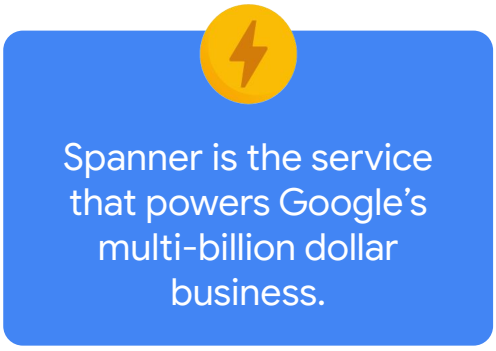
- Is trusted by thousands of the largest enterprises around the world.
- Doesn't require software installation or maintenance.
- Supports managed backups.
- Encrypts customer data.
- Includes a network firewall.

Trusted by thousands of the largest enterprises around the world, organizations that use Cloud SQL obtain various benefits.

- It doesn't require any software installation or maintenance.
- It supports *managed backups*, so backed-up data is securely stored and accessible if a restore is required.
- It encrypts customer data when on Google's internal networks and when stored in database tables, temporary files, and backups.
- And it includes a network firewall, which controls network access to each database instance.

Spanner

A fully managed, mission-critical, relational database service that scales horizontally to handle unexpected business spikes



Spanner is a fully managed, mission-critical, relational database service that scales horizontally to handle unexpected business spikes. Battle tested by Google's own mission-critical applications and services, Spanner is the service that powers Google's multi-billion dollar business.

Spanner user cases

Spanner

Especially suited for applications that require:

- SQL relational database management system with joins and secondary indexes
- Built-in high availability
- Strong global consistency
- High numbers of input/output operations per second

Spanner is especially suited for applications that require:

- A SQL relational database management system with joins and secondary indexes
- Built-in high availability, which provides data redundancy to reduce downtime when a zone or instance becomes unavailable (the goal is to prevent a single point of failure)
- Strong global consistency, which ensures that all locations where data is stored are updated to the most recent data version quickly
- And high numbers of input and output operations per second (tens of thousands of reads and writes per second or more)

Differences between Cloud SQL and Spanner

Cloud SQL

It's a fully managed relational database service for MySQL, PostgreSQL, and SQL Server.

Its availability is greater than 99.95%.

DMS makes it easy to migrate your production databases to Cloud SQL with minimal downtime.

Spanner

It's a fully managed relational database with unlimited scale, strong consistency.

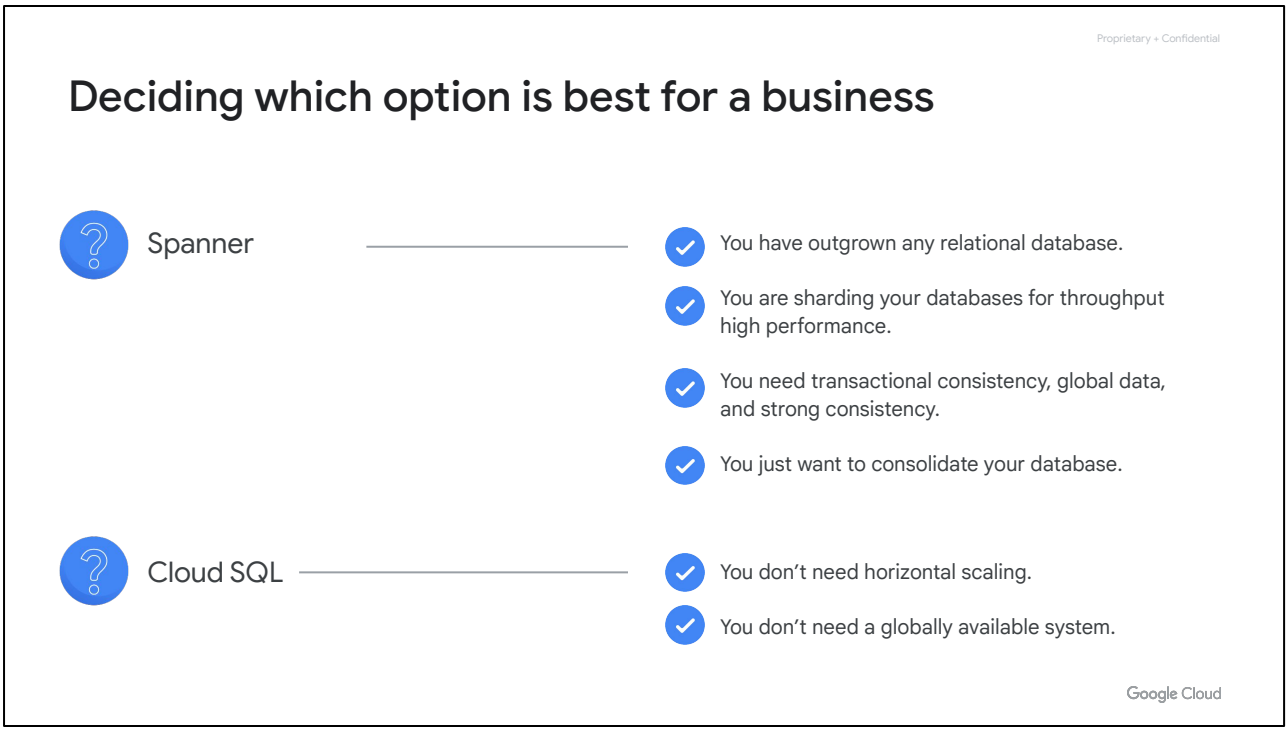
It's availability reaches 99.999%.

It handles replicas, sharding, and transaction processing, so you can quickly scale to meet any usage pattern and ensure success of products.

Both Cloud SQL and Spanner are fully managed database services, but how do they differ?

Cloud SQL is a fully managed relational database service for MySQL, PostgreSQL, and SQL Server with greater than 99.95% availability. Database Migration Service (DMS) makes it easy to migrate your production databases to Cloud SQL with minimal downtime.

And then there is Spanner, which is a fully managed relational database with unlimited scale, strong consistency, and up to 99.999% availability with zero downtime for planned maintenance and schema changes. This globally distributed, ACID-compliant cloud database automatically handles replicas, sharding, and transaction processing, so you can quickly scale to meet any usage pattern and ensure success of products.

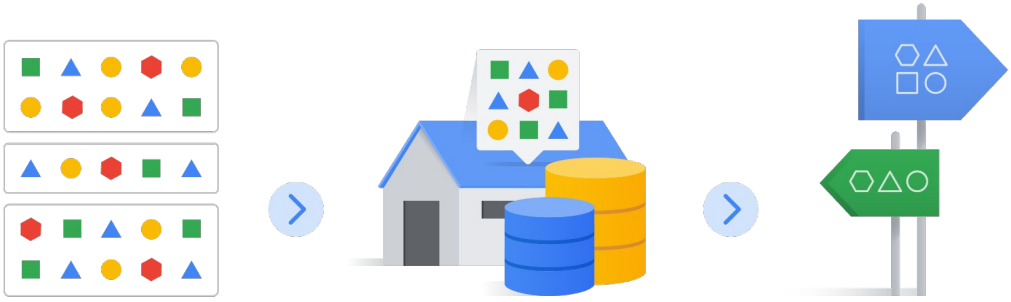


When considering which option is best for your business, consider this:

If you have outgrown any relational database; are sharding your databases for throughput high performance; need transactional consistency, global data, and strong consistency; or just want to consolidate your database, consider using **Spanner**.

If you don't need horizontal scaling or a globally available system, **Cloud SQL** is a cost-effective solution.

BigQuery

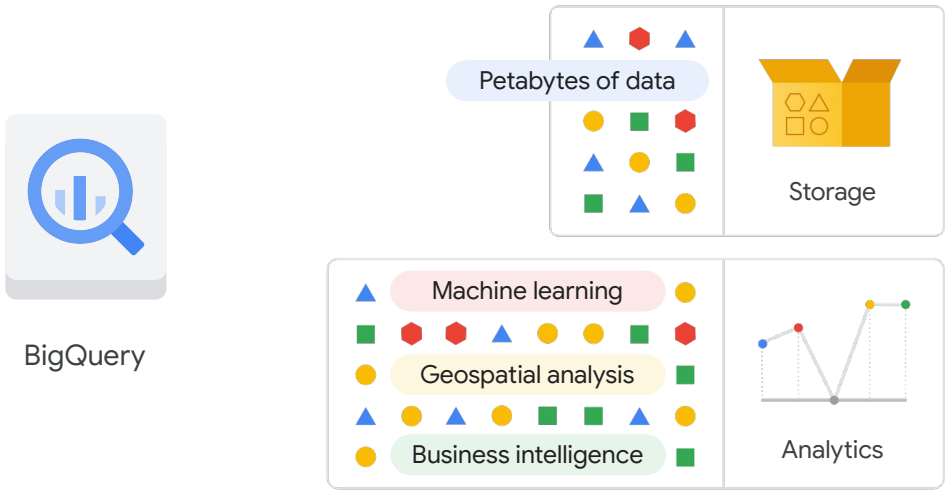


BigQuery is a fully-managed data warehouse.

The final structured data storage solution that we'll explore is **BigQuery**. BigQuery is a fully-managed data warehouse.

As you've already heard, a data warehouse is a large store that contains **petabytes** of data gathered from a wide range of sources within an organization and is used to guide management decisions. Because it's fully managed, BigQuery takes care of the underlying infrastructure, so users can focus on using SQL queries to answer business questions, without having to worry about deployment, scalability, and security.

BigQuery provides two services in one



BigQuery provides two services in one: **storage and analytics**.

It's a place to *store* petabytes of data. For reference, one petabyte is equivalent to 11,000 movies at 4k quality.

BigQuery is also a place to *analyze* data, with built-in features like machine learning, geospatial analysis, and business intelligence.

BigQuery features

BigQuery

- Data in BigQuery is encrypted at rest by default.
- BigQuery seamlessly integrates with the partner ecosystem.
- BigQuery works in a multicloud environment.
- BigQuery has built-in machine learning features.

Data in BigQuery is **encrypted** at rest by default without any action required from a user. Encryption at rest is encryption used to protect data that's stored on a disk, including solid-state drives, or backup media.

BigQuery provides **seamless integration** with the existing partner ecosystem. Businesses can tap into our ecosystem of system integrators and data integration partners to help enhance analytics and reporting. These integrations mean that BigQuery lets organizations make the most of existing investments in business intelligence, data ingestion, and data integration tools.

Industry research shows that 90% of organizations have a multicloud strategy, which adds complexity to data integration, orchestration, and governance. BigQuery works in a **multicloud environment**, which lets data teams eradicate data silos by using BigQuery to securely and cost effectively analyze data across multiple cloud providers.

BigQuery also has **built-in machine learning** features so that ML models can be written directly in BigQuery by using SQL. And if other professional tools—such as Vertex AI from Google Cloud—are used to train ML models, datasets can be exported from BigQuery directly into Vertex AI for a seamless integration across the data-to-AI lifecycle

Quiz

Question

Which is the best SQL-based storage option for a transactional workload that requires local or regional scalability?

A. Cloud Storage

B. Spanner

C. BigQuery

D. Cloud SQL

Which is the best SQL-based storage option for a transactional workload that requires local or regional scalability?

A. The on-premises networking is more complicated.
B. Scaling processing is too difficult due to power consumption.
C. Maintenance workers do not have physical access to the servers.
D. The on-premises hardware procurement process can take a long time.

