Google Cloud

Student Slides

Innovating with Data with Google Cloud

Modules

01 The Value of Data

02 Data Consolidation and Analytics

03 Innovation with Machine Learning

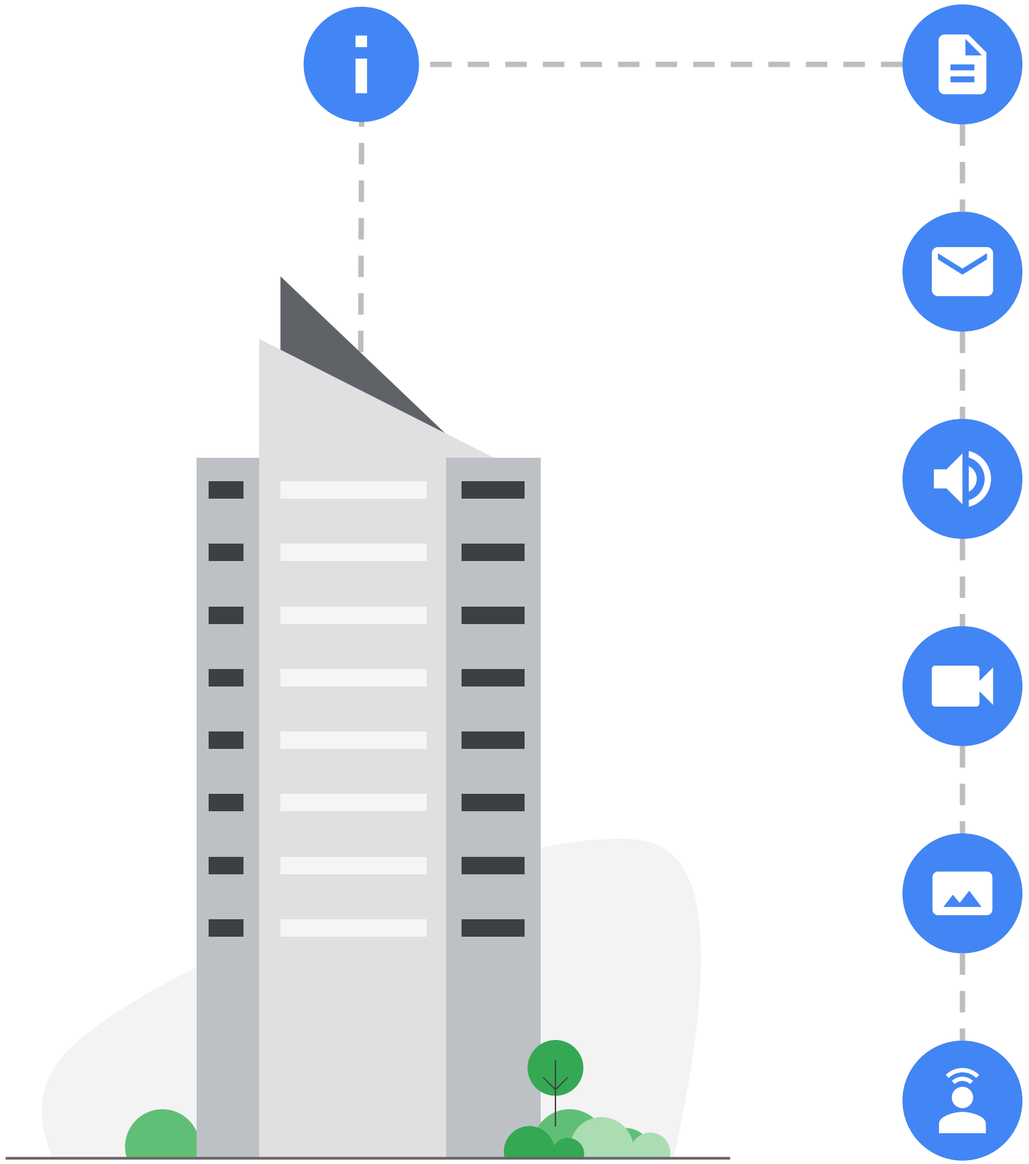
Google Cloud

Module 1: Student Slides

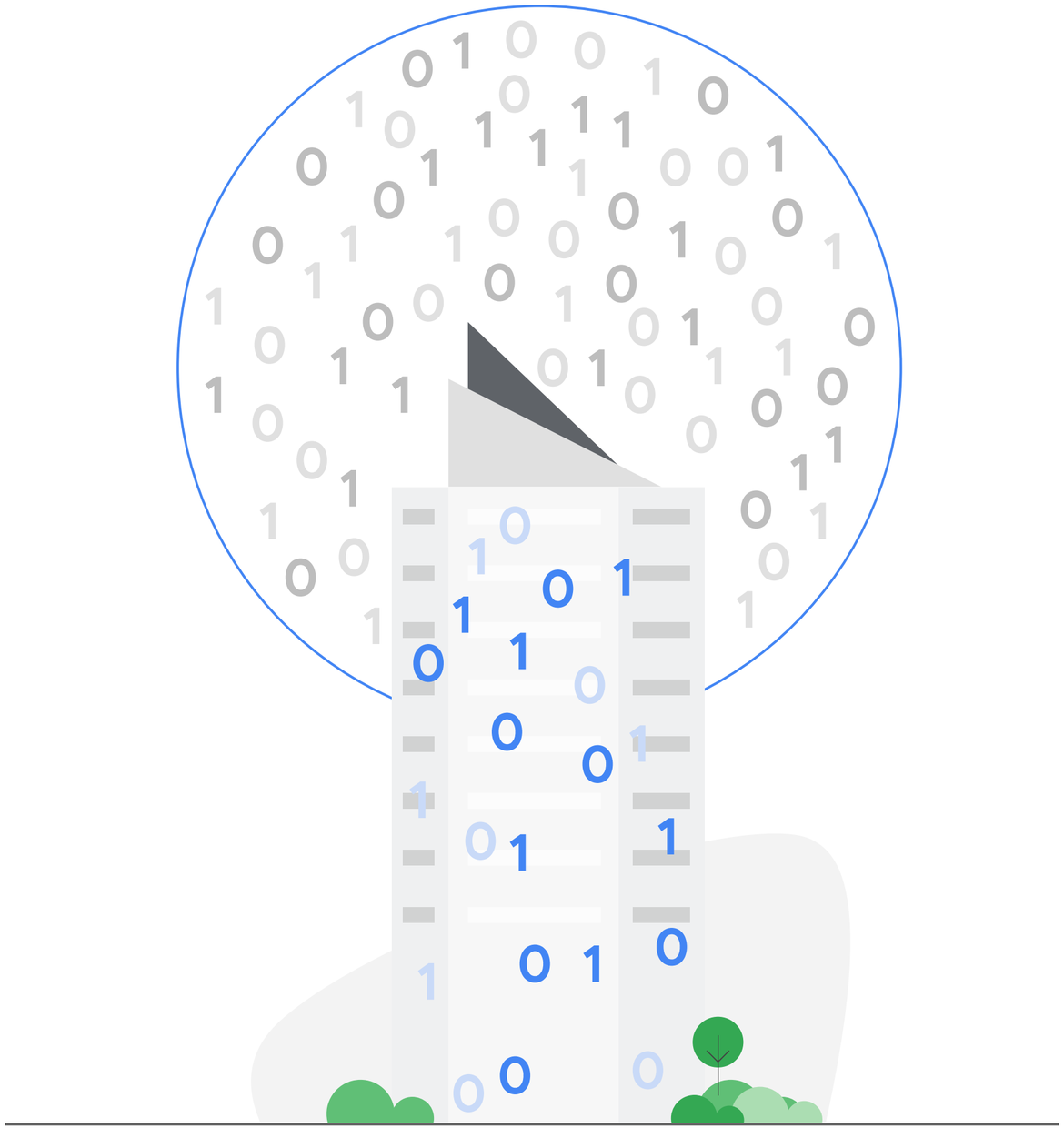
The Value of Data

Topics covered

- The role of data in digital transformation
- Leveraging data in your organization
- Types of data
- Important considerations for using data in the cloud



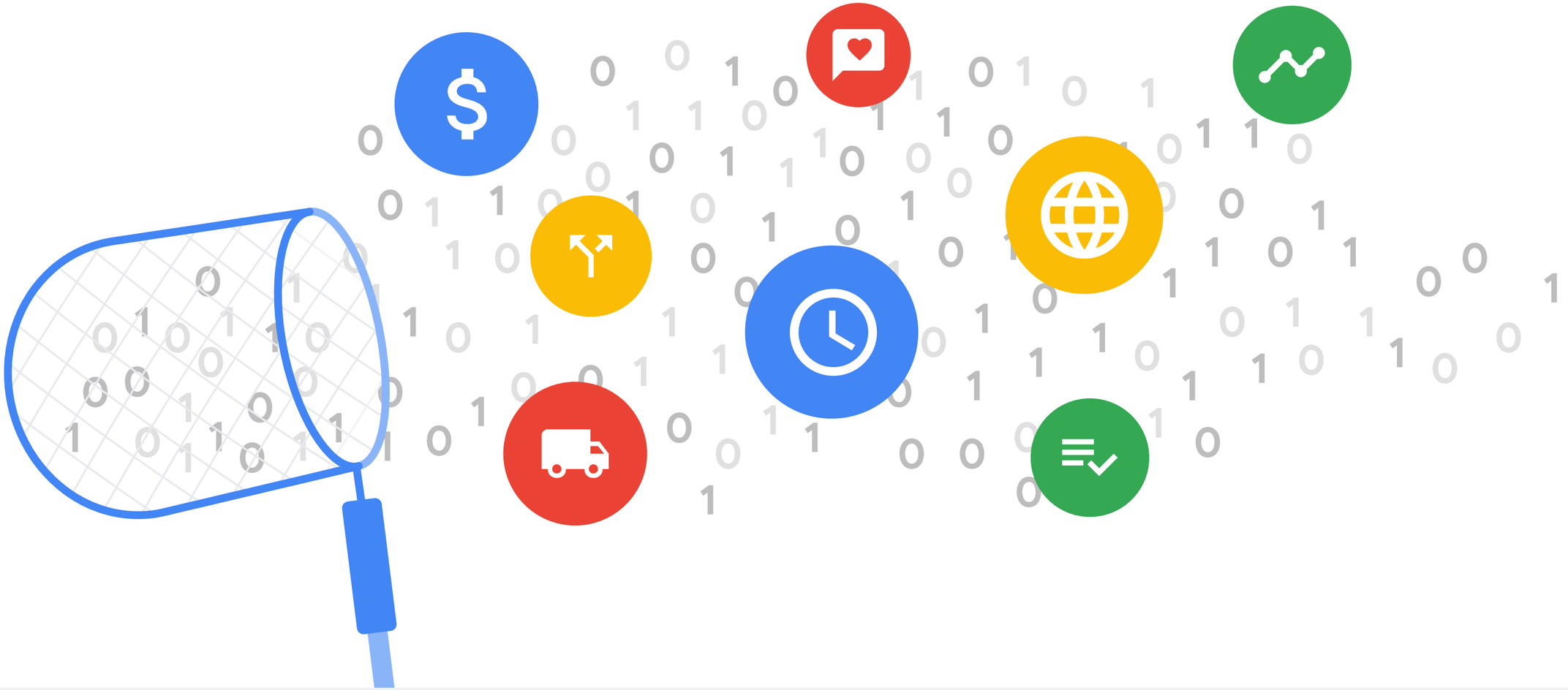
Data is any information that is useful to an organization. Examples include: documents, emails, audio files, video files, images, and even ideas in users' minds.



Businesses now have access to data like never before. This includes internal information (data from inside their organization) and external information (customer and industry data).

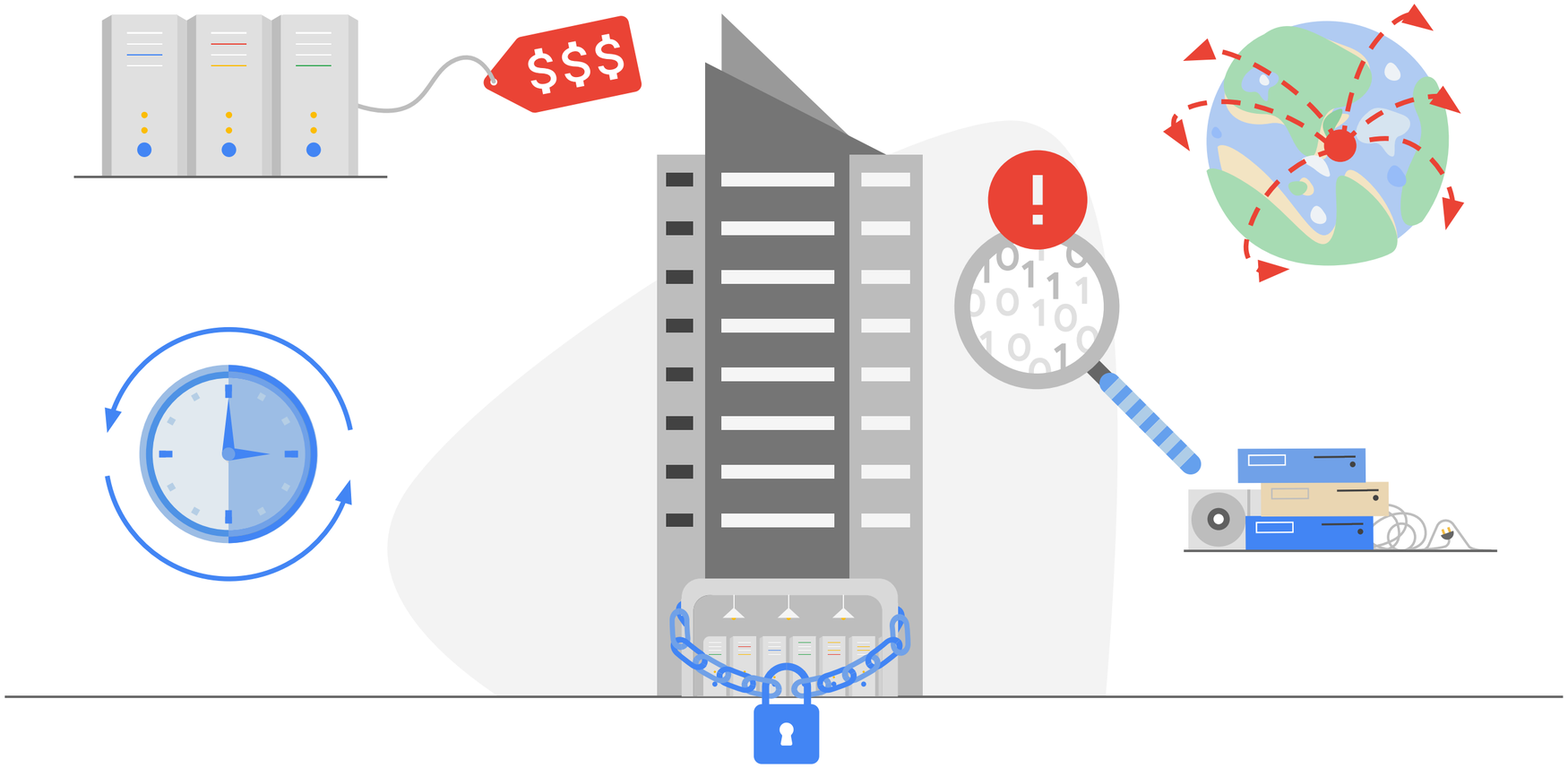
As organizations have digitized their operations, various forms of business data have become available. This includes financial information, logistics data, production output, quality reports, etc. Businesses also have access to new kinds of data about their customers.