Quiz

Answer

Which is the best SQL-based storage option for a transactional workload that requires local or regional scalability?

A. Cloud Storage

B. Spanner

C. BigQuery

D. Cloud SQL 

The correct answer is D.

A. Cloud Storage
   - Why this is the **incorrect** answer: Cloud Storage is an object storage service. It's highly scalable and great for images, videos, and backup files, but not designed for structured, transactional data that requires SQL queries.

B. Spanner
   - Why this is the **incorrect** answer: Spanner is a globally distributed, strongly consistent database. While highly scalable, it's overkill for a workload only requiring regional scaling and adds operational complexity over a traditional SQL database.

C. BigQuery
   - Why this is the **incorrect** answer: BigQuery is a serverless data warehouse designed for analytical queries on large datasets. It's an excellent choice for analytics, but not ideal for traditional transactional database operations.

D. Cloud SQL
   - Why this is the **correct** answer: Cloud SQL is the best fit for transactional workloads, particularly when local or regional scalability is important because it's a **fully-managed relational database** offering MySQL, PostgreSQL, and SQL Server engines, tailor-made for transactional data, it **supports both vertical (scaling up resources for greater capacity) and horizontal scaling** (adding read replicas) for

- regional workloads, and it uses **standard SQL syntax**, making it easy to work with for those familiar with SQL databases.

Quiz

Question

Which would be the best SQL-based storage option for a transactional workload that requires global scalability?

A. Cloud Storage
B. Spanner
C. BigQuery
D. Cloud SQL

Which would be the best SQL-based storage option for a transactional workload that requires global scalability?

A. Cloud Storage
B. Spanner
C. BigQuery
D. Cloud SQL

Quiz

Answer

Which would be the best SQL-based storage option for a transactional workload that requires global scalability?

A. Cloud Storage

B. Spanner 

C. BigQuery

D. Cloud SQL

The correct answer is B.

A. Cloud Storage
   - Why this is the **incorrect** answer: This is object storage, not optimized for transactional data and relational querying required by transactional workloads.

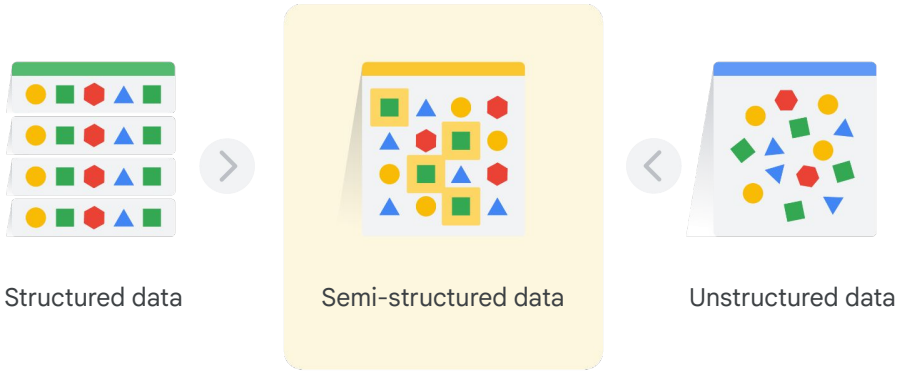
B. Spanner
   - Why this is the **correct** answer: It is the best choice for transactional workloads needing global scalability because it is designed for global distribution, with automatic data replication across continents. This provides low latency access and resilience for users worldwide. For highly consistent transactions that must be reflected everywhere instantly, Spanner excels. Spanner seamlessly handles data growth across regions, accommodating globally distributed user bases.

C. BigQuery
   - Why this is the **incorrect** answer: This is a data warehousing solution primarily suited for analyzing large-scale datasets and not real-time transaction processing.

D. Cloud SQL
   - Why this is the **incorrect** answer: While great for transactional workloads, its regional scalability, though easily configured, limits it compared to Spanner's global reach.

Semi-structured data storage

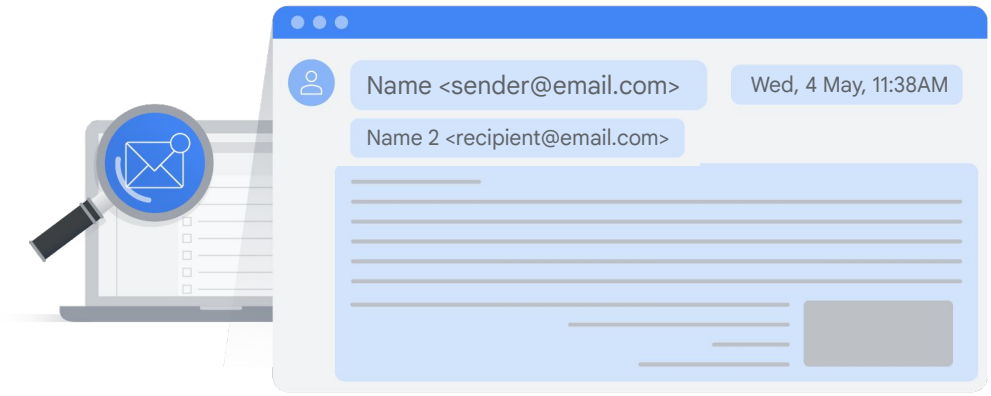
Semi-structured data



Semi-structured data contains elements of both structured and unstructured data. It does have some defining or consistent characteristics, but generally doesn't follow a structure as rigid as a relational database.

Semi-structured data is easier to organize because it usually contains some organizational properties, such as tags or metadata.

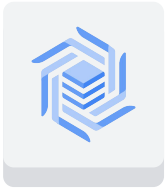
An email has unstructured data and structured data



An example of semi-structured data is email. While the actual content of the email is unstructured, it does contain structured data such as the name and email address of the sender and recipient, the time sent, and so on.

Google Cloud semi-structured data storage products

Firestore



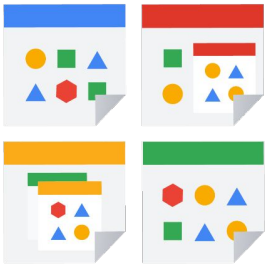
Bigtable

Google Cloud offers two semi-structured data storage products, **Firestore** and **Bigtable**.

Firestore

Firestore

Firestore is a flexible, horizontally scalable, NoSQL cloud database for storing and syncing data in real-time..



It performs data storage in the form of documents.

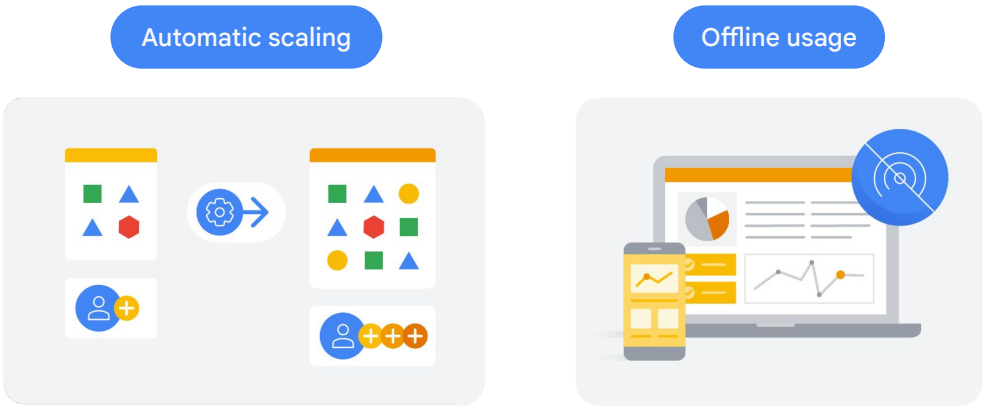
Documents are stored in collections.

Firestore is a flexible, horizontally scalable, NoSQL cloud database for storing and syncing data in real-time. Firestore can be directly accessed by mobile and web applications.

Firestore performs data storage in the form of documents, with the documents being stored in collections.

Documents support a wide variety of data types, such as nested objects, numbers, and strings.

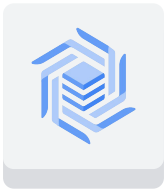
Firestore features



One of Firestore's main features is automatic scaling. It's been designed to scale automatically depending on user demand, but retains the same level of performance irrespective of database size.

Firestore also provides offline usage through a comprehensive database on users' devices. Offline data access ensures that applications run without interruption, even if the user gets disconnected from the internet.

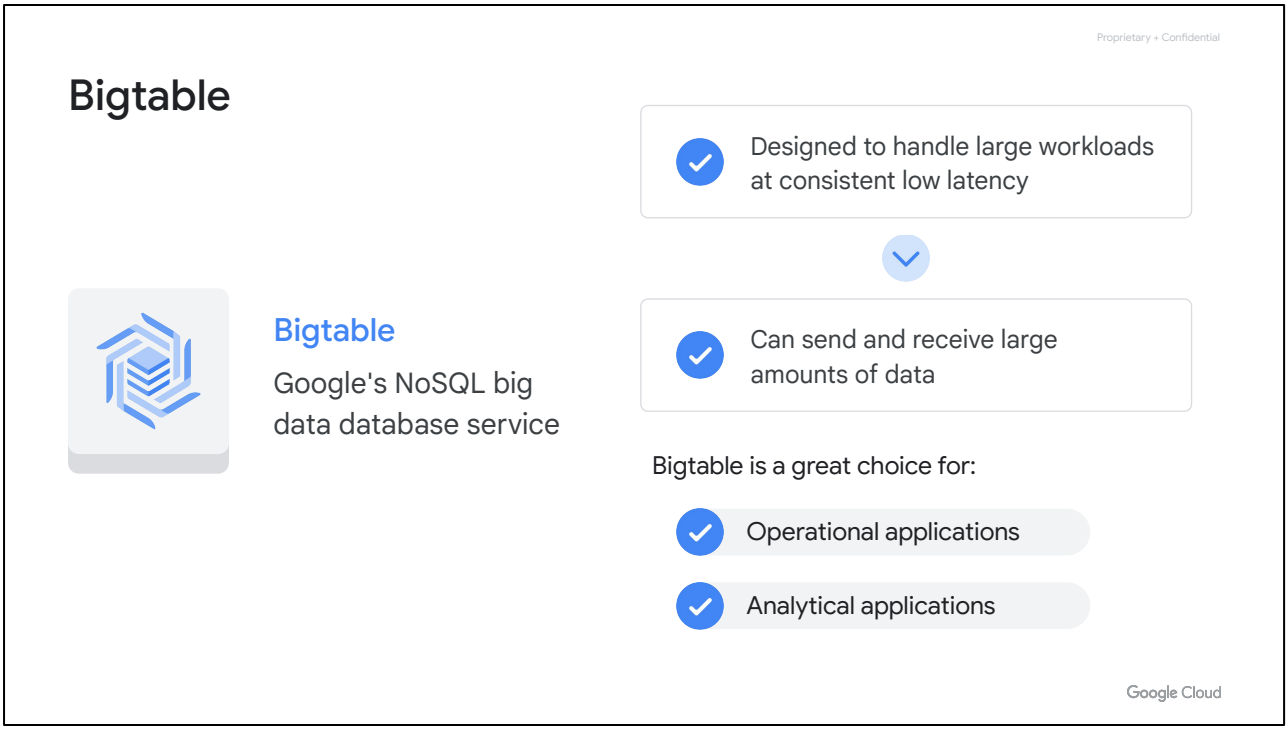
Bigtable



Bigtable

Google's NoSQL big data database service

And then there is **Bigtable,** Google's NoSQL big data database service. It's the same database that powers many core Google services, including Search, Analytics, Maps, and Gmail.



Bigtable is designed to handle large workloads at consistent low latency, which means Bigtable responds to requests quickly, with high throughput, which means it can send and receive large amounts of data.

For this reason, it's a great choice for both operational and analytical applications, including Internet of Things, user analytics, and financial data analysis.

You might use Bigtable if ...

Bigtable

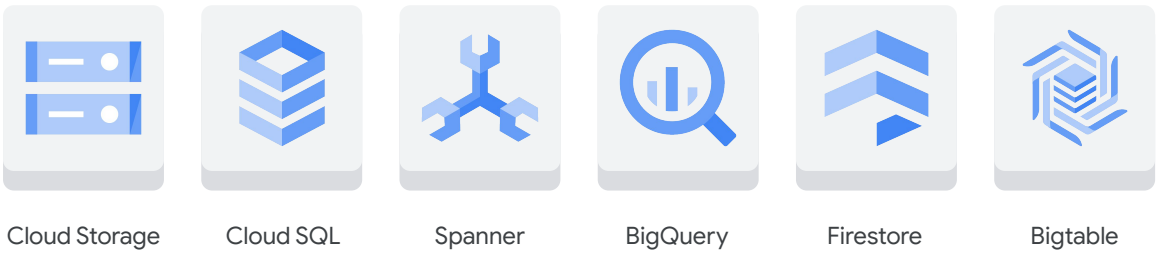
- You're working with more than 1 TB of semi-structured or structured data.
- Data is fast with high throughput, or it's rapidly changing.
- You're working with NoSQL data.
- Data is a time-series or has natural ordering.
- You're working with big data and running batch or real-time processing on the data.
- You're running machine learning algorithms on the data.

When deciding on a storage option, you might choose Bigtable if:

- You're working with more than 1 TB of semi-structured or structured data.
- Data is fast with high throughput, or it's rapidly changing.
- You're working with NoSQL data.
- Data is a time-series or has natural ordering.
- You're working with big data and running batch or real-time processing on the data.
- Or you're running machine learning algorithms on the data.

Choosing the right storage product

What's the right scenario for each storage option?

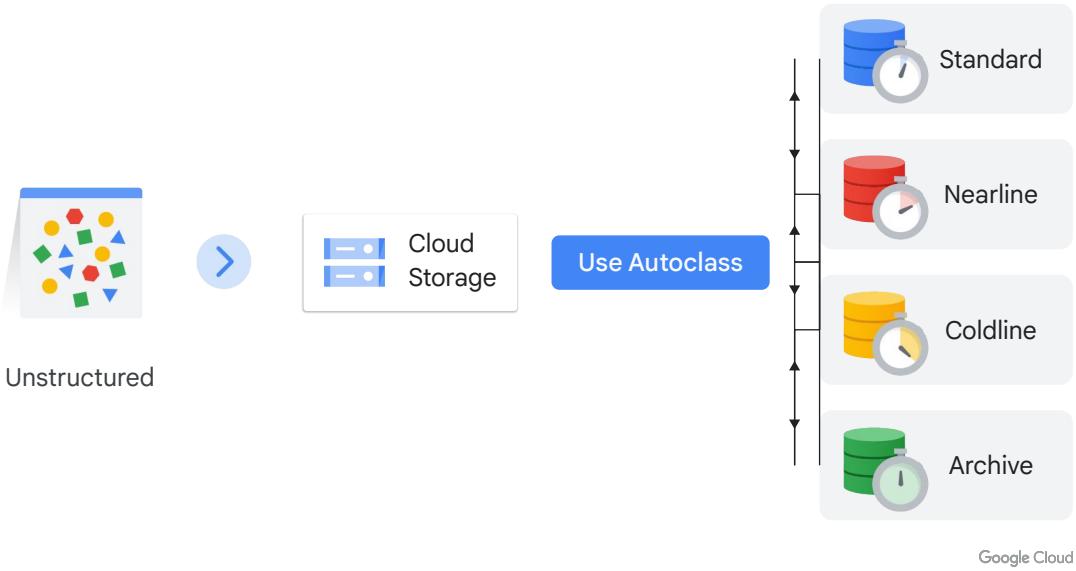


Data type + Business need

So, you've learned about the different storage options that Google Cloud offers, but in what scenarios should you use each one?

Ultimately, it's a combination of the **data type** that needs to be stored and the **business need**.

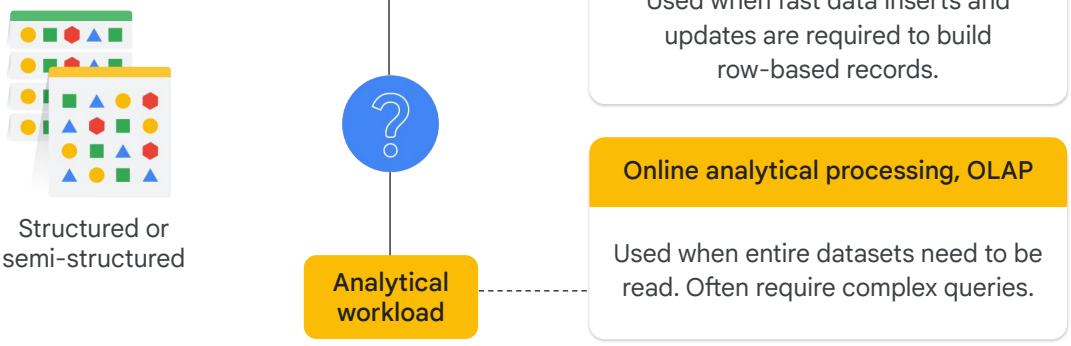
Use Cloud Storage for unstructured data



If data is unstructured, then **Cloud Storage** is the most appropriate option.

You have to decide a storage class: Standard, Nearline, Coldline, or Archive. Or whether to let the Autoclass feature decide that for you.

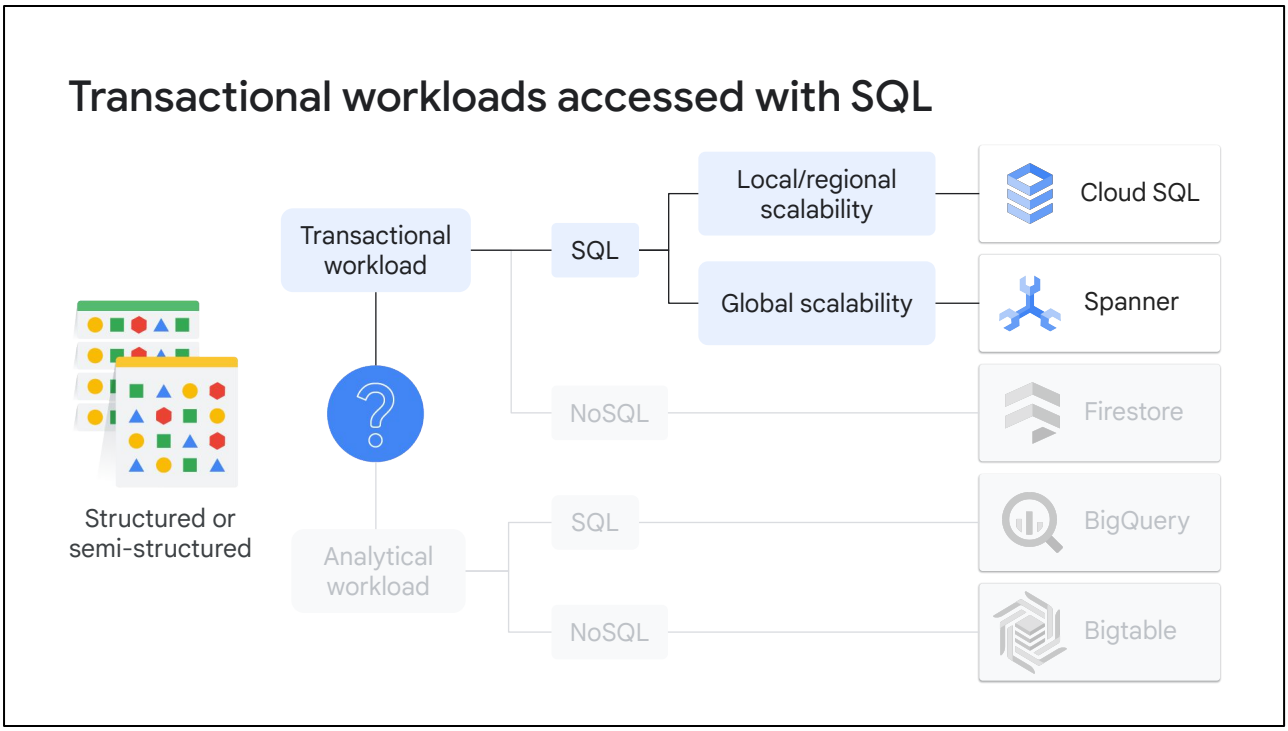
Structured or semi-structured data



If data is structured or semi-structured, choosing a storage product will depend on whether workloads are **transactional** or **analytical**.

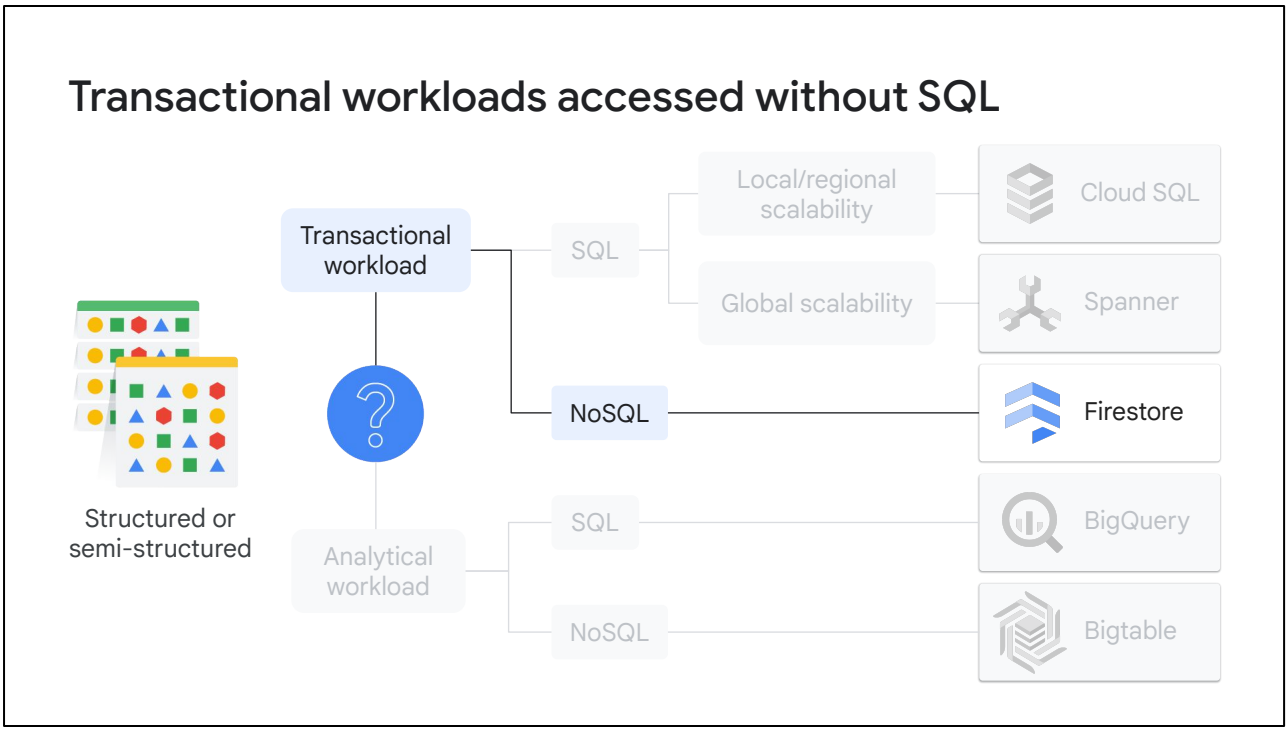
Transactional workloads stem from online transaction processing, or OLTP, systems, which are used when fast data inserts and updates are required to build row-based records. An example of this is point-of-sale transaction records.