The internet has also enabled access to external data, such as industry benchmarking reports.



Capturing and leveraging this data to unlock business value is central to digital transformation.

Large enterprises with traditional IT infrastructures face several limitations when it comes to leveraging the value of data. These limitations include: processing volumes and varieties of new data, finding cost effective solutions, scaling resource capacity up or down, accessing historical data, and deriving insights from historical and new data.



Google Cloud offers

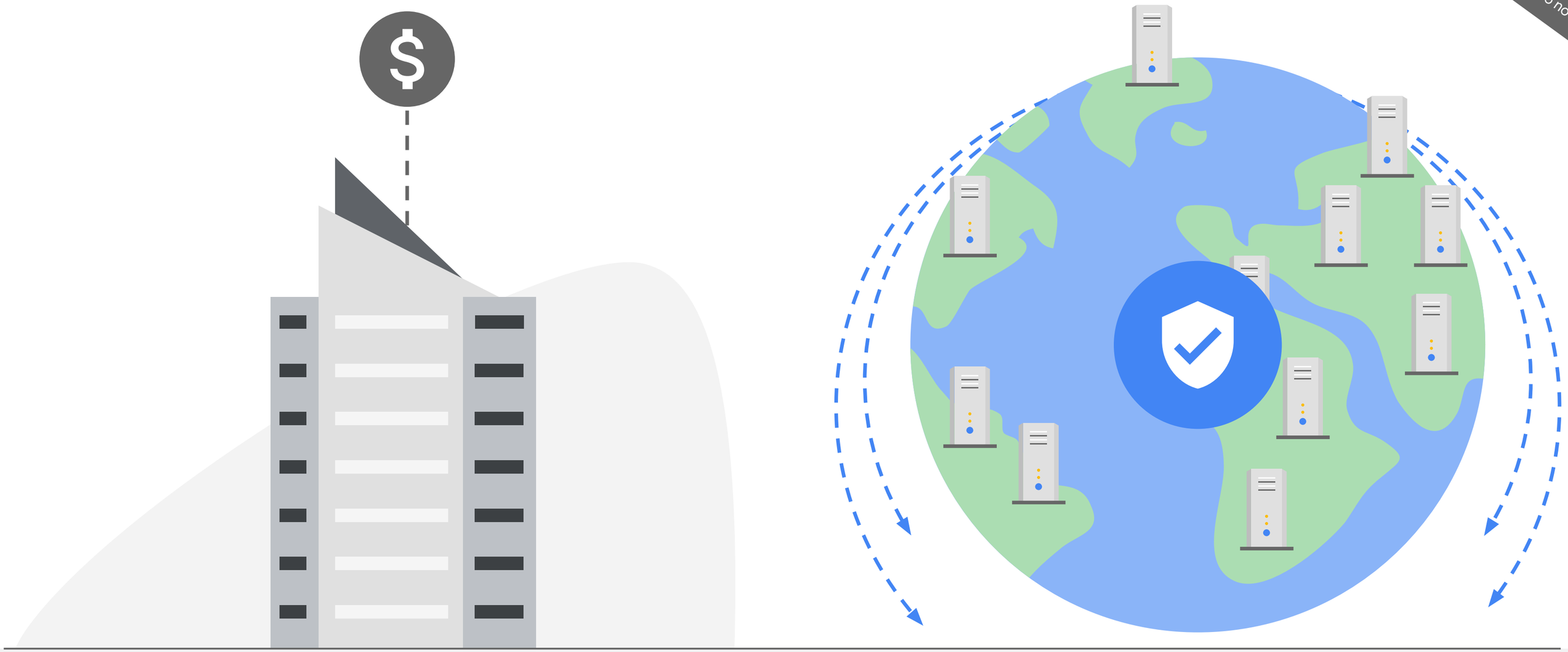
Economies of scale

Automation

Rapid elasticity

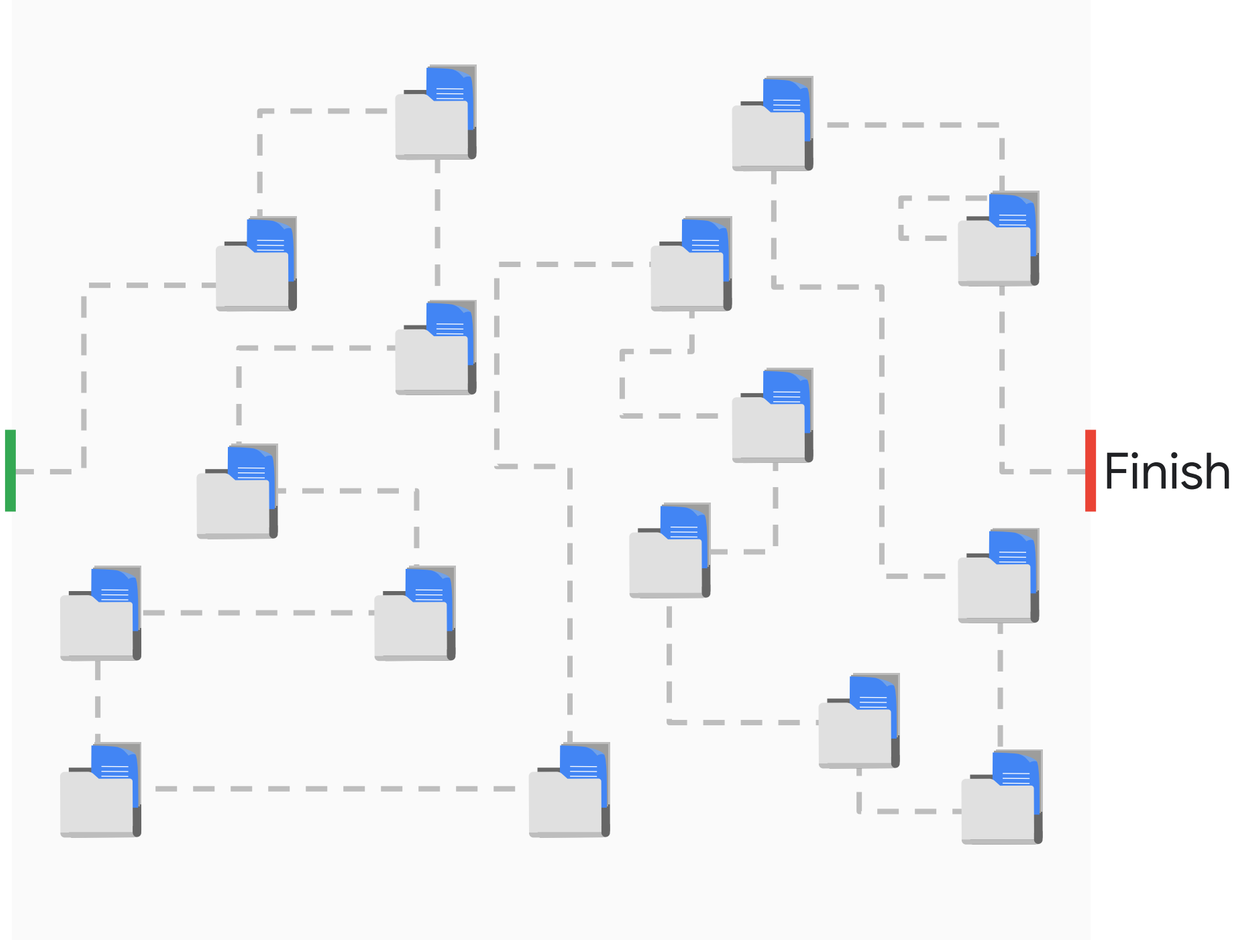
Data access

Businesses can now consume, store and process terabytes of data in real-time, and run queries—that is, requests to retrieve and use data, instantly.

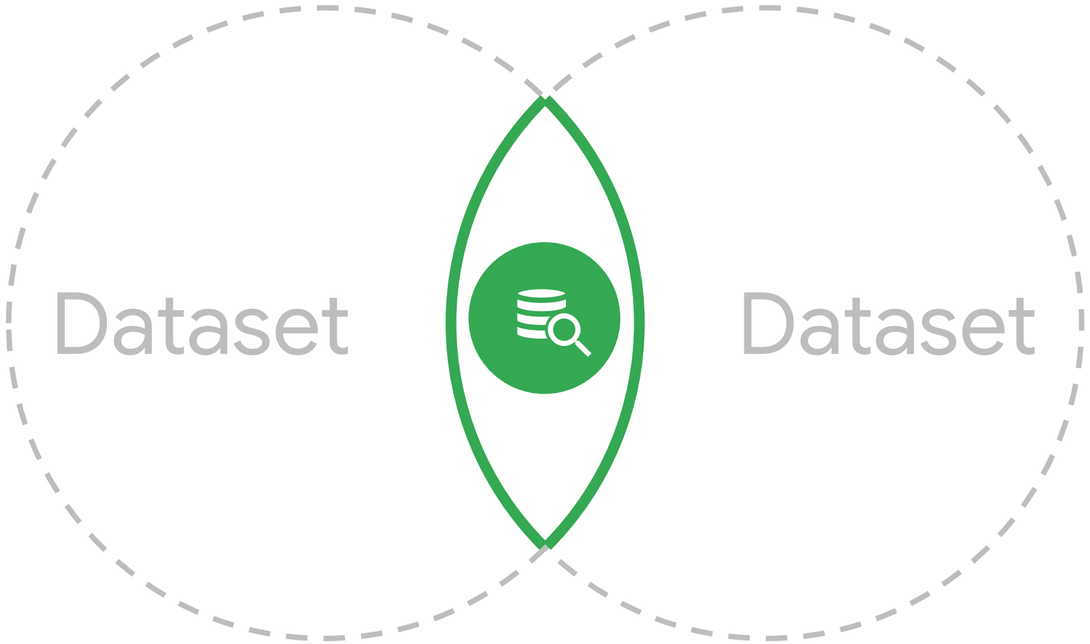


Resources are now distributed across a global network. Multiple data centers create resilience against data loss or service disruption, without any extra overhead for businesses.

And data can be combined, analyzed and served to business teams quickly and cost effectively.

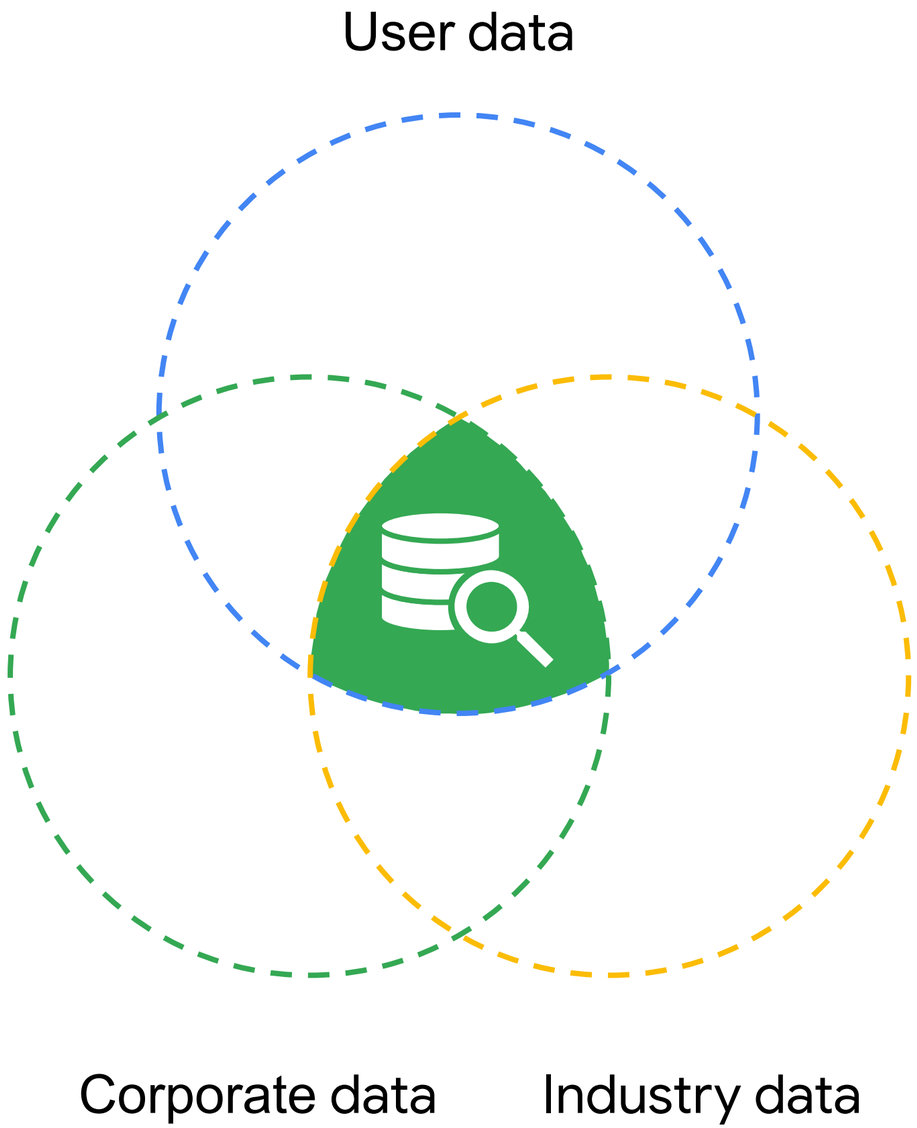


A “data map” is a chart of all the data used in end-to-end business processes.