Then there are analytical workloads, which stem from online analytical processing, or OLAP systems, which are used when entire datasets need to be read. They often require complex queries, for example, aggregations. An example here would be analyzing sales history to see trends and aggregated views.



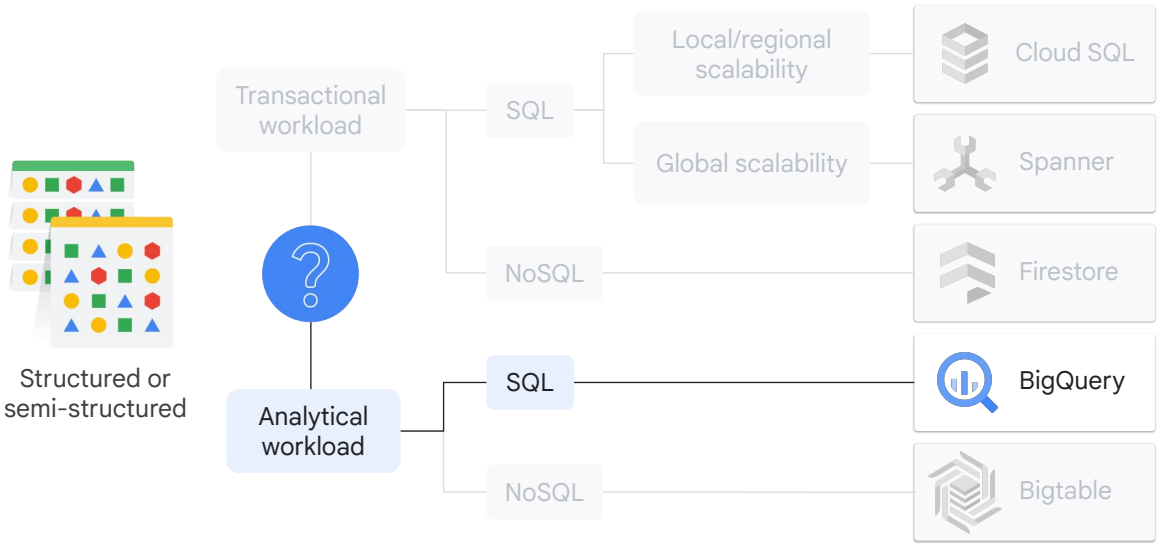
After you determine if the workloads are transactional or analytical, you must determine whether the data will be accessed by using SQL.

So, if your data is transactional and you need to access it by using SQL, then **Cloud SQL** and **Spanner** are two options. Cloud SQL works best for local to regional scalability, and Spanner is best to scale a database globally.



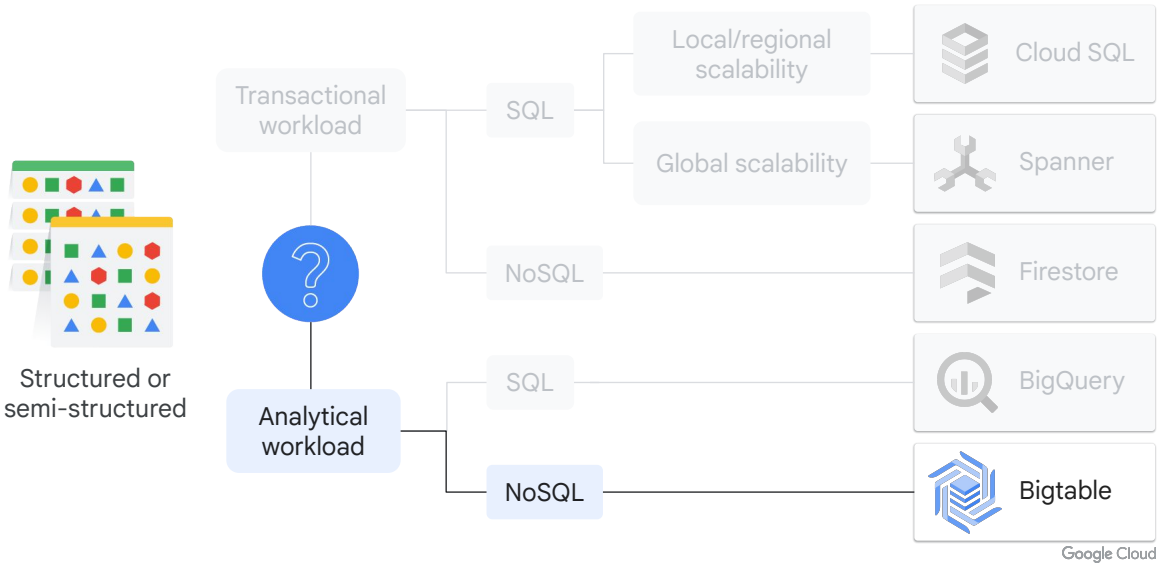
If the transactional data will be accessed without SQL, **Firestore** might be the best option. Firestore is a transactional NoSQL, document-oriented database. (less rigid then a schema; doesn't use SQL query, more programmatic language)

Analytical workloads that require SQL commands



If you have analytical workloads that require SQL commands, **BigQuery** might be the best option. BigQuery, Google's data warehouse solution, lets you analyze petabyte-scale datasets.

Analytical workloads that don't require SQL commands



Alternatively, **Bigtable** provides a scalable NoSQL solution for analytical workloads. It's best for real-time, high-throughput applications that require only **millisecond latency**.

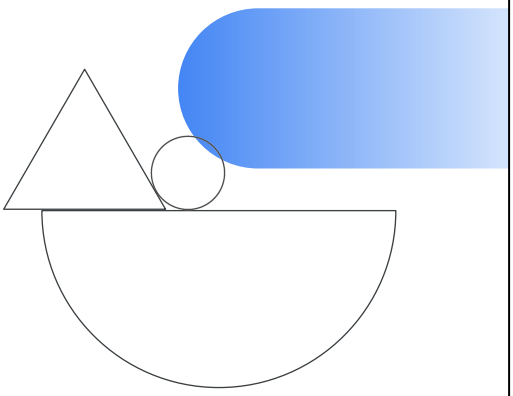
You can use Bigtable to store and query all of the following types of data:

- Time-series data, such as CPU and memory usage over time for multiple servers
- Marketing data, such as purchase histories and customer preferences
- Financial data, such as transaction histories, stock prices, and currency exchange rates
- Internet of Things data, such as usage reports from energy meters and home appliances
- Graph data, such as information about how users are connected to one another

Activity

5 min | Class | Page 15

Identify the Google Cloud storage product that matches the listed criteria.



Let's put what you've just learned into practice. On the slides that follow, you'll be show a list of criteria. Identify the best Google Cloud storage product.

Which storage product...?

Let's start with the first column.

Which storage product can handle structured or unstructured data, and can access transactional workloads without SQL?

Options include: Bigtable, Cloud Storage, Firestore, Spanner, BigQuery, or Cloud SQL.

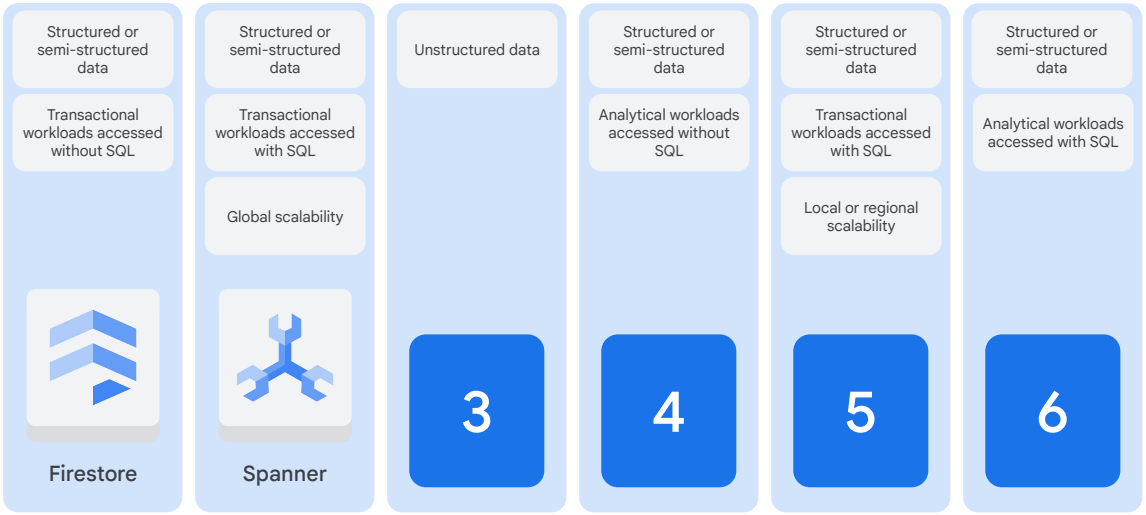
Which storage product...?

The correct answer is **Firestore**. Now let's move on to the second column.

Which storage product can handle structured or semi-structured data, access transactional workloads with SQL, and scale globally?

Options include: Bigtable, Cloud Storage, Firestore, Spanner, BigQuery, or Cloud SQL.

Which storage product...?

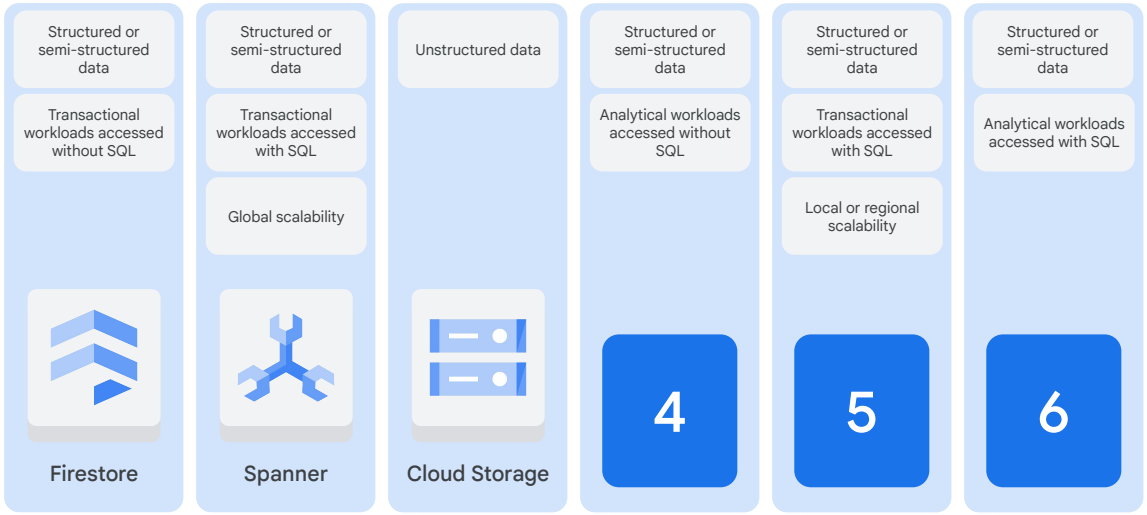


The correct answer is **Spanner**. Now let's move on to the third column.