How can you make your data actionable?

Take two or more datasets and ask yourself, “What insight could I gain if these datasets were combined?”



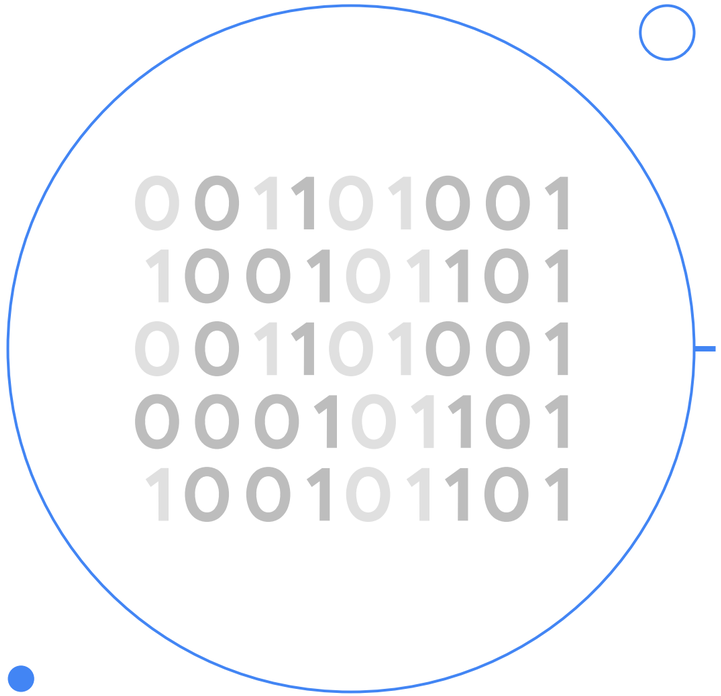
Think about what dataset you have in each category. Then, consider how different datasets can be combined to create valuable insights.

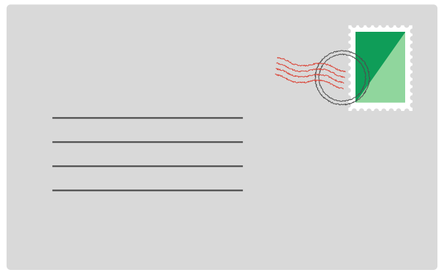
Structured

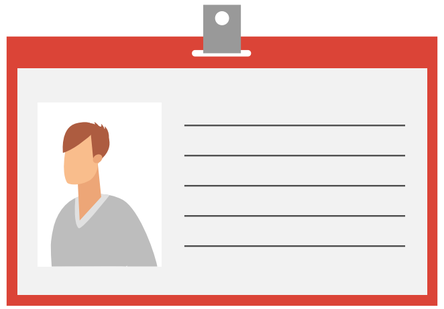
Unstructured

We can categorize data into two main types:

1. Structured
2. Unstructured



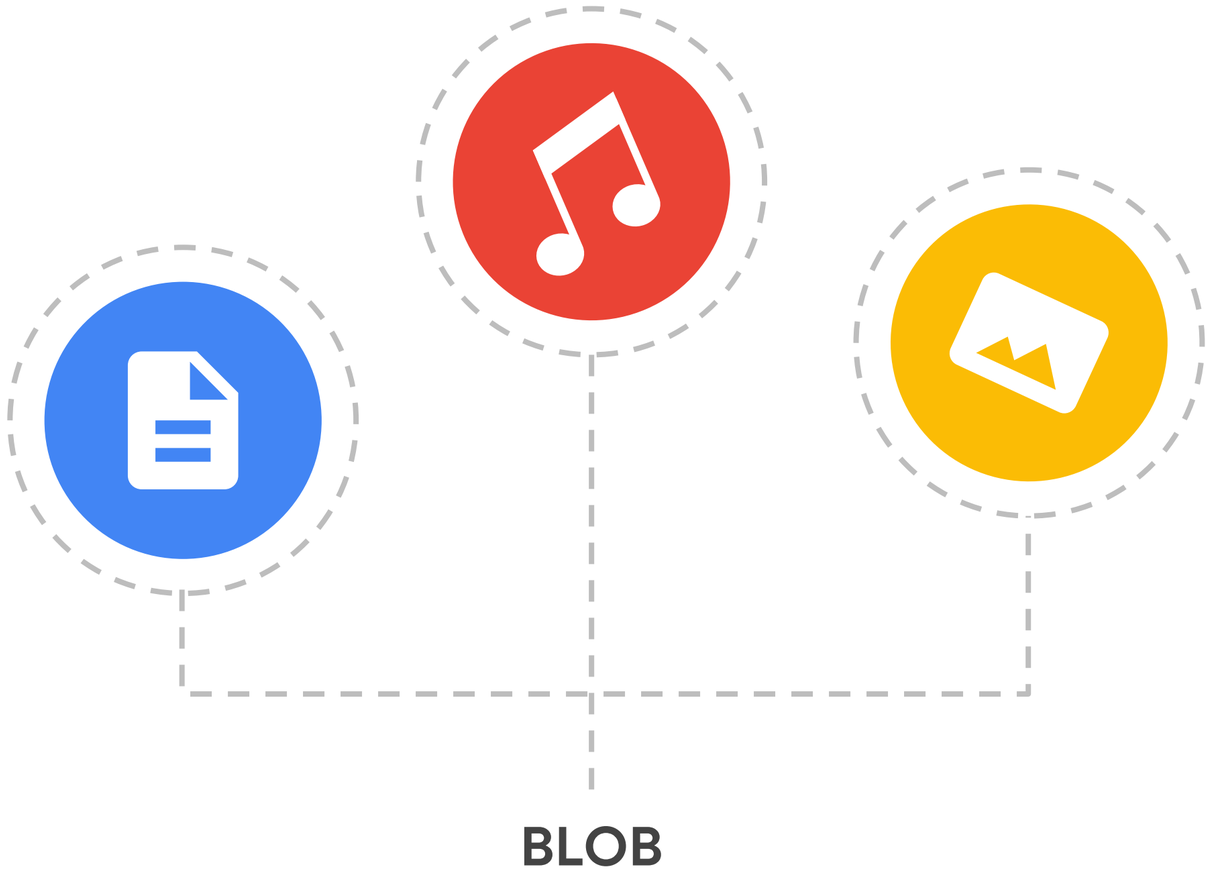


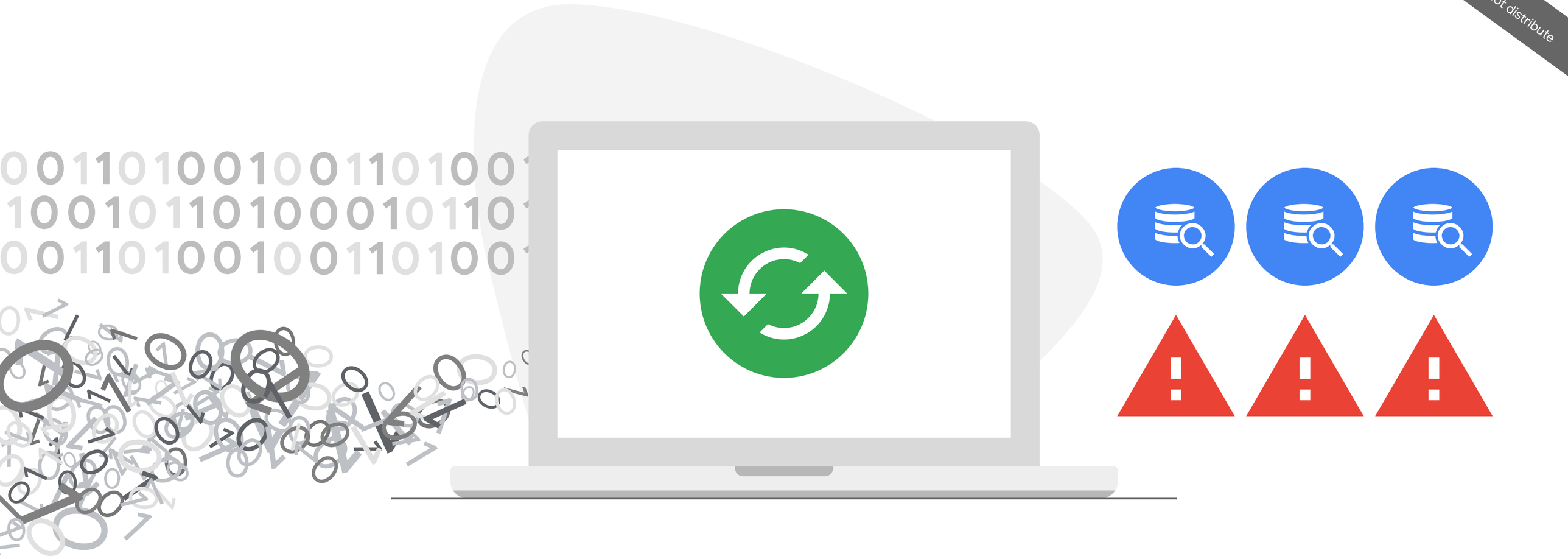


Structured data is highly organized. Examples include customer records consisting of names, addresses, credit card numbers, and other quantitative data. Structured data can be easily stored and managed in databases.

Unstructured data has no organization and tends to be qualitative. Examples of unstructured data include word processing documents, audio files, images, and videos.

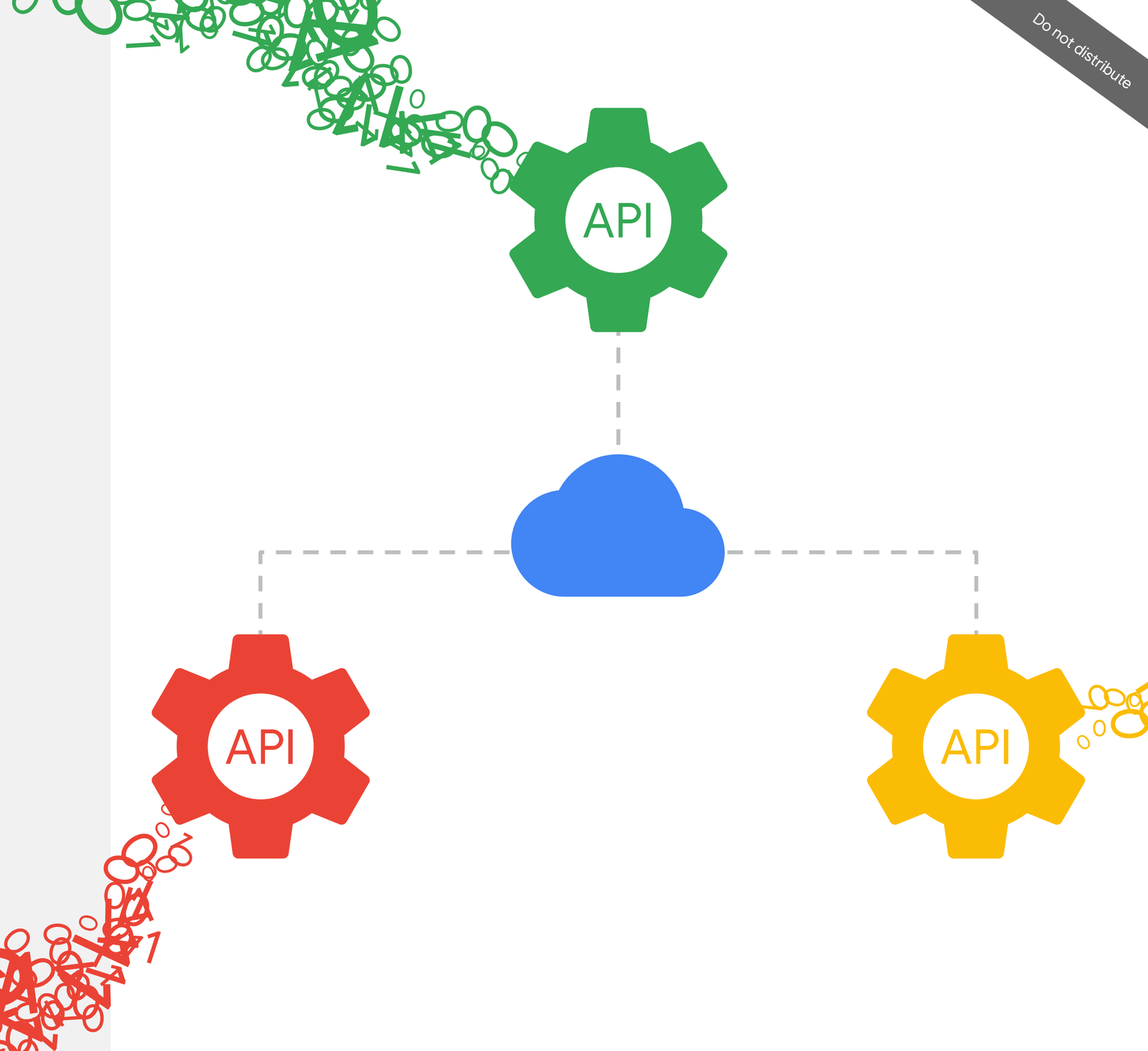
Some unstructured data can be stored in a format called a BLOB. This stands for **b**inary **l**arge **ob**ject. Images, audio, and multimedia files can all be stored as BLOBs.