Which storage product can store unstructured data?

Options include: Bigtable, Cloud Storage, Firestore, Spanner, BigQuery, or Cloud SQL.

Which storage product...?

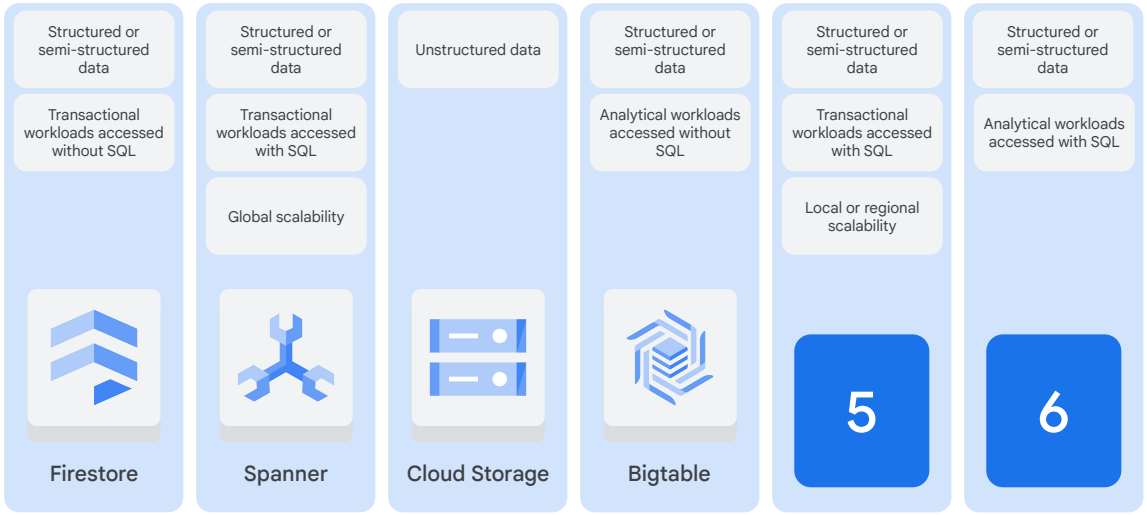


The correct answer is **Cloud Storage**. Now let's move on to the fourth column.

Which storage product can handle structured or semi-structured data, and can access analytical workloads without SQL?

Options include: Bigtable, Cloud Storage, Firestore, Spanner, BigQuery, or Cloud SQL.

Which storage product...?

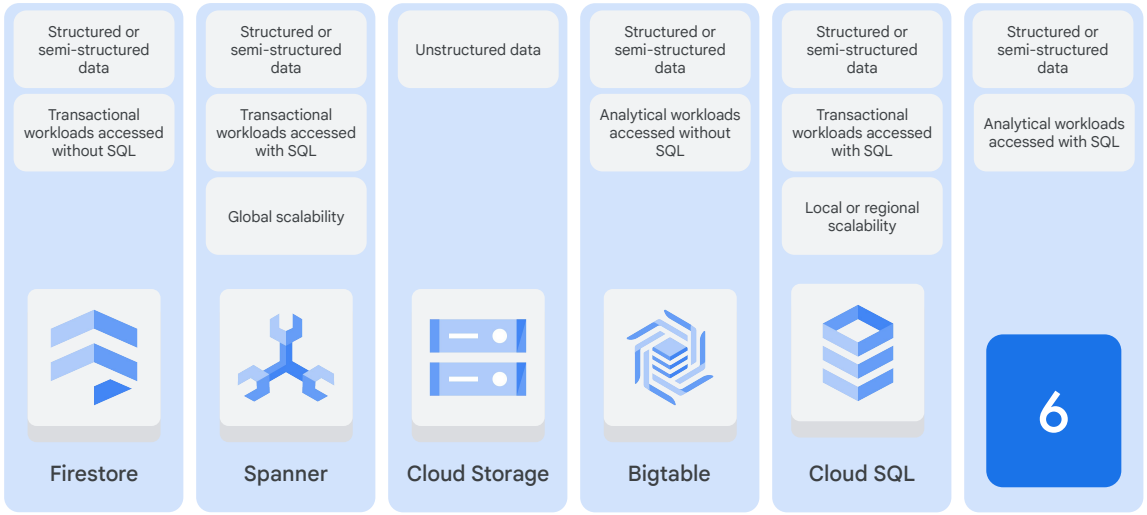


The correct answer is **Bigtable**. Now let's move on to the fifth column.

Which storage product can handle structured or semi-structured data, can access transactional workloads with SQL, and can scale locally or regionally?

Options include: Bigtable, Cloud Storage, Firestore, Spanner, BigQuery, or Cloud SQL.

Which storage product...?

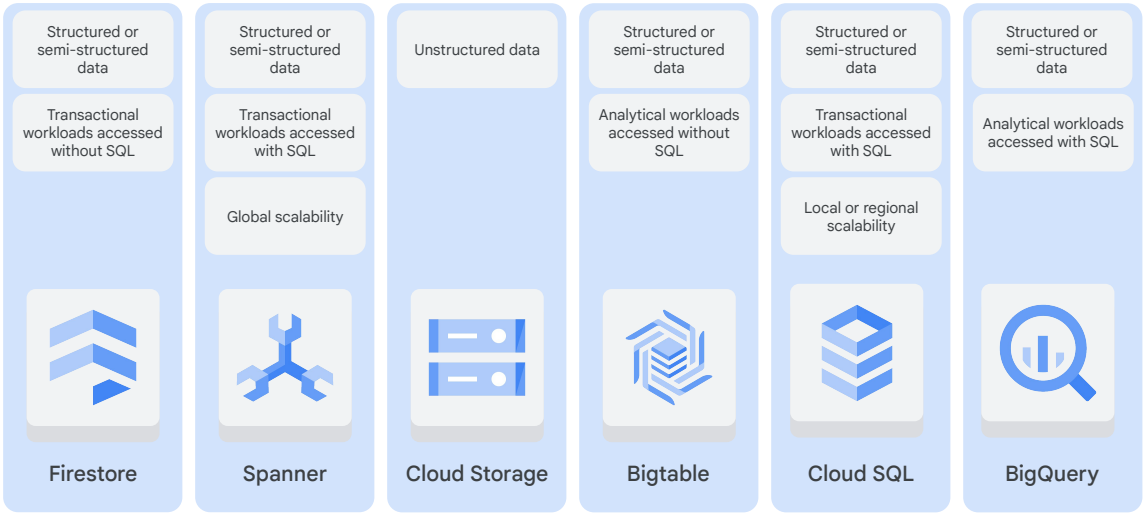


The correct answer is **Cloud SQL**. Now let's move on to the sixth final column.

Which storage product can handle structured or semi-structured data, can access analytical workloads with SQL?

Options include: Bigtable, Cloud Storage, Firestore, Spanner, BigQuery, or Cloud SQL.

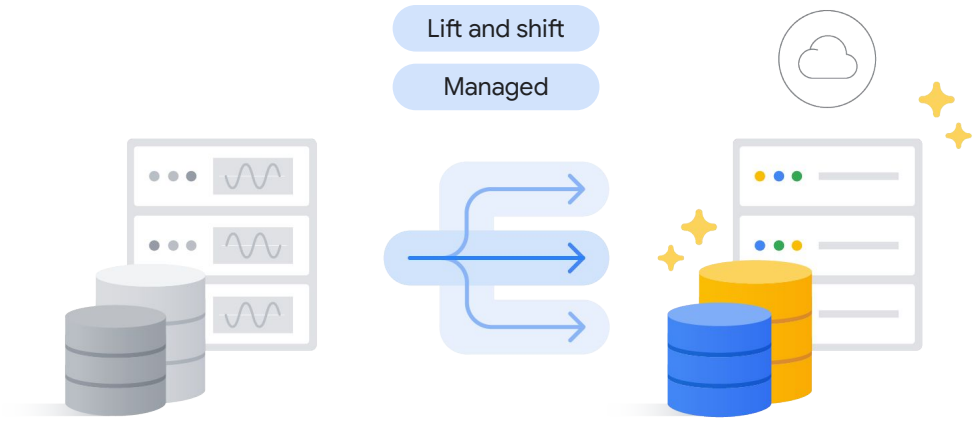
Which storage product...?



The correct answer is **BigQuery**.

Database migration and modernization

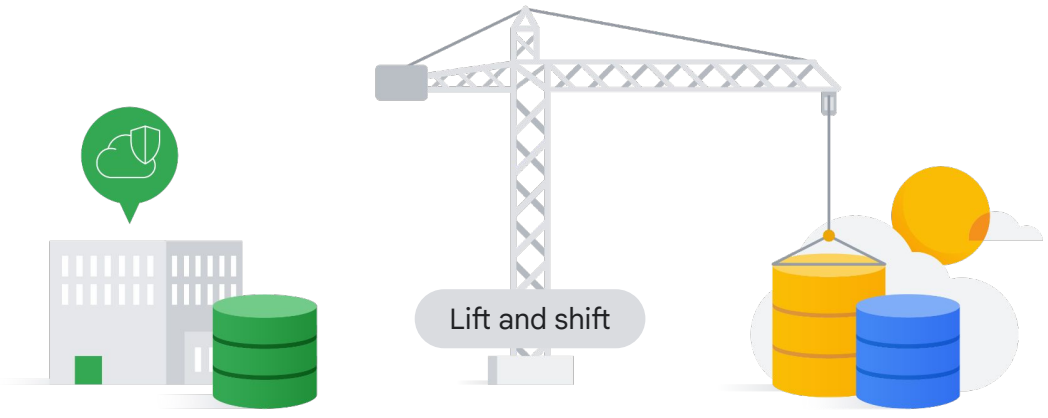
Ways to migrate or modernize databases in the cloud



Running modern applications on legacy, on-premises databases requires overcoming expensive, time-consuming challenges around latency, throughput, availability, and scaling.

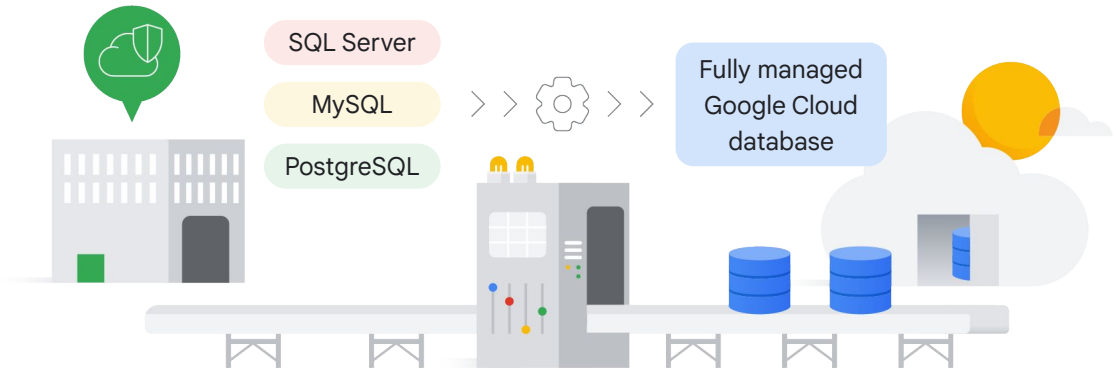
With database modernization, organizations can move data from traditional databases to fully managed or modern databases with relative ease. There are different ways that an organization can migrate or modernize their current database in the cloud.