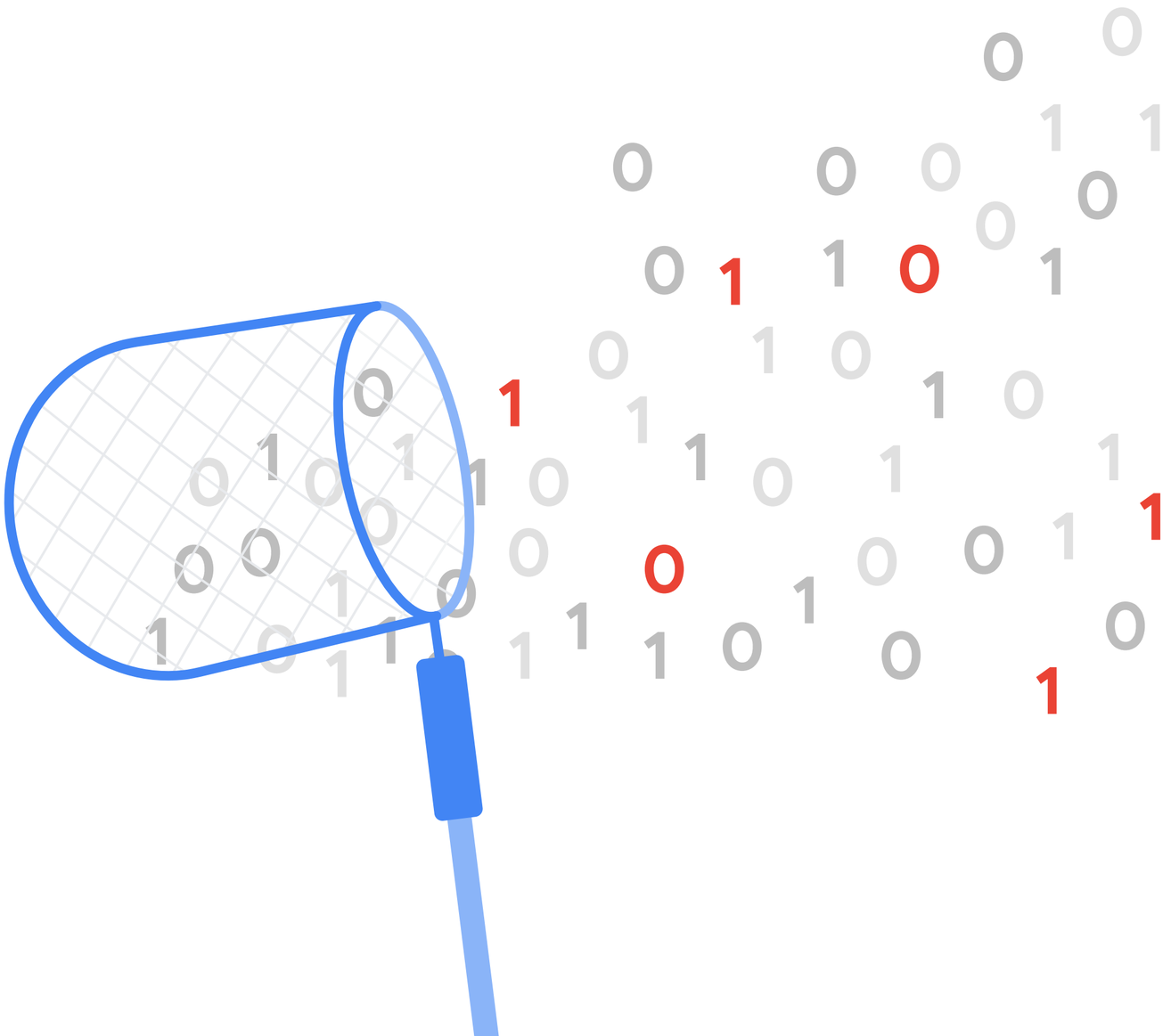




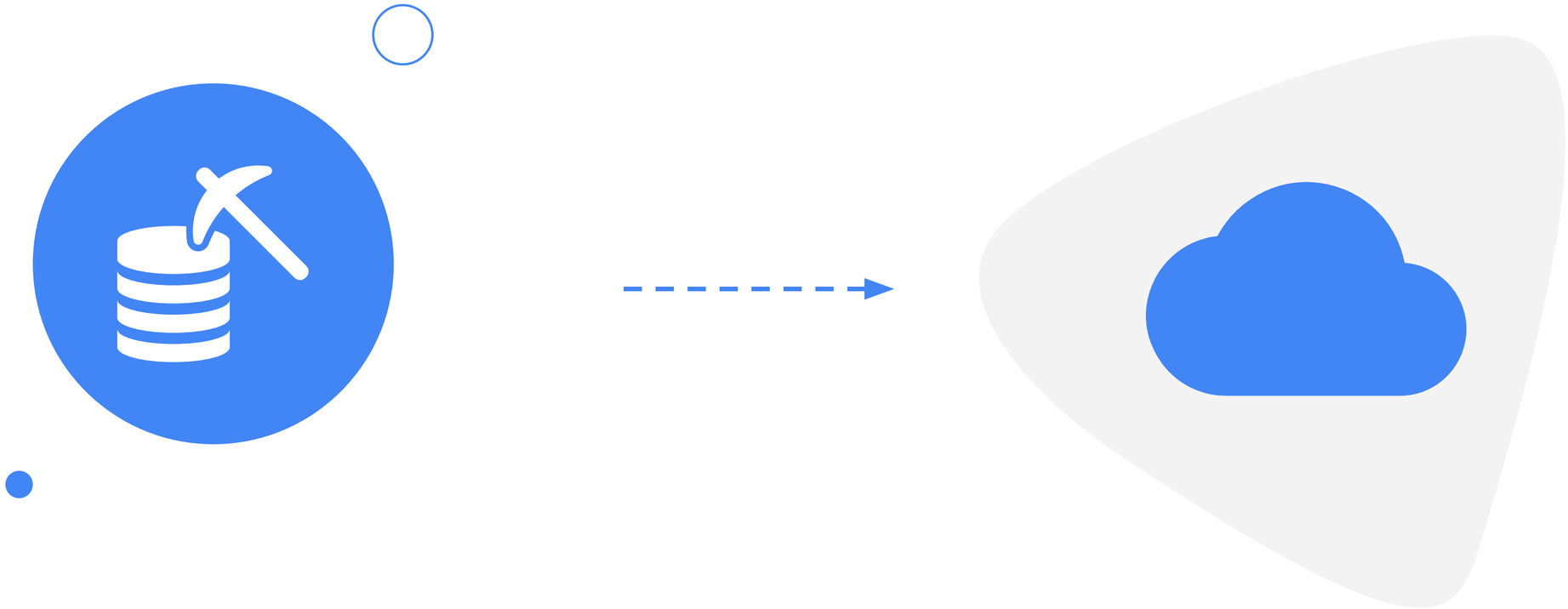
Organizations rely on both structured and unstructured data to gain insight and make intelligent decisions. However, unstructured data has historically been very difficult to analyze.

With the right cloud tools, businesses can extract value from unstructured data by using APIs to create structure. **APIs** are a set of functions that integrate different platforms, with different types of data, so that new insights can be uncovered.

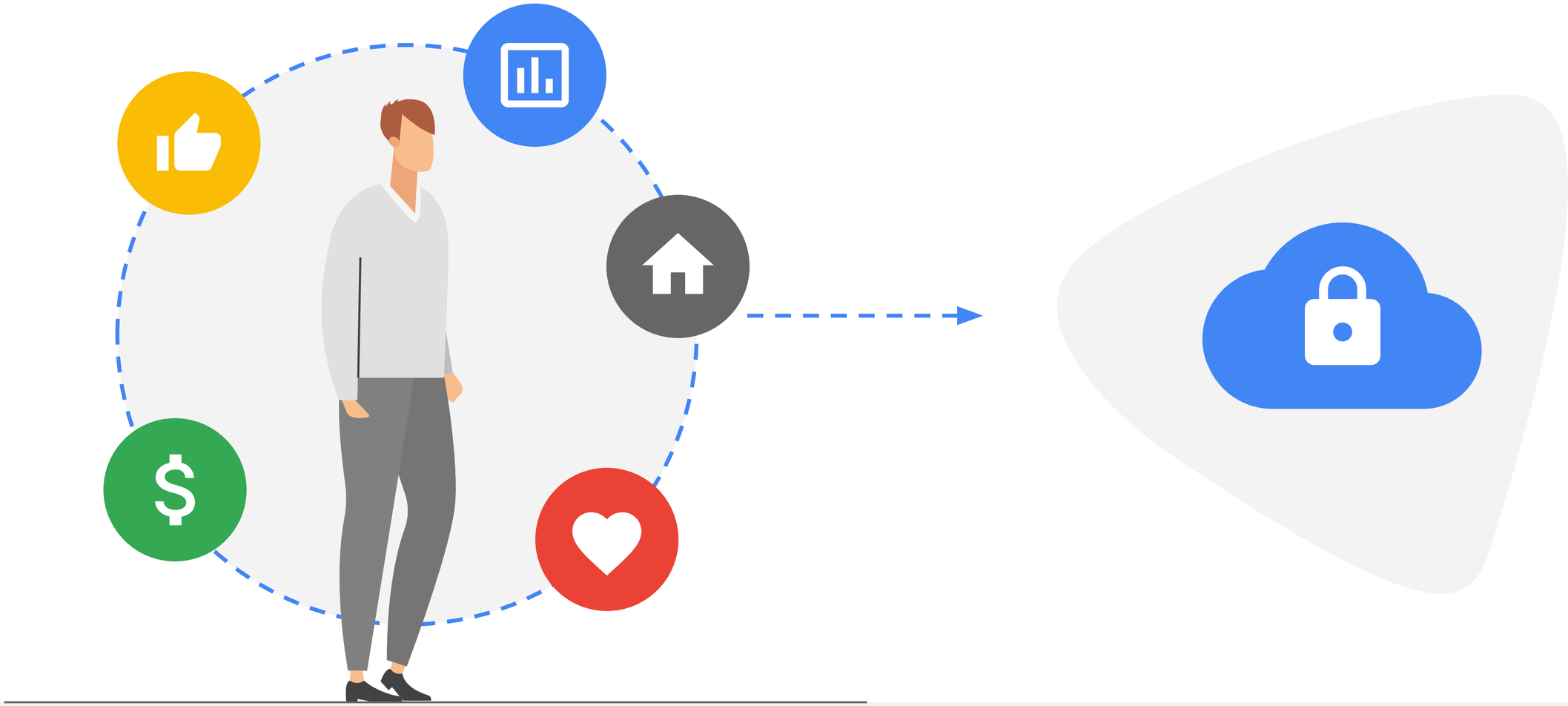




Handling volumes and diversity of data comes with its own ethical considerations and requires alternative ways of thinking about security. Not all information that *can be* captured, *should* be captured. Businesses are accountable for making responsible decisions about which data they collect, store, and analyze.

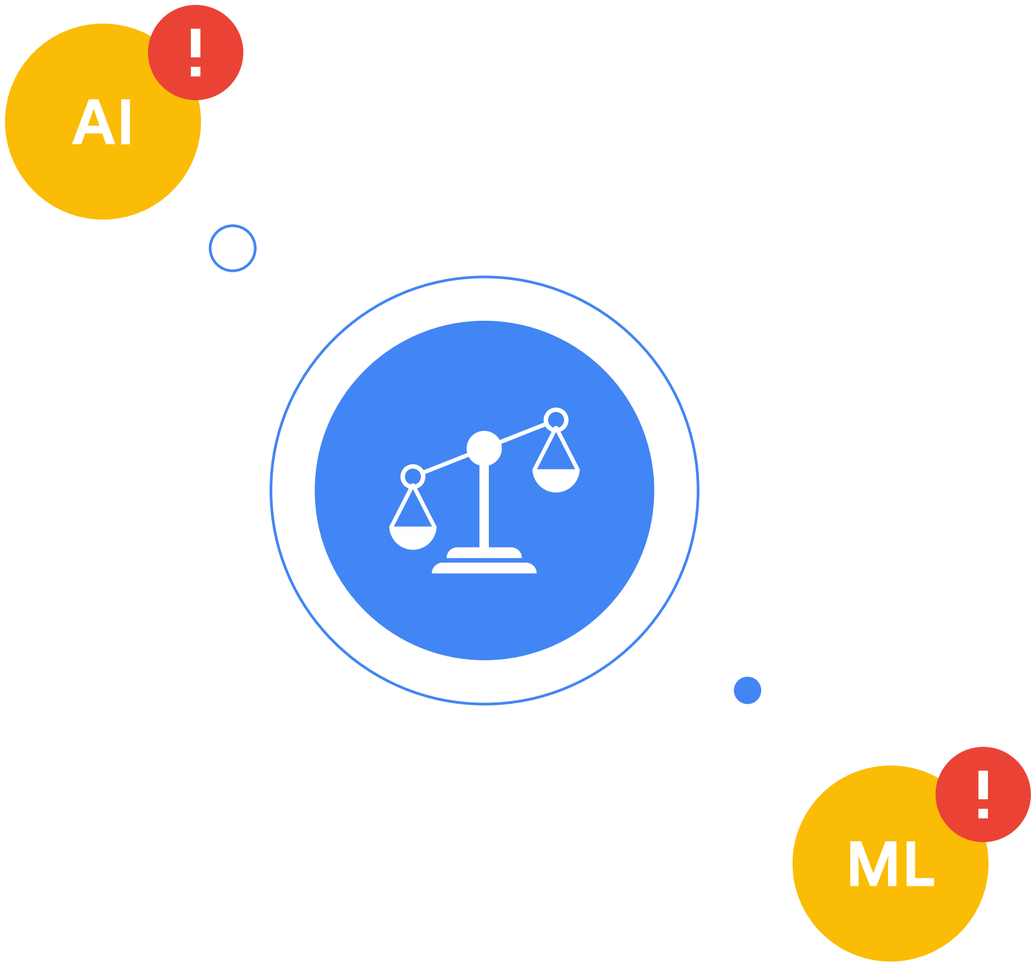


Consider the source of the data, how it is being collected and where it’s stored.