Lift and shift databases



The most straightforward method is a **lift and shift** platform migration. This is where databases are migrated from on-premises and private cloud environments to the same type of database hosted by a public cloud provider, such as Google Cloud.

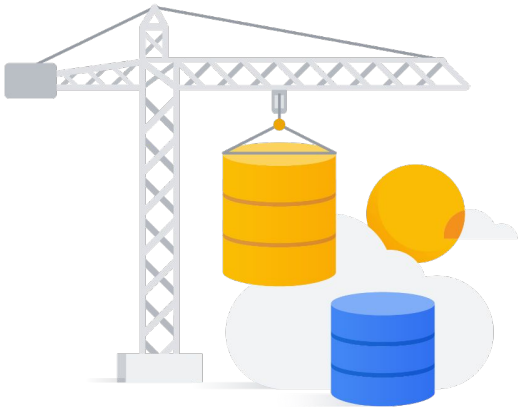
Managed database migration



Alternatively, a **managed database migration** allows the migration of databases from SQL Server, MySQL, PostgreSQL, and others to a fully managed Google Cloud database.

Although this migration requires careful planning and might cause slight upheaval, a fully managed solution lets you focus on higher priority work that really adds value to your organization.

Lift and shift databases



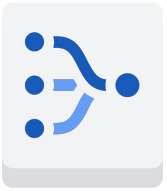
- More difficult to modernize databases
- Minimal upheaval
- Managed by the cloud provider

Although this solution makes the database more difficult to modernize, it does bring with it the benefits of minimal upheaval, and having data and infrastructure managed by the cloud provider.

Managed database migration services

Database Migration Service (DMS)

Easily migrate your databases to Google Cloud.



Datastream

It's used to synchronize data across databases, storage systems, and applications.

Google Cloud's **Database Migration Service (DMS)** can easily migrate your databases to Google Cloud, or **Datastream** can be used to synchronize data across databases, storage systems, and applications.

Module 2

Exploring Data Transformation with Google Cloud

Lessons

01 The value of data

02 Google Cloud data management solutions

03 Making data useful and accessible

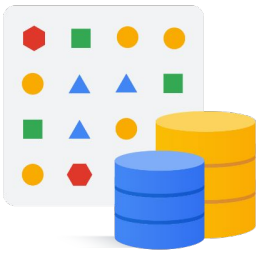
It's not always easy for organizations to make smart business decisions based on the data they've collected or produced. And too often there can be blockers in place that make analyzing it difficult for part, or all, of a workforce. With Google Cloud, that doesn't need to be the case.

In this section of this course, you'll explore:

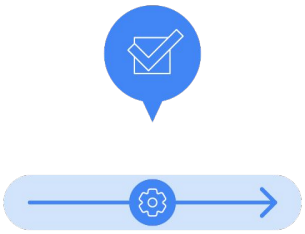
- How Looker makes it easy for a workforce to access the data they need, when they need it.
- How streaming analytics in real time can make data more useful.
- And two Google Cloud products that modernize data pipelines: Pub/Sub and Dataflow.

Business intelligence and insights using Looker

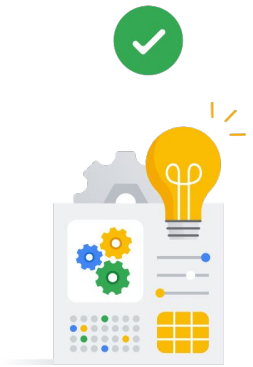
Finding the right business intelligence (BI) solution



Database



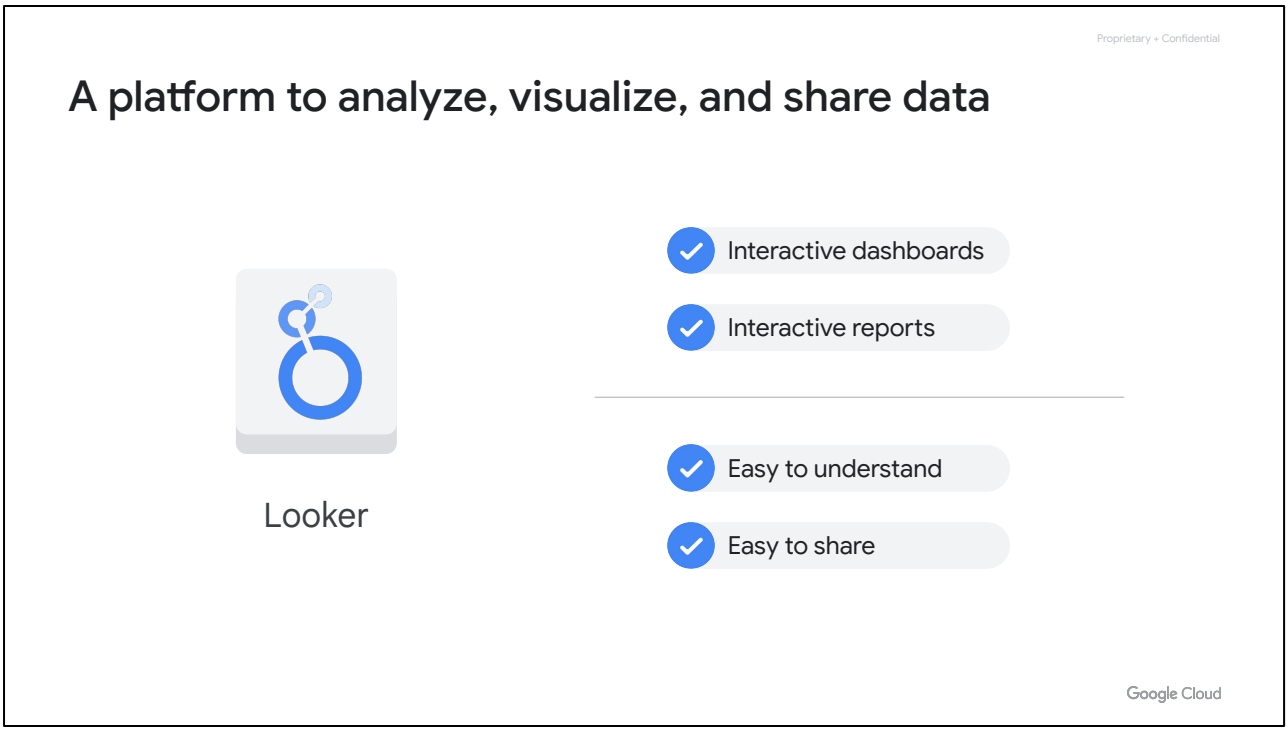
Effort and expertise might still be required to uncover insights.



It can be hard to identify the right BI solution.

When data is in a database, a fair amount of effort and expertise might still be required to uncover insights. This goal can be achieved by using a business intelligence solution. However, the challenge that organizations often face is identifying the *right* business intelligence solution.

Some solutions are too complex and not accessible by those outside the data engineering or data analysis teams. This means other teams have to put in requests and wait for answers, which defeats the purpose of gaining real-time insights. Other solutions let *everyone* in the business perform their own data analysis, but they can only perform their analysis with a selection of the available data. This means that only a few people, or possibly no one, has a full view of the organization's business data.



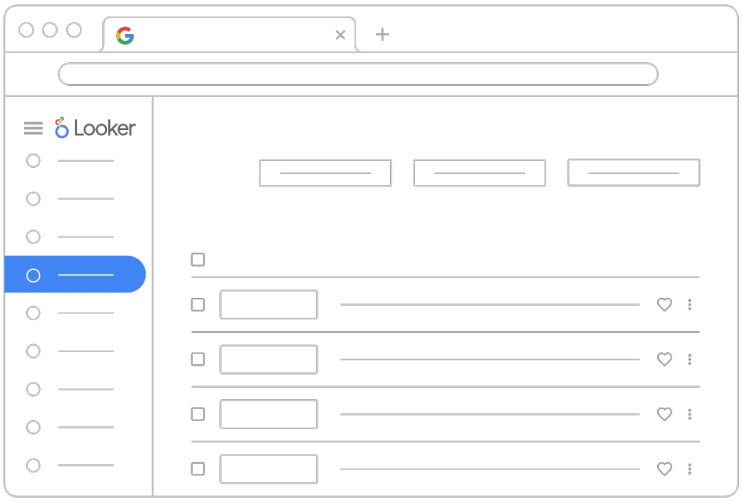
Looker is a Google Cloud business intelligence platform designed to help individuals and teams analyze, visualize, and share data. This includes creating interactive dashboards and reports that are easy to understand and share.

By having a reliable authority for business data, anyone on a team can explore it, ask and answer their own questions, and create visualizations. This approach empowers organizations to not just uncover insights, but also act on them.

Looker supports BigQuery and 60+ SQL databases

Looker supports BigQuery, along with more than 60 different SQL databases. Together, BigQuery and Looker provide rich, interactive dashboards and reports without compromising performance, scale, security, or data freshness.

Looker is 100% web-based



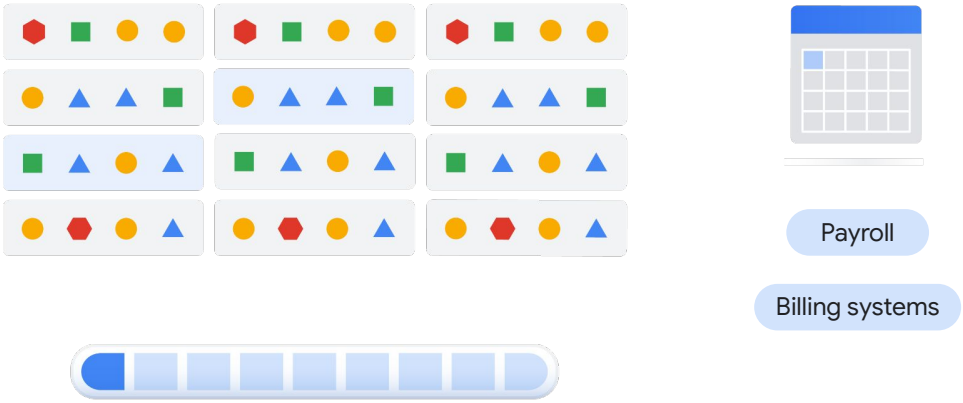
Easy to integrate into existing workflows

Easy to share with teams

Looker is also 100% web-based, which makes it easy to integrate into existing workflows and share with multiple teams at an organization.