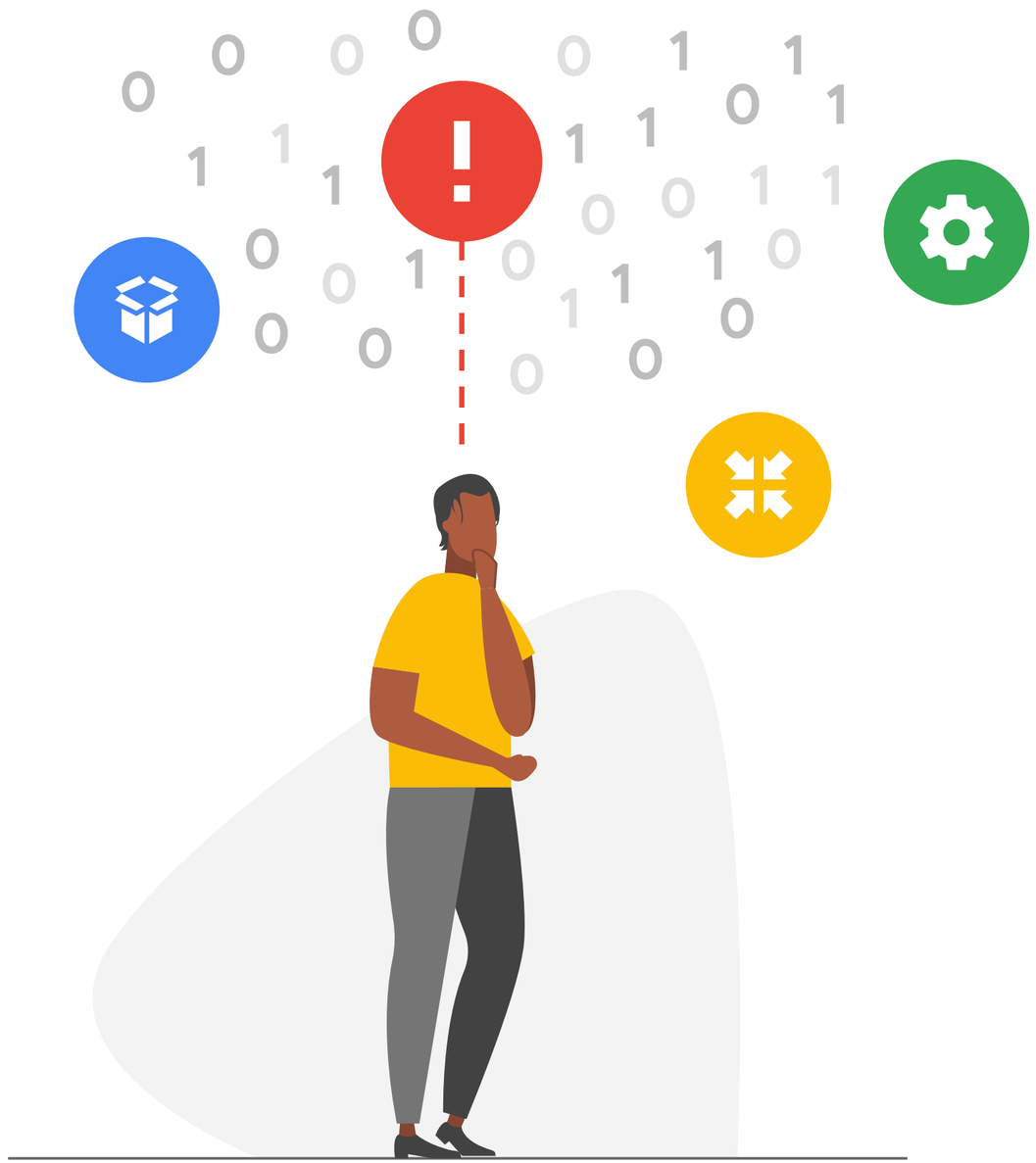


If it’s personal or sensitive data about a customer or an employee, it needs to be securely collected, encrypted when stored in the cloud, and protected from external threats.

Regional or industry specific regulations often guide data policies.



Ethical and fair considerations are particularly important and applicable when you work with Artificial Intelligence (AI) and Machine Learning.



Human bias can influence the way datasets are collected, combined, and used. It's always important to include strategies to remove unconscious biases as you start to leverage data to build new business value.

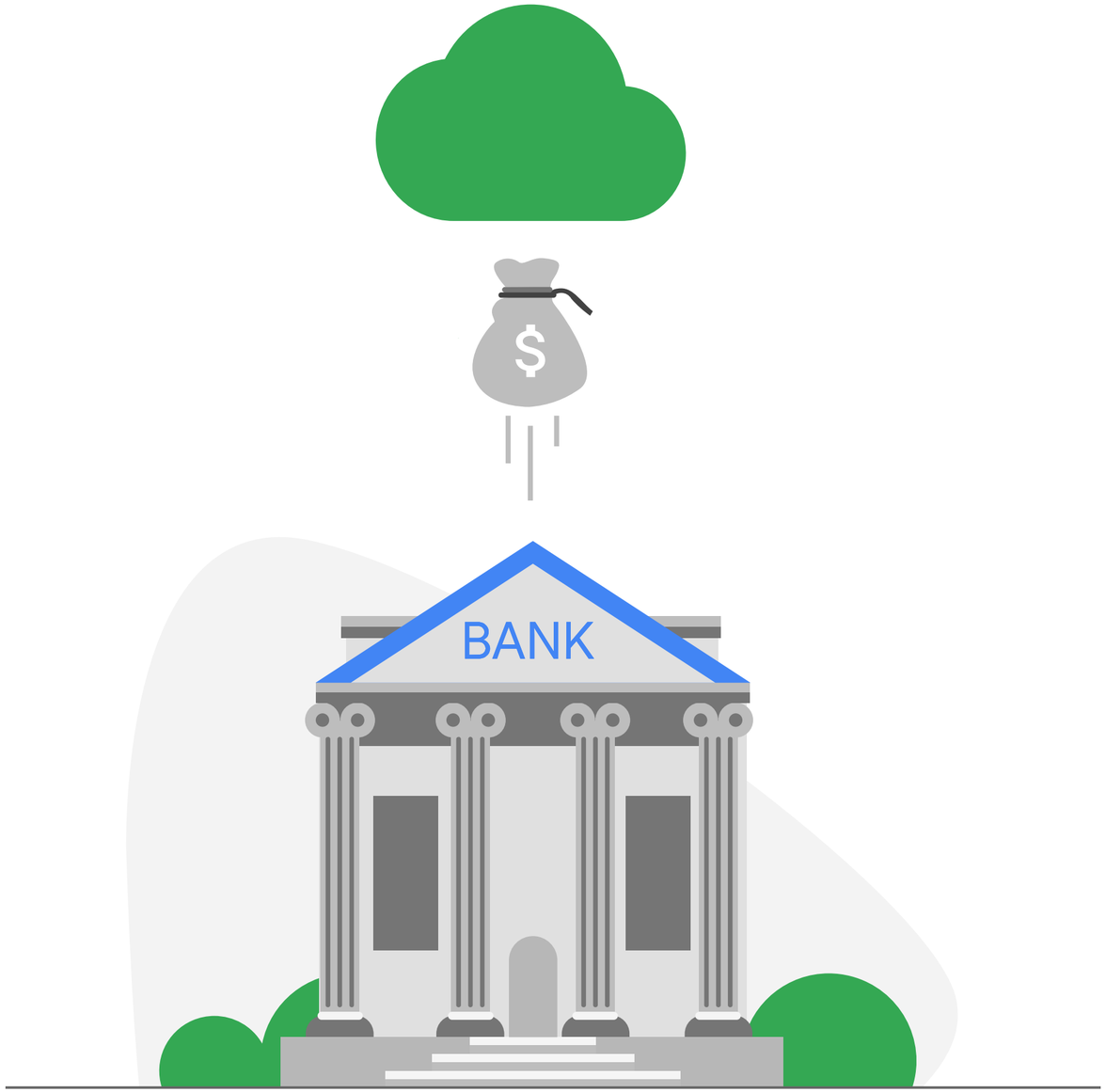
Google Cloud

Module 2: Student Slides

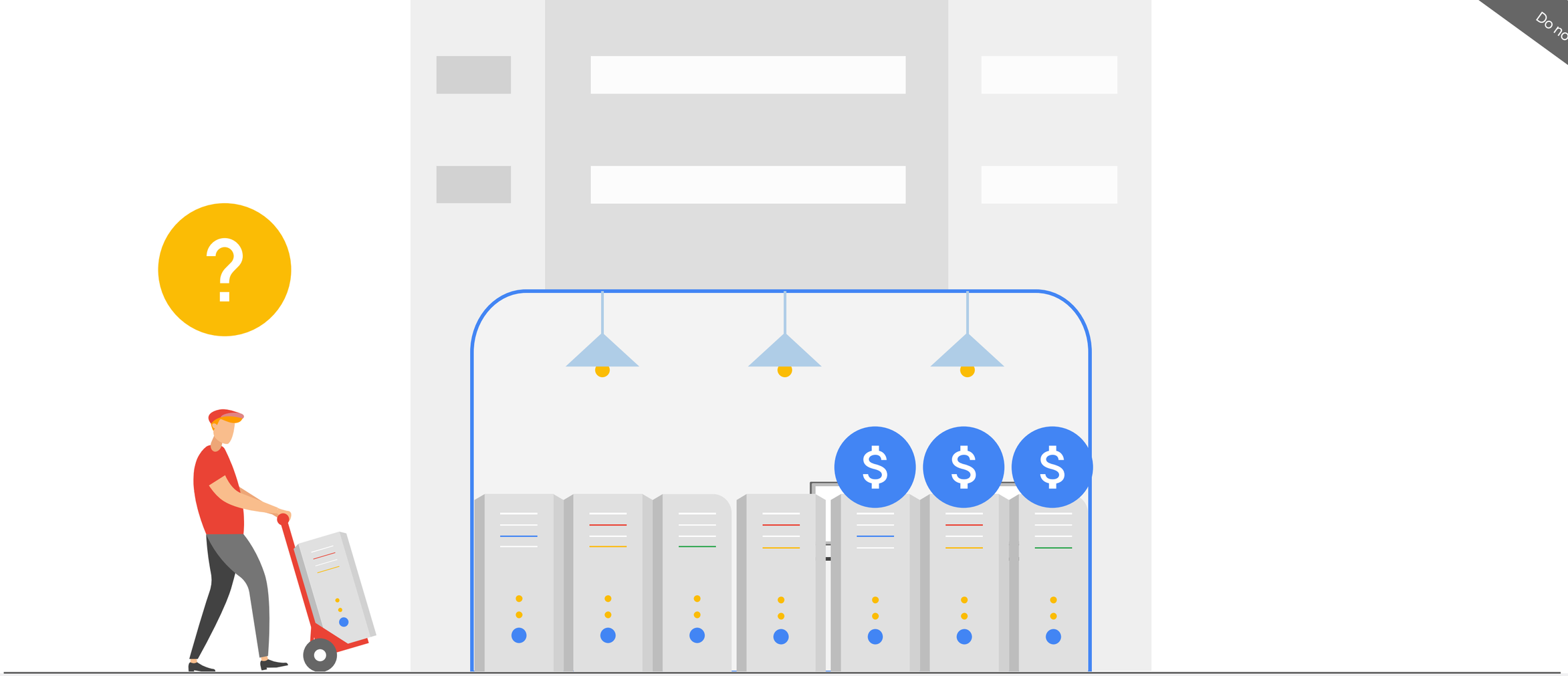
Data Consolidation and Analytics

Topics covered

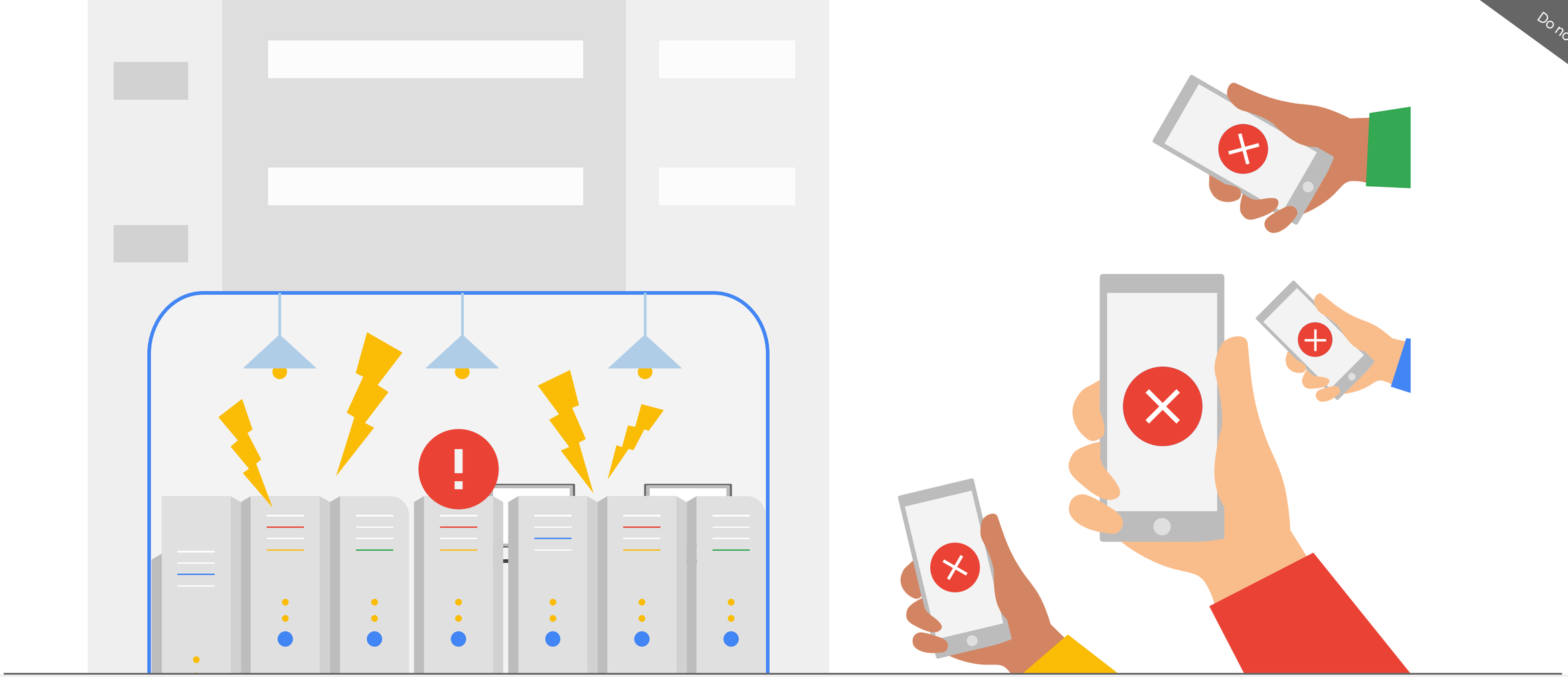
- Migrating data to the cloud
- Databases, data warehouses, and data lakes
- Business intelligence solutions