Streaming analytics

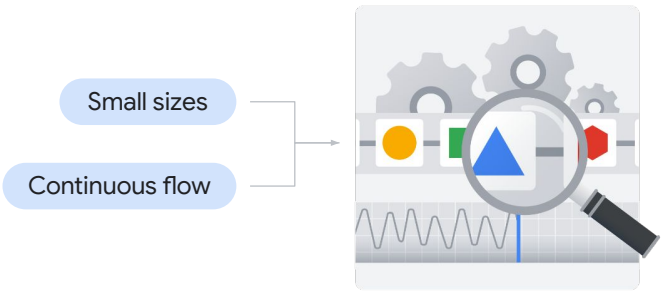
Data traditionally is moved in batches



Data traditionally is moved in batches. Batch processing often processes large volumes of data at the same time, with long periods of latency. An example is payroll and billing systems that have to be processed on either a weekly or monthly basis.

Although this approach can be efficient to handle large volumes of data, it doesn't work with time-sensitive data that's meant to be streamed, because that data can be stale by the time it's processed.

Streaming analytics

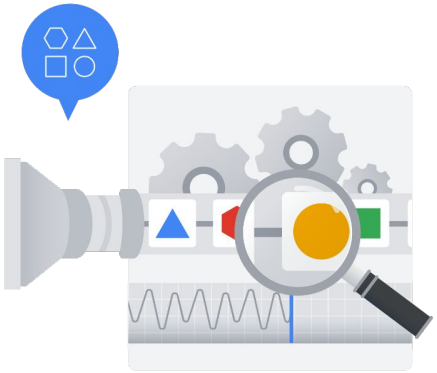


Streaming analytics is the processing and analyzing of data records continuously instead of in batches.

Streaming analytics is the processing *and* analyzing of data records **continuously** instead of in batches. Generally, streaming analytics is useful for the types of data sources that send data in small sizes, often in kilobytes, in a continuous flow as the data is generated.

This results in the analysis and reporting of events as they happen.

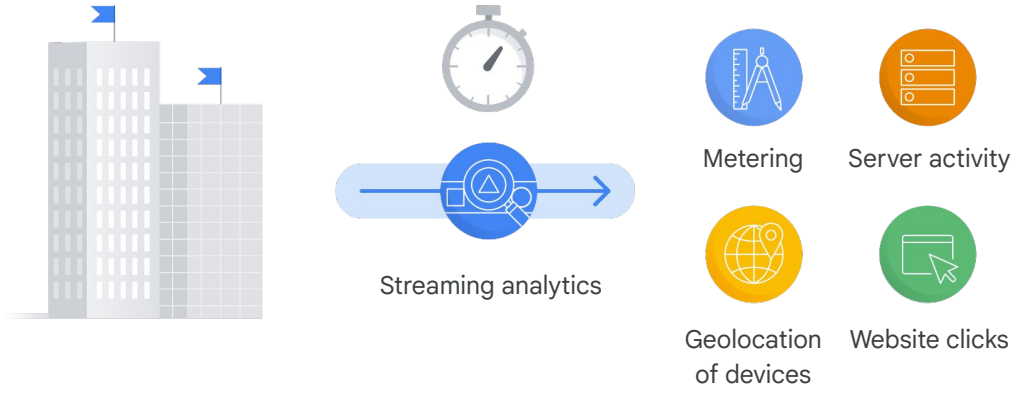
Sources of streaming data



- Equipment sensors
- Clickstreams
- Social media feeds
- Stock market quotes
- App activity

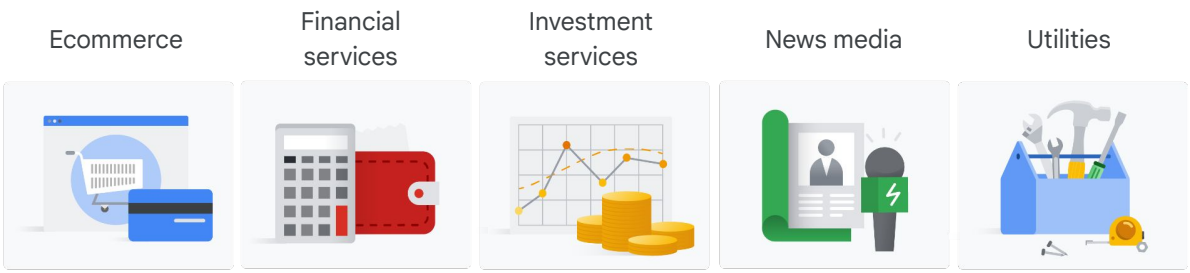
Sources of streaming data include equipment sensors, clickstreams, social media feeds, stock market quotes, app activity, and more.

Companies use streaming analytics to analyze data in real time and provide insights



Companies use streaming analytics to analyze data in real time and provide insights into a wide range of activities, such as metering, server activity, geolocation of devices, or website clicks.

Streaming analytics use cases

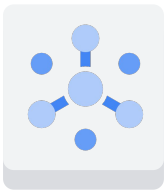


Use cases include:

- **Ecommerce**: user clickstreams can be analyzed to optimize the shopping experience with real-time pricing, promotions, and inventory management.
- **Financial services**: account activity can be analyzed to detect abnormal behavior in the data stream and generate a security alert.
- **Investment services**: market changes can be tracked and settings adjusted to customer portfolios based on configured constraints, such as selling when a certain stock value is reached.
- **News media**: user click records can be streamed from various news source platforms and the data can then be enriched with demographic information to better serve articles that are relevant to the targeted audience.
- **Utilities**: throughput across a power grid can be monitored and alerts generated or workflows initiated when established thresholds are reached.

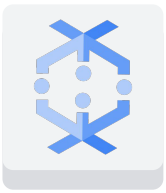
Google Cloud's streaming analytics products

Pub/Sub



Ingests hundreds of millions of events per second.

Dataflow

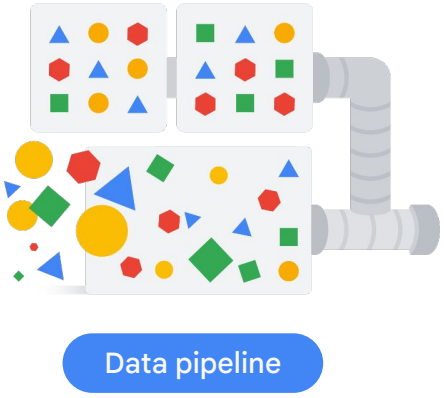


Unifies streaming and batch data analysis and builds cohesive data pipelines.

Google Cloud offers two main streaming analytics products to ingest, process, and analyze event streams in real time, which makes data more useful and accessible from the instant it's generated.

Pub/Sub ingests hundreds of millions of events per second, but **Dataflow** unifies streaming and batch data analysis and builds cohesive data pipelines.

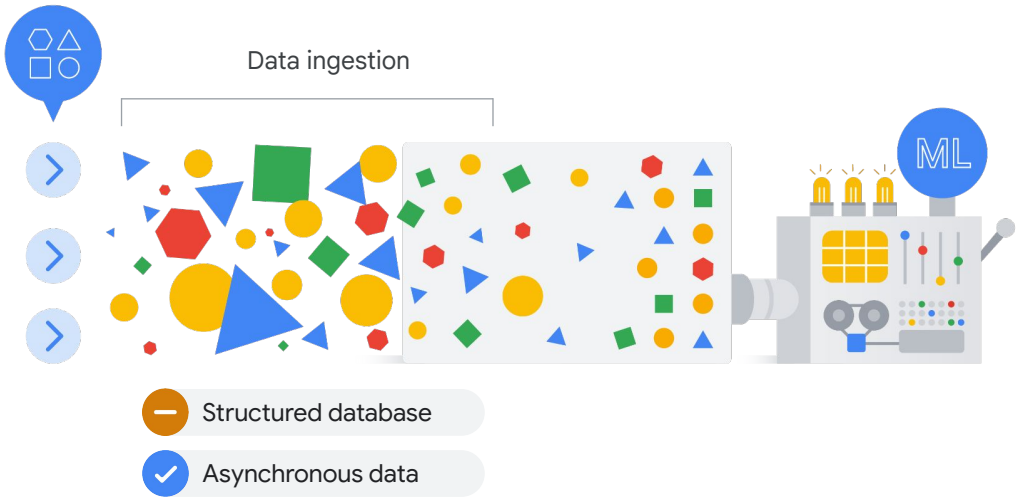
What is a data pipeline?



A series of actions, or stages, that ingest raw data from different sources and then move that data to a destination for storage and analysis.

A data pipeline represents a series of actions, or stages, that ingest raw data from different sources and then move that data to a destination for storage and analysis.

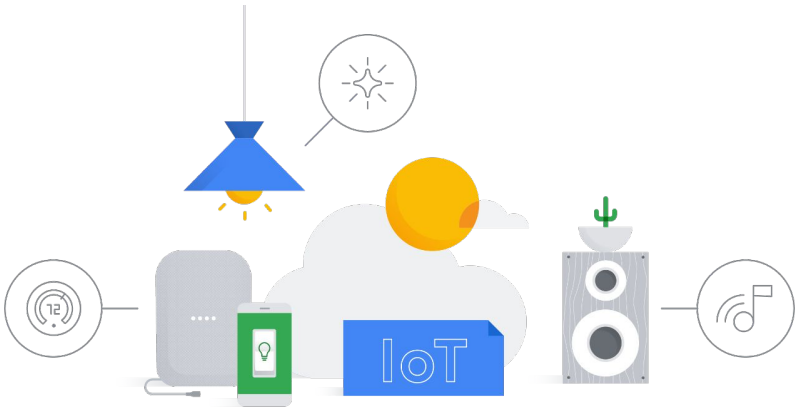
A data pipeline



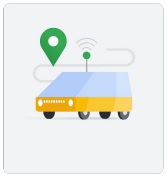
One of the early stages in a data pipeline is data ingestion, which is where large amounts of streaming data are received.

Data, however, might not always come from a single, structured database. Instead, the data might stream from a thousand, or even a million, different events that are all happening asynchronously.

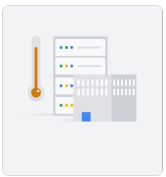
Asynchronous data



Internet of Things applications



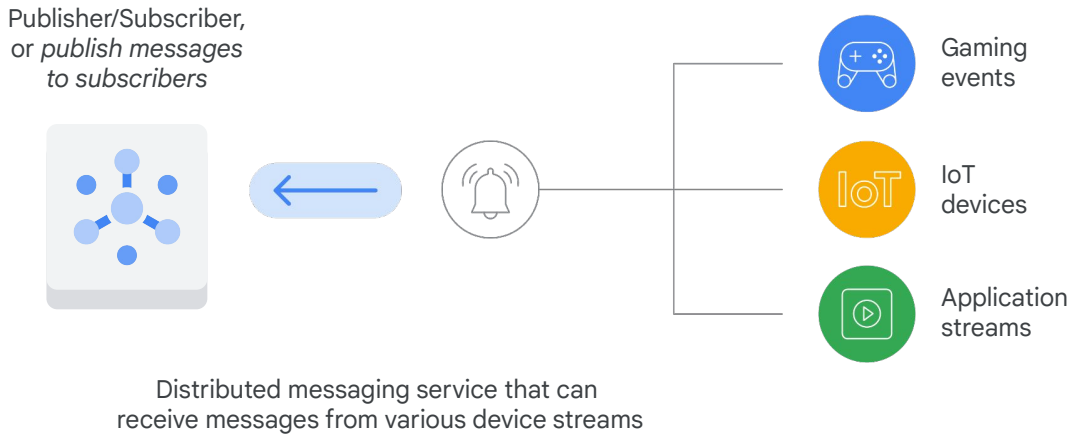
Location sensor data



Temperature sensor data

A common example of this is data from IoT, or Internet of Things, applications. These can include sensors on taxis that send out location data every 30 seconds, or temperature sensors around a data center to help optimize heating and cooling.