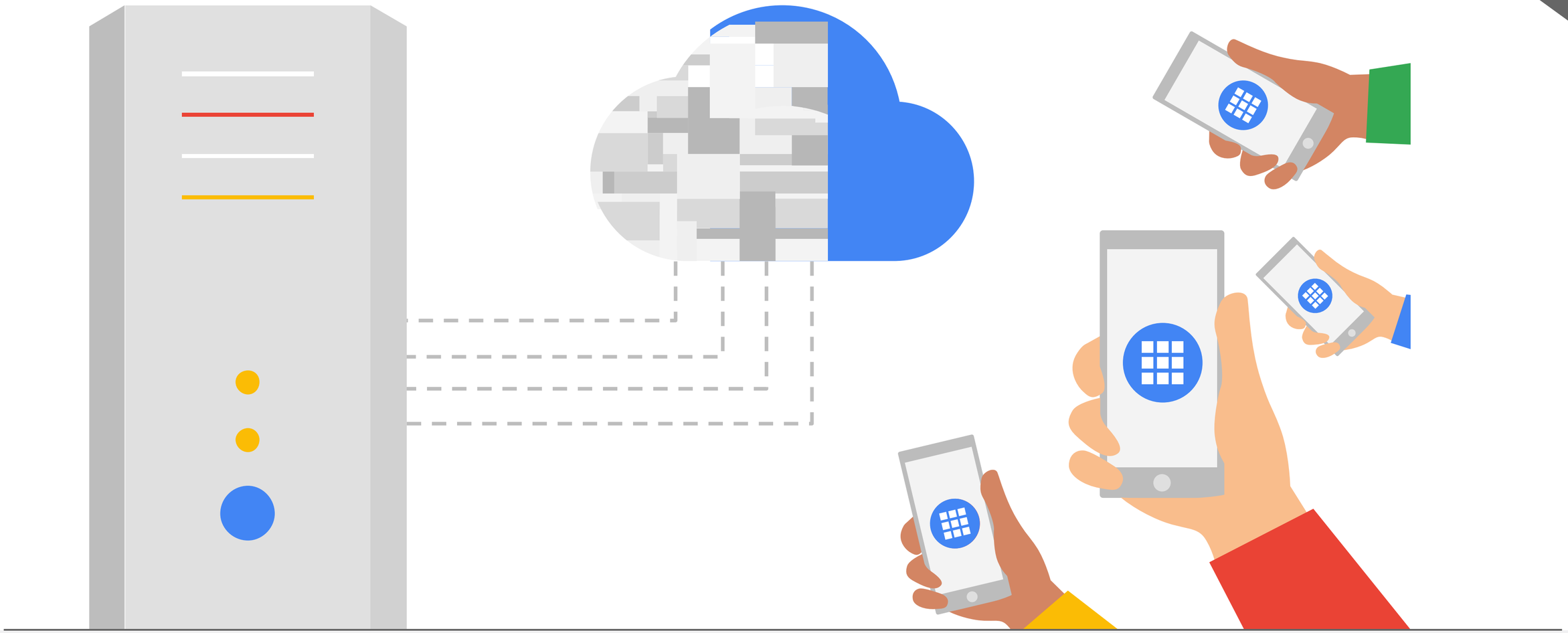
If you're storing your data on-premises, you'll need to start thinking about taking some or all of it to the "bank"—in other words, the cloud. It will provide a greater return on investment.



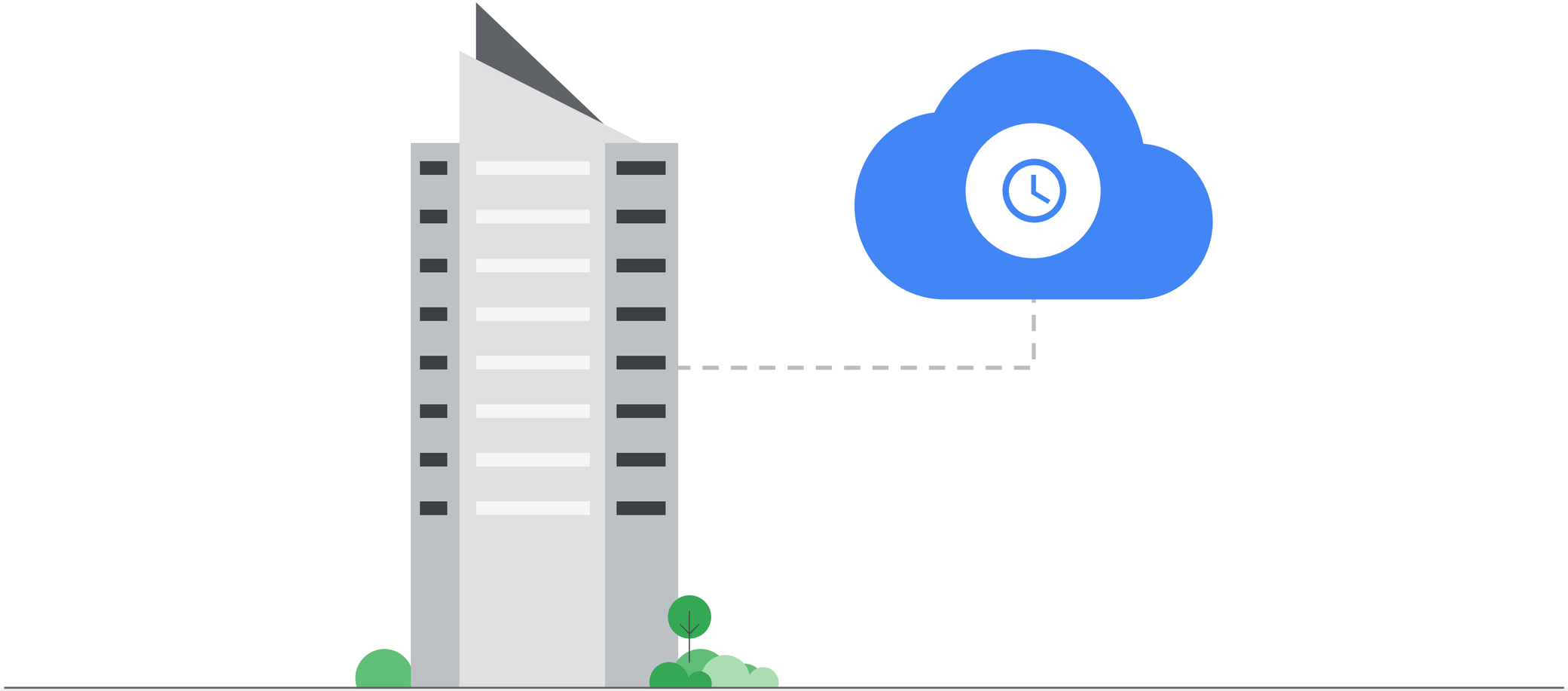
When you store your data on-premises, you are responsible for the IT infrastructure that supports the collection, security, and processing of that data. You're also responsible for maintaining and expanding the capacity of your IT infrastructure.



You also risk downtime, resulting in dissatisfied users.



With cloud, you can ‘rent’ space from public cloud providers like Google Cloud. This means that their data storage and compute power is elastic.



Another way that migrating data to the cloud provides a better return on investment is the speed at which you can ingest and use data. Businesses can now ingest data in real-time.

What is a database?

An organized collection of data, generally stored in tables and accessed electronically from a computer system

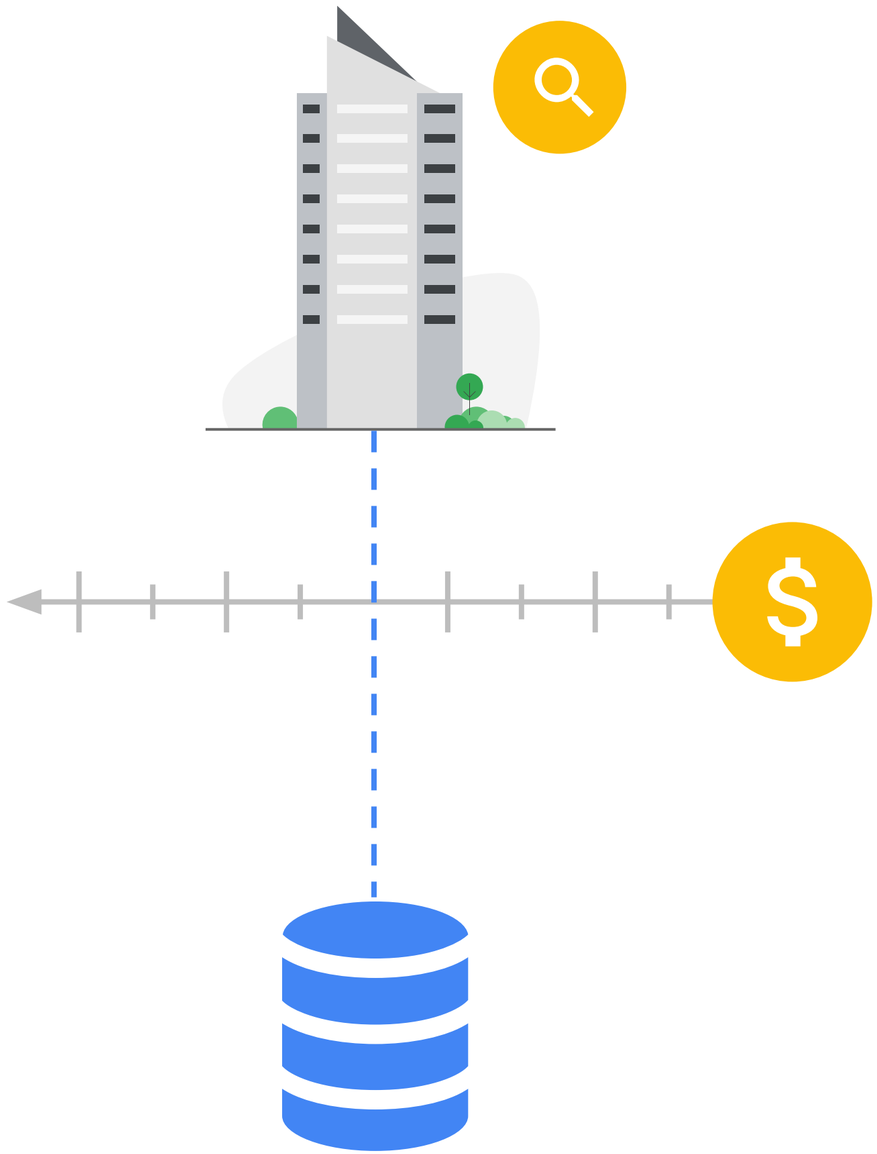
Data management priorities

Data integrity

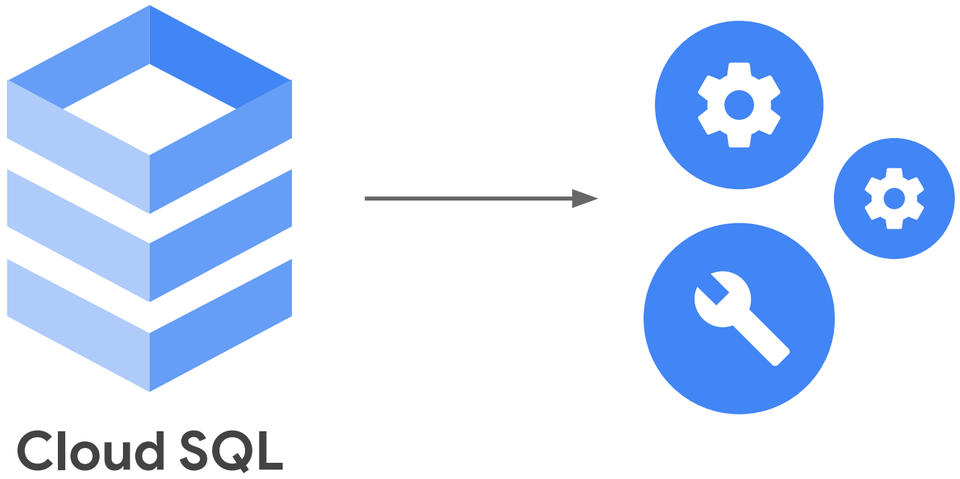
Scale

What is data integrity?

Data integrity, or transactional integrity, refers to the accuracy and consistency of data stored in a database. Data integrity is achieved by implementing a set of rules when a database is first designed and through ongoing error checking and validation routines as data is collected.

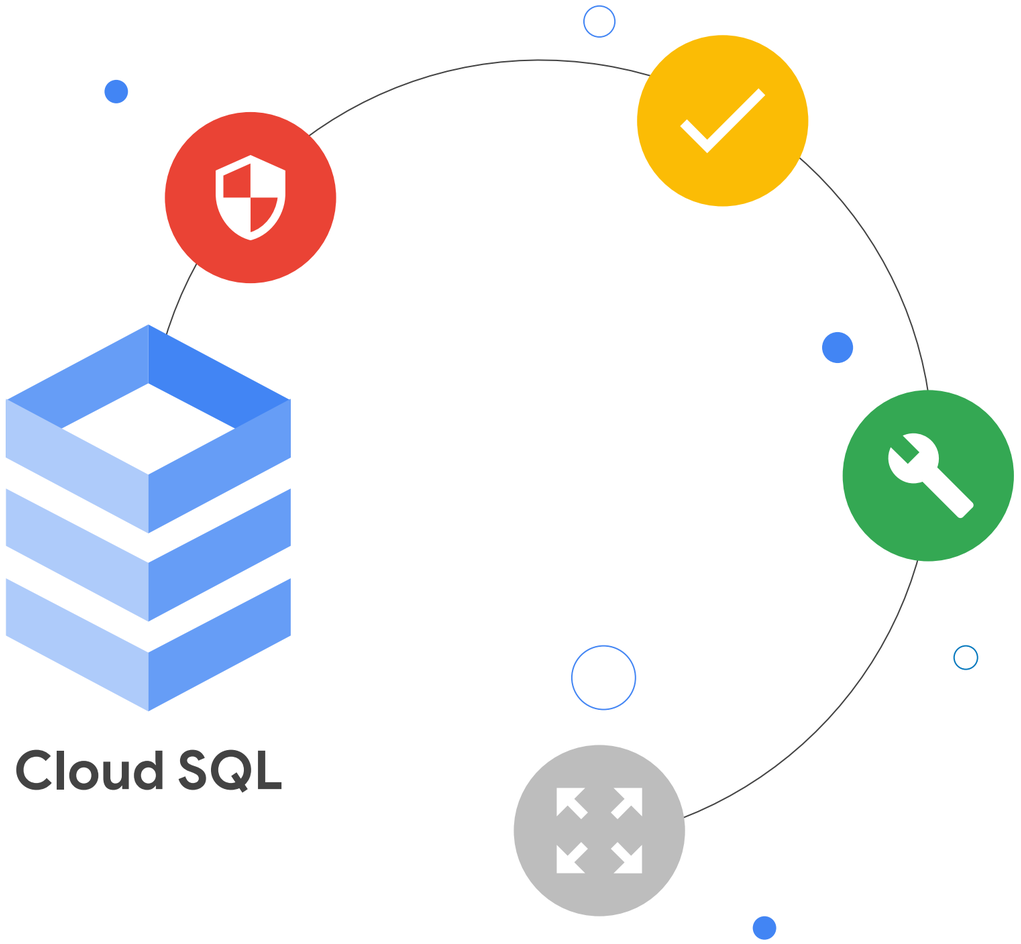


Databases also allow businesses to rollback transactions to see data history.



What is Cloud SQL?

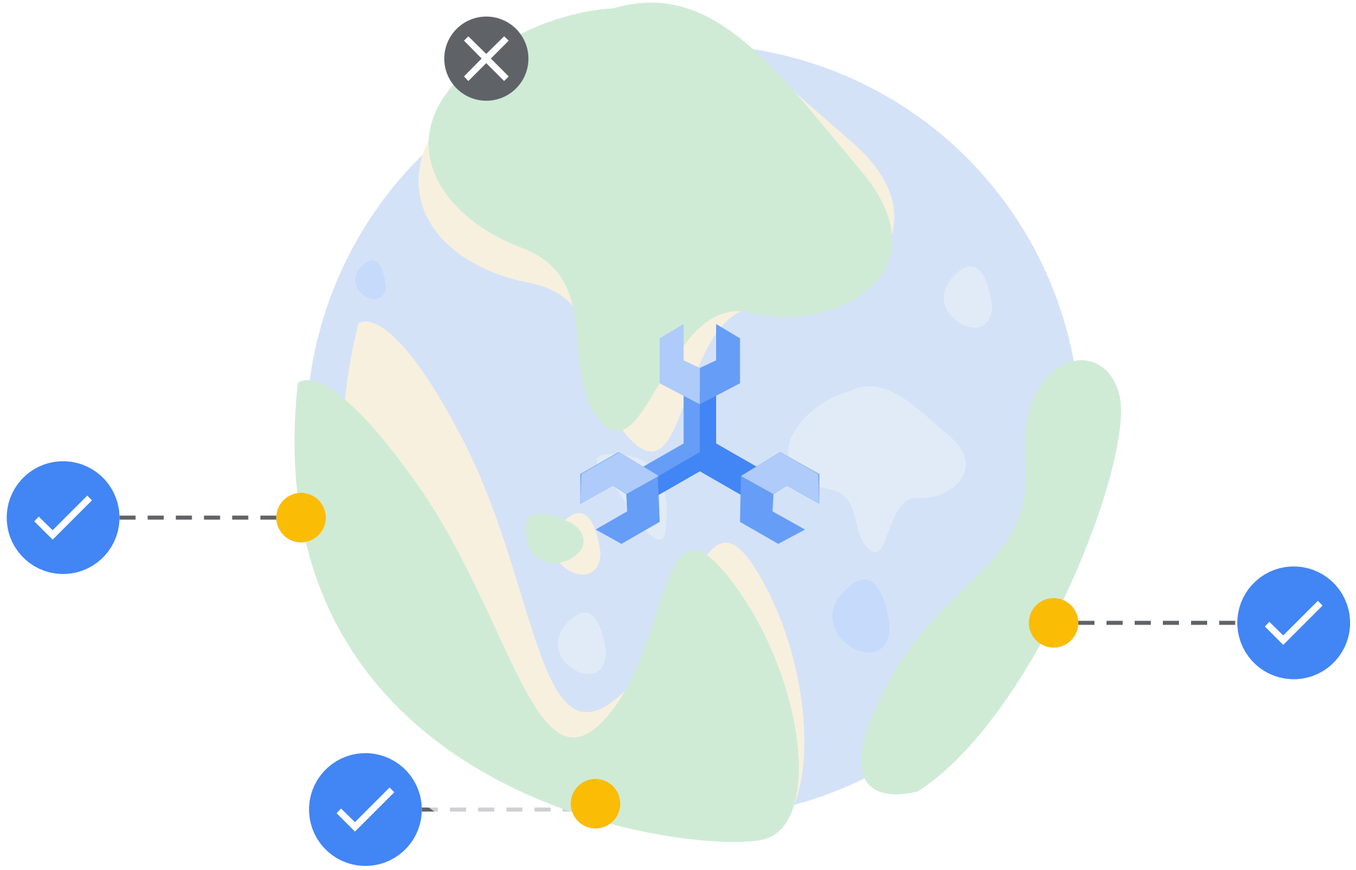
It's a fully managed relational database management service, or RDBMS. It easily integrates with existing applications and Google Cloud services like Google Kubernetes Engine and BigQuery.

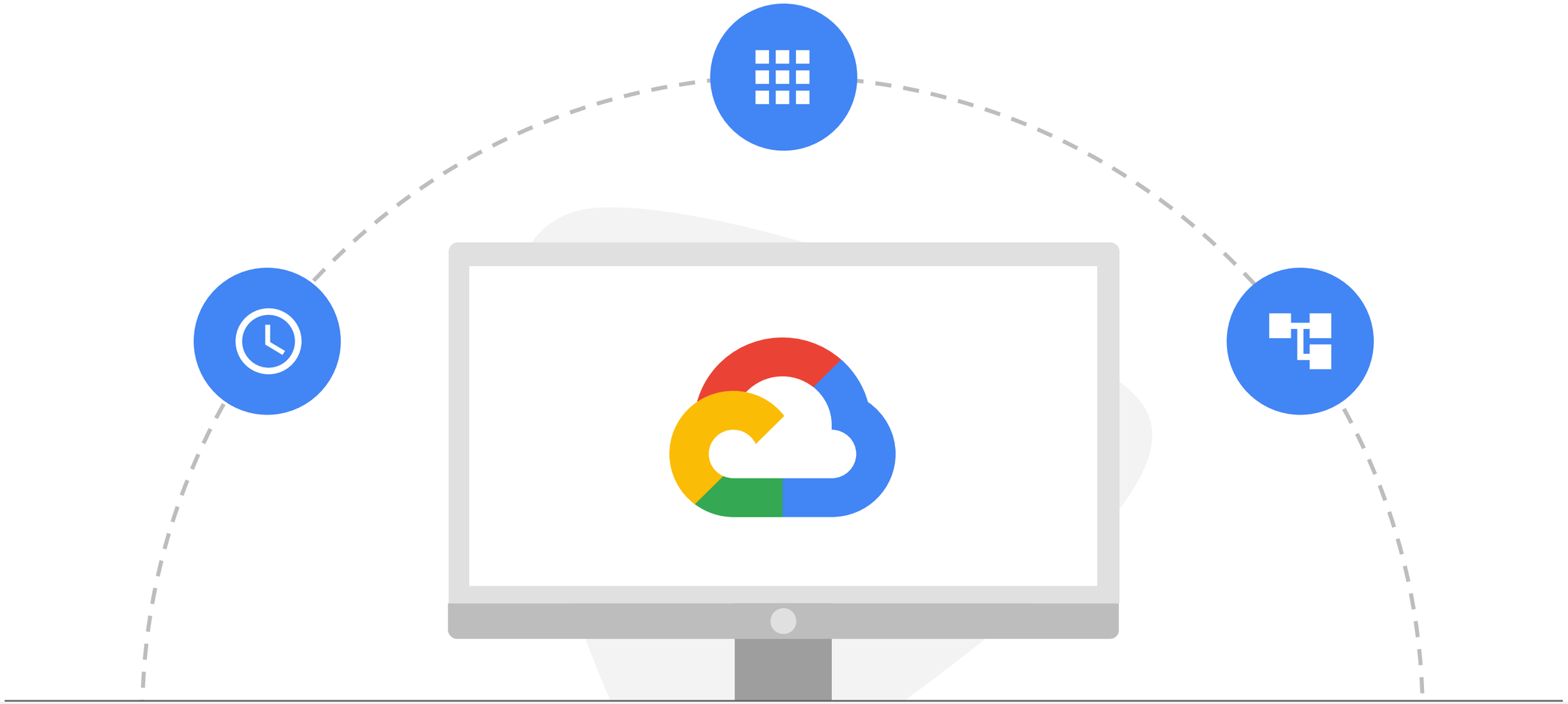


Cloud SQL offers security, availability and durability, and storage scales up automatically when enabled.

What is Cloud Spanner?

It's another fully managed database service, and it's designed for global scale. With Cloud Spanner, data is automatically and instantly copied across regions. This replication means that if one region goes offline, the organization's data can still be retrieved from another region.



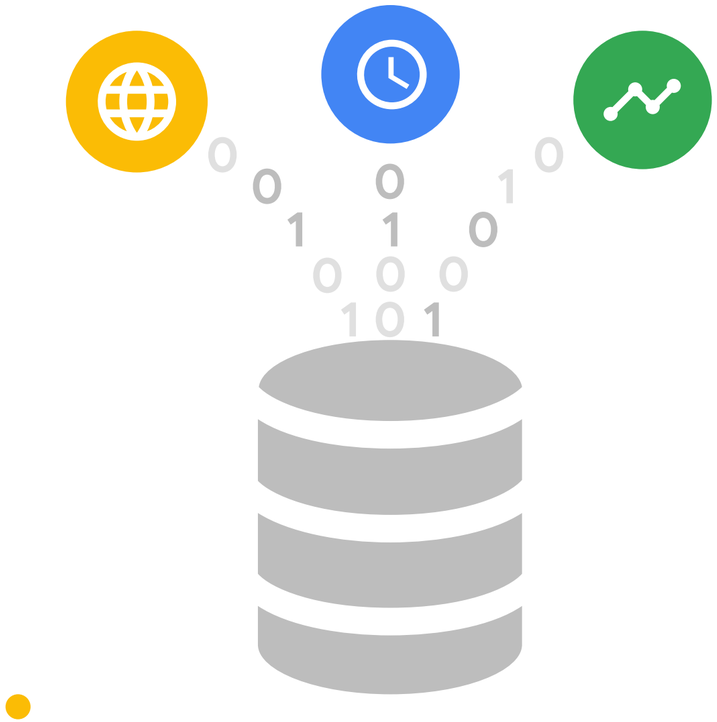


With Google Cloud databases, businesses can build and deploy faster, deliver transformative applications, and maintain portability and control of their data.

Database

Data warehouse

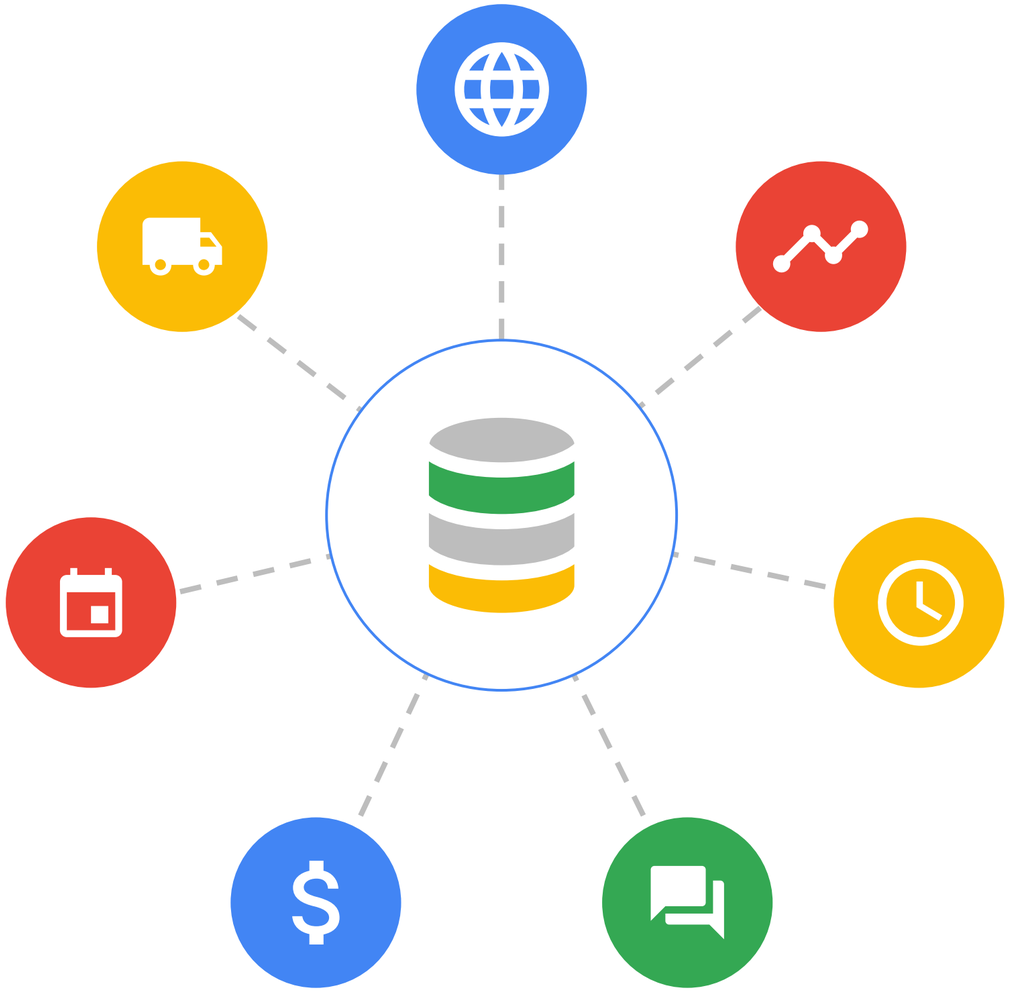
While databases **store** transactional data in an online fashion, data warehouses **assemble** data from multiple sources including databases.