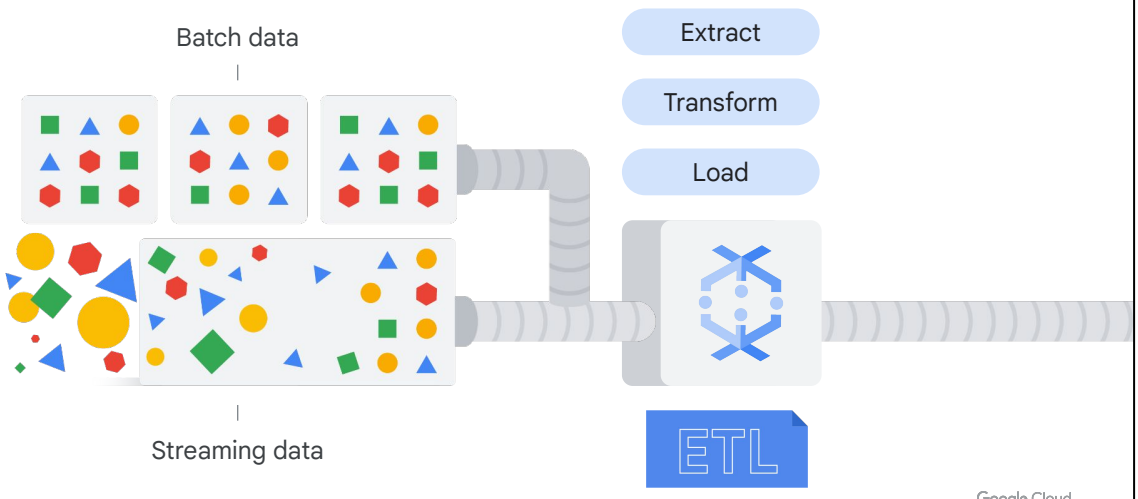
Pub/Sub



Pub/Sub is a distributed messaging service that can receive messages from various device streams such as gaming events, IoT devices, and application streams.

The name is short for Publisher/Subscriber, or *publish messages to subscribers*.

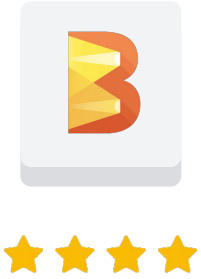
Dataflow pipes data captured by Pub/Sub into a data warehouse



After messages have been captured from the streaming input sources, you need a way to pipe that data into a data warehouse for analysis. This is where Dataflow comes in.

Dataflow creates a pipeline to process both streaming data and batch data. “Process” in this case refers to the steps to extract, transform, and load data, sometimes referred to as **ETL**.

Apache Beam is a popular solution for pipeline design

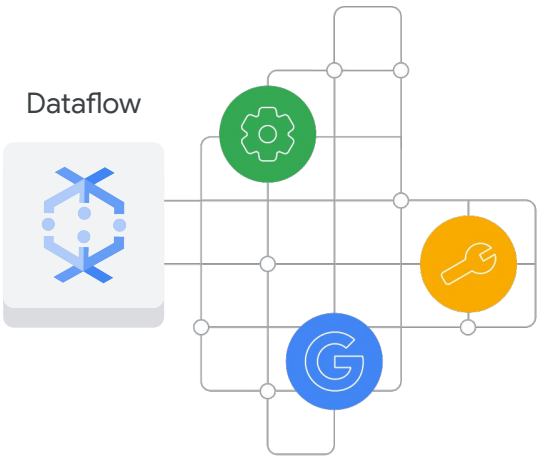


An open source, unified programming model to define and execute data processing pipelines, including ETL, batch, and stream processing

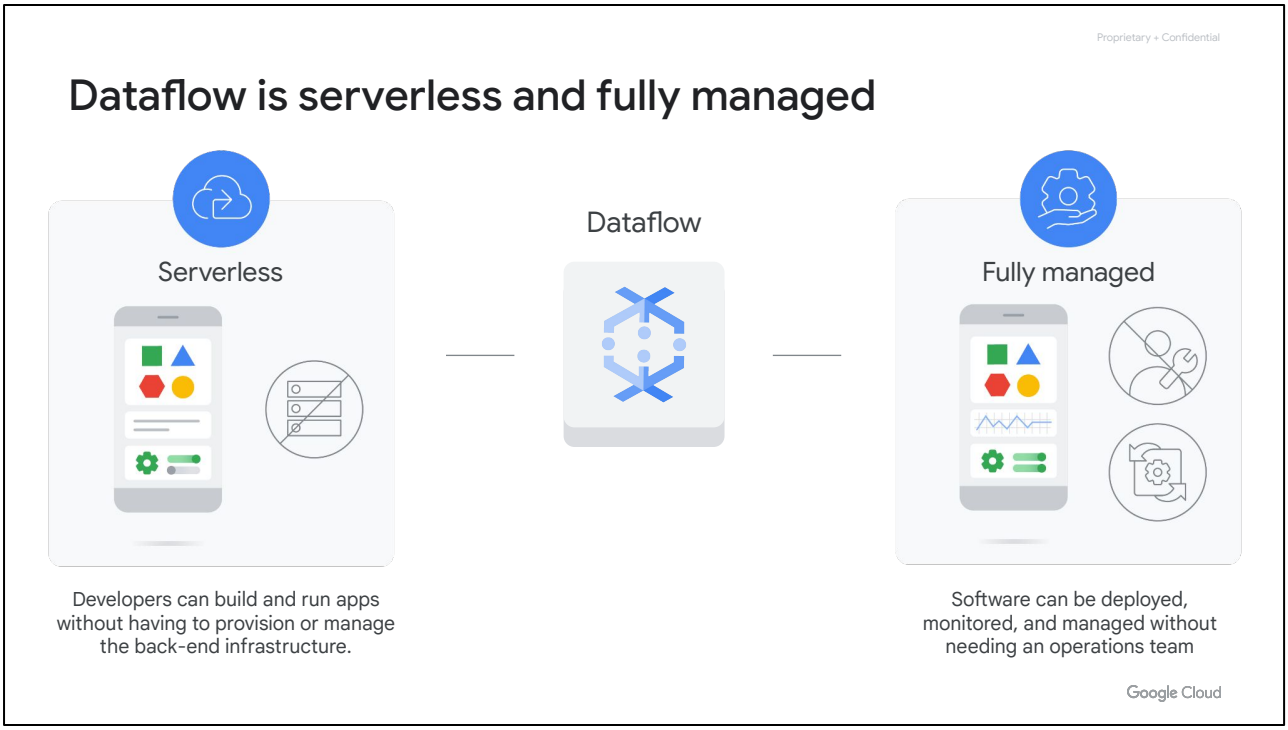
A popular solution for pipeline design is Apache Beam. It's an open source, unified programming model to define and execute data processing pipelines, including ETL, batch, and stream processing.

Dataflow

- Handles infrastructure setup
- Handles maintenance
- Is built on Google infrastructure
- Allows for reliable auto scaling
- Meets data pipeline demands



Dataflow handles much of the complexity for infrastructure setup and maintenance, and is built on Google's infrastructure. This product allows for reliable auto scaling to meet data pipeline demands.



Dataflow is serverless and full managed.

- Serverless computing means that software developers can build and run applications without having to provision or manage the back-end infrastructure. For example, Google Cloud manages infrastructure tasks on behalf of the users, like resource provisioning, performance tuning, and ensuring pipeline reliability.
- And a fully managed environment is one where software can be deployed, monitored, and managed without needing an operations team. You can create this environment by using automation tools and technologies.

Using a serverless and fully managed solution like Dataflow means that you can spend more time analyzing the insights from your datasets and less time provisioning resources to ensure that your pipeline will successfully complete its next cycles.

Quiz

Question

New cloud tools make it possible to harness the potential of unstructured data. Which of these use cases best demonstrates this?

A. Creating visualizations from seasonal weather data
B. Analyzing historical sales figures to predict future trends
C. Analyzing social media posts to identify sentiment toward a brand
D. Using GPS coordinates to power a ride-sharing app

New cloud tools make it possible to harness the potential of unstructured data. Which of these use cases best demonstrates this?

A. Creating visualizations from seasonal weather data
B. Analyzing historical sales figures to predict future trends
C. Analyzing social media posts to identify sentiment toward a brand
D. Using GPS coordinates to power a ride-sharing app

Quiz

Answer

New cloud tools make it possible to harness the potential of unstructured data. Which of these use cases best demonstrates this?

A. Creating visualizations from seasonal weather data

B. Analyzing historical sales figures to predict future trends

C. Analyzing social media posts to identify sentiment toward a brand 

D. Using GPS coordinates to power a ride-sharing app

The correct answer is C.

A. Creating visualizations from seasonal weather data
   - Why this is the **incorrect** answer: While weather data can involve complex models and large datasets, it typically fits within a structured or semi-structured format with standardized units of measurement.

B. Analyzing historical sales figures to predict future trends
   - Why this is the **incorrect** answer: Sales data often follows traditional structured formats – figures, numbers, dates, making it well-suited for established analysis techniques, even without modern cloud tools.

C. Analyzing social media posts to identify sentiment toward a brand
   - Why this is the **correct** answer: This use case highlights the ability to tackle unstructured data. Social media posts contain a mix of text, images, emojis, and more – a classic example of unstructured data. Sentiment analysis tools powered by AI and natural language processing in the cloud are key for interpreting this complex data source.

D. Using GPS coordinates to power a ride-sharing app
   - Why this is the **incorrect** answer: Geospatial (GPS) data typically has predictable structure (latitude, longitude). Advanced analysis can certainly improve such apps, but structuring this data doesn't rely on tools designed specifically for unstructured data.

Quiz

Question

Data in the form of video, pictures, and audio recordings is well suited to object storage. Which product is best for storing this kind of data?

A. Firestore

B. Cloud Storage

C. Cloud SQL

D. BigQuery

Data in the form of video, pictures, and audio recordings is well suited to object storage. Which product is best for storing this kind of data?

A. Firestore
B. Cloud Storage
C. Cloud SQL
D. BigQuery

Quiz

Answer

Data in the form of video, pictures, and audio recordings is well suited to object storage. Which product is best for storing this kind of data?

A. Firestore

B. Cloud Storage 

C. Cloud SQL

D. BigQuery

The correct answer is B.

A. Firestore
   - Why this is the **incorrect** answer: Firestore is a NoSQL document database primarily used to store application data in a structured format. It's not optimal for storing large media files.

B. Cloud Storage
   - Why this is the **correct** answer: Cloud Storage is the ideal choice for storing objects like videos, pictures, and audio recordings because it's fundamentally designed to store files as a whole, making uploads and downloads of multimedia easy and efficient. It can also handle massive amounts of data, offering virtually unlimited storage for a growing media library, and supports content delivery with low latency.

C. Cloud SQL
   - Why this is the **incorrect** answer: Cloud SQL is a relational (SQL-based) database service aimed at structured data (think tables, rows, columns) in transactional workloads. It's not designed for storing and managing multimedia object files.

D. BigQuery
   - Why this is the **incorrect** answer: BigQuery is a data warehouse solution optimized for analyzing enormous datasets using SQL queries. While it could store metadata about multimedia files, it's not ideal for storing the large video, picture, and audio files themselves.