Database

Data warehouse

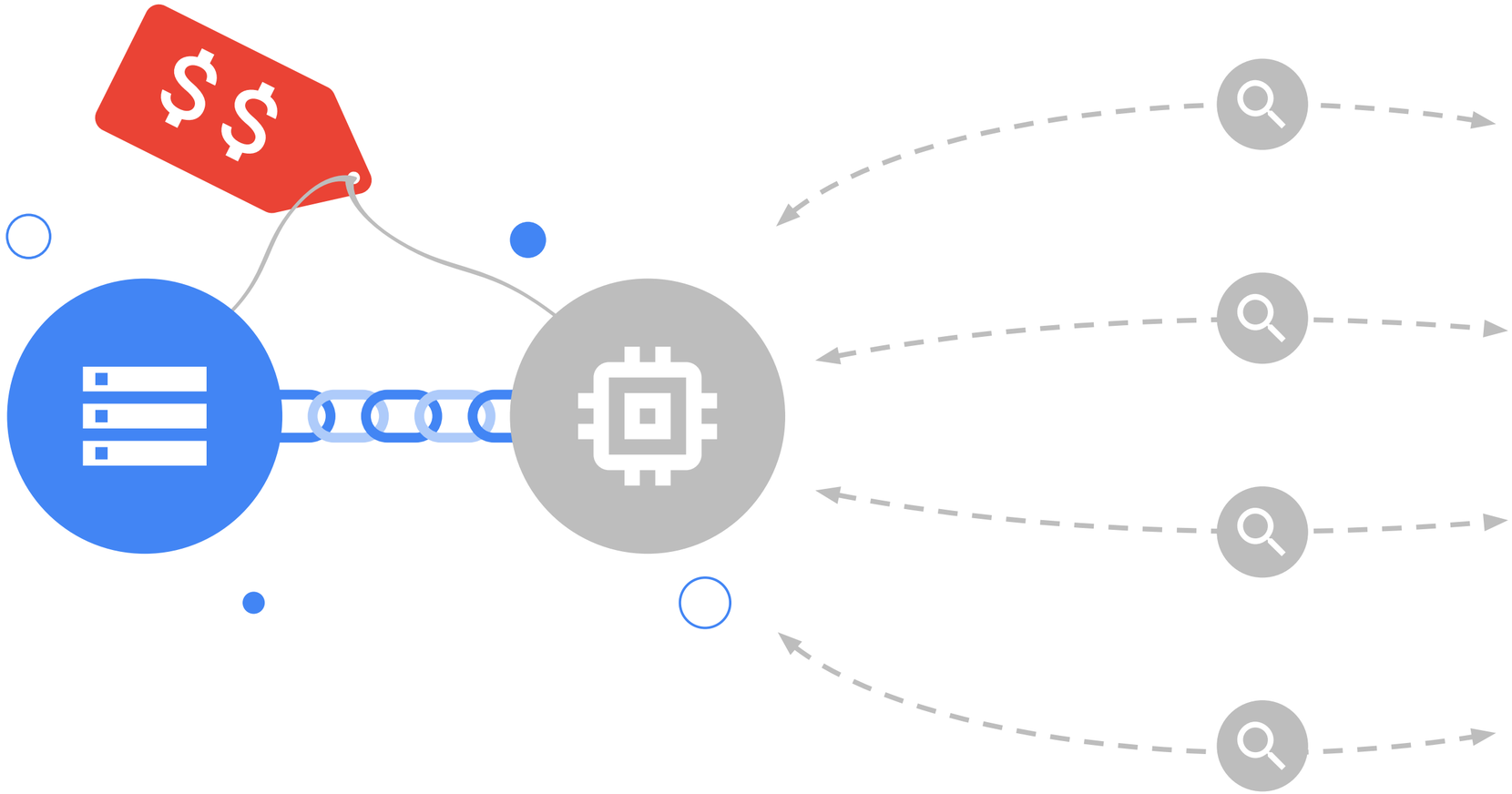
Databases are built and optimized to enable ingesting large amounts of data from many different sources efficiently. However, data warehouses are built to enable rapid analysis of large and multi-dimensional datasets.



Think of the data warehouse as the central hub for all business data. Different types of data can be transformed and consolidated into the warehouse so that they are useful for analysis.

In particular, a cloud data warehouse allows businesses to consolidate data that is structured *and* semi-structured.

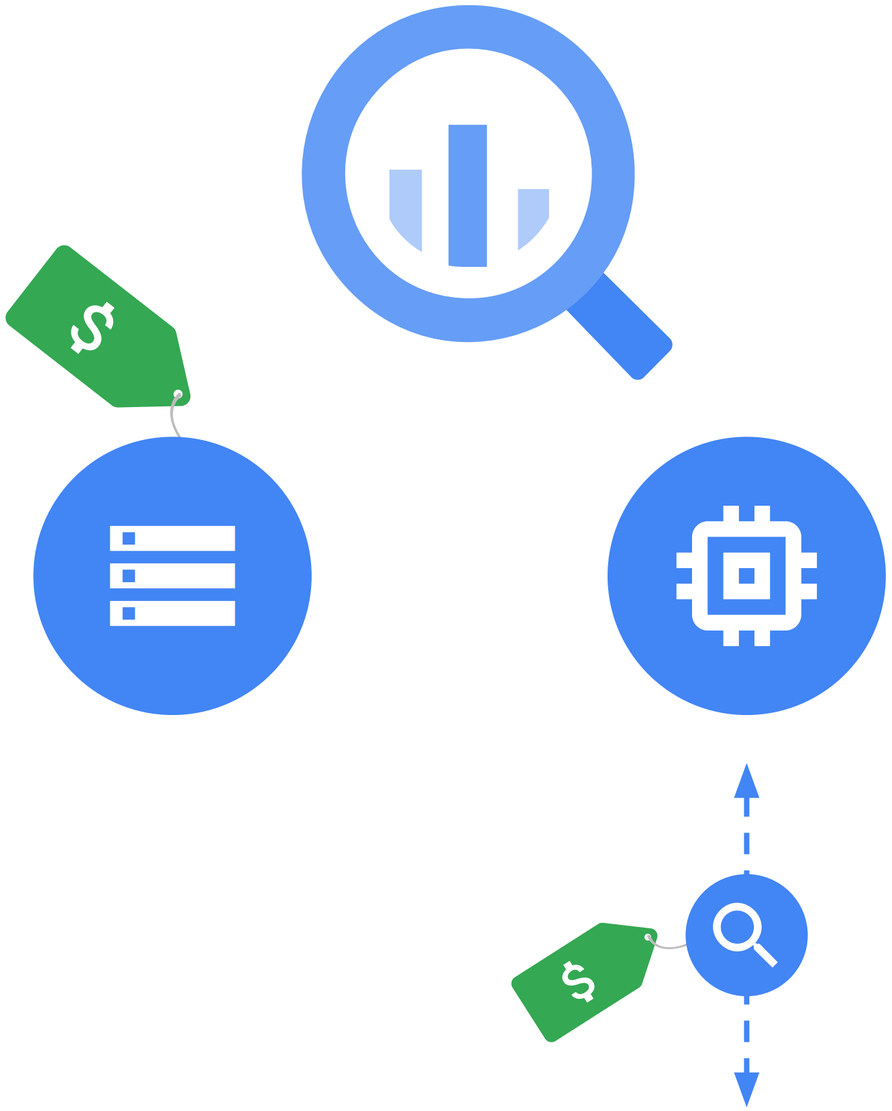
When combined with connector tools, data warehouses can transform unstructured data into semi-structured data that can be used for analysis.



Most data warehouse providers link storage and compute together, so customers are charged for compute capacity whether they are running a query or not.

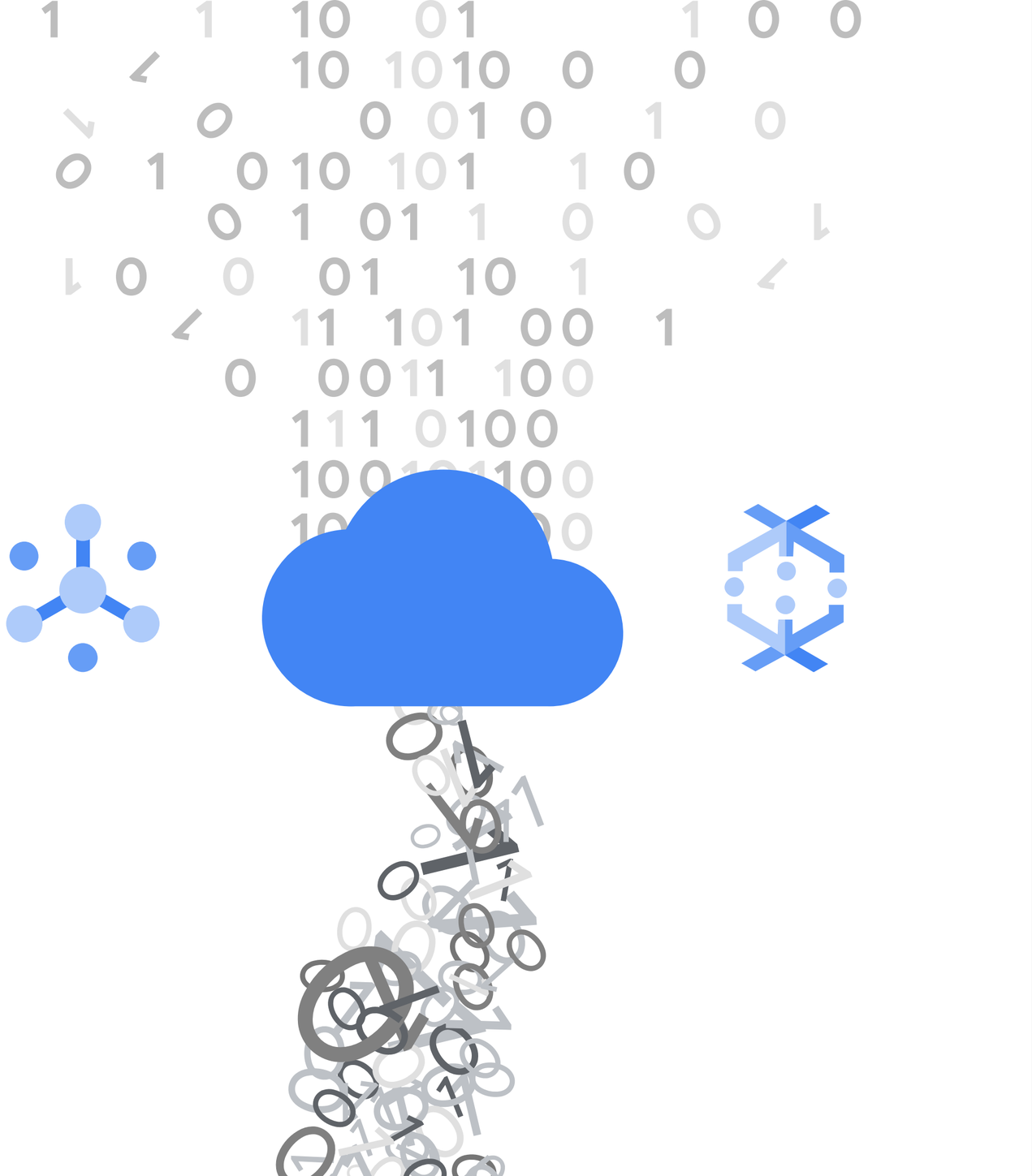
What is BigQuery?

BigQuery is serverless. This doesn't mean that there's no server!

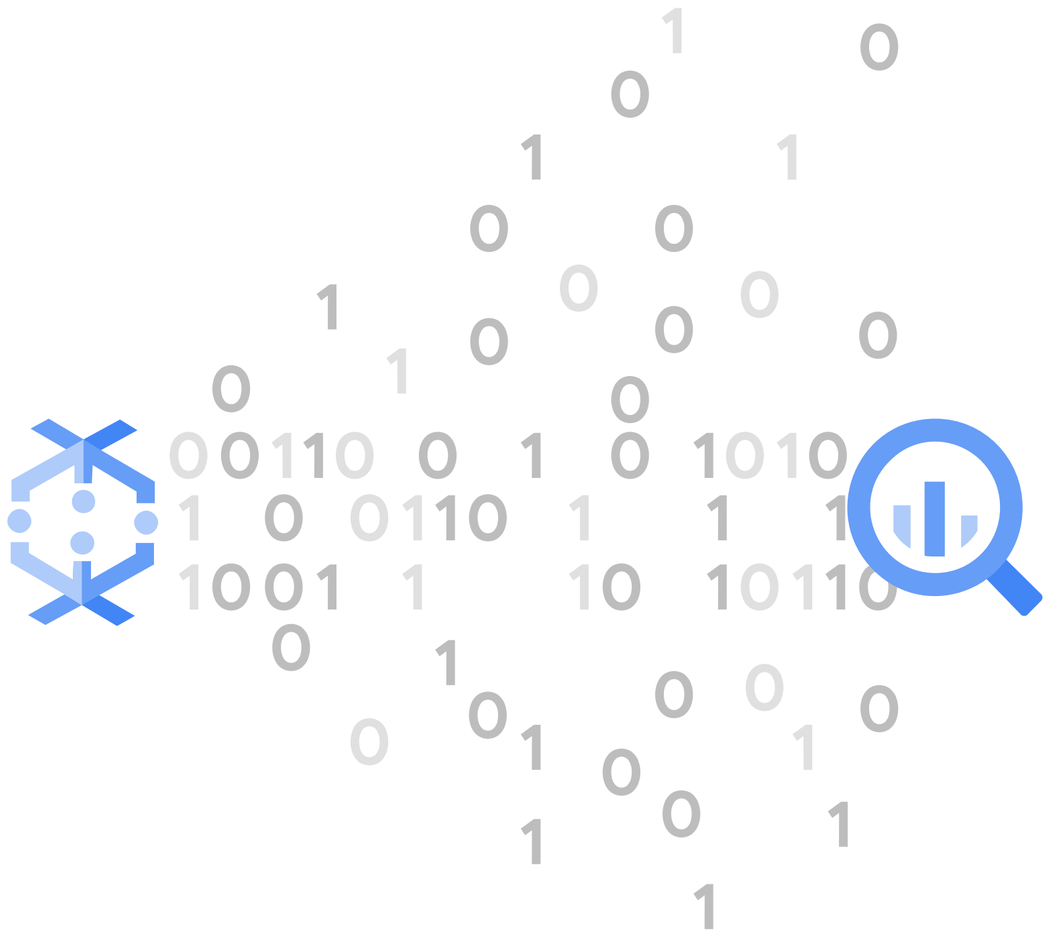


What is BigQuery?

It means that resources, such as compute power, are automatically provisioned behind the scenes as needed to run your queries. So businesses do not pay for compute power unless they are actually running a query.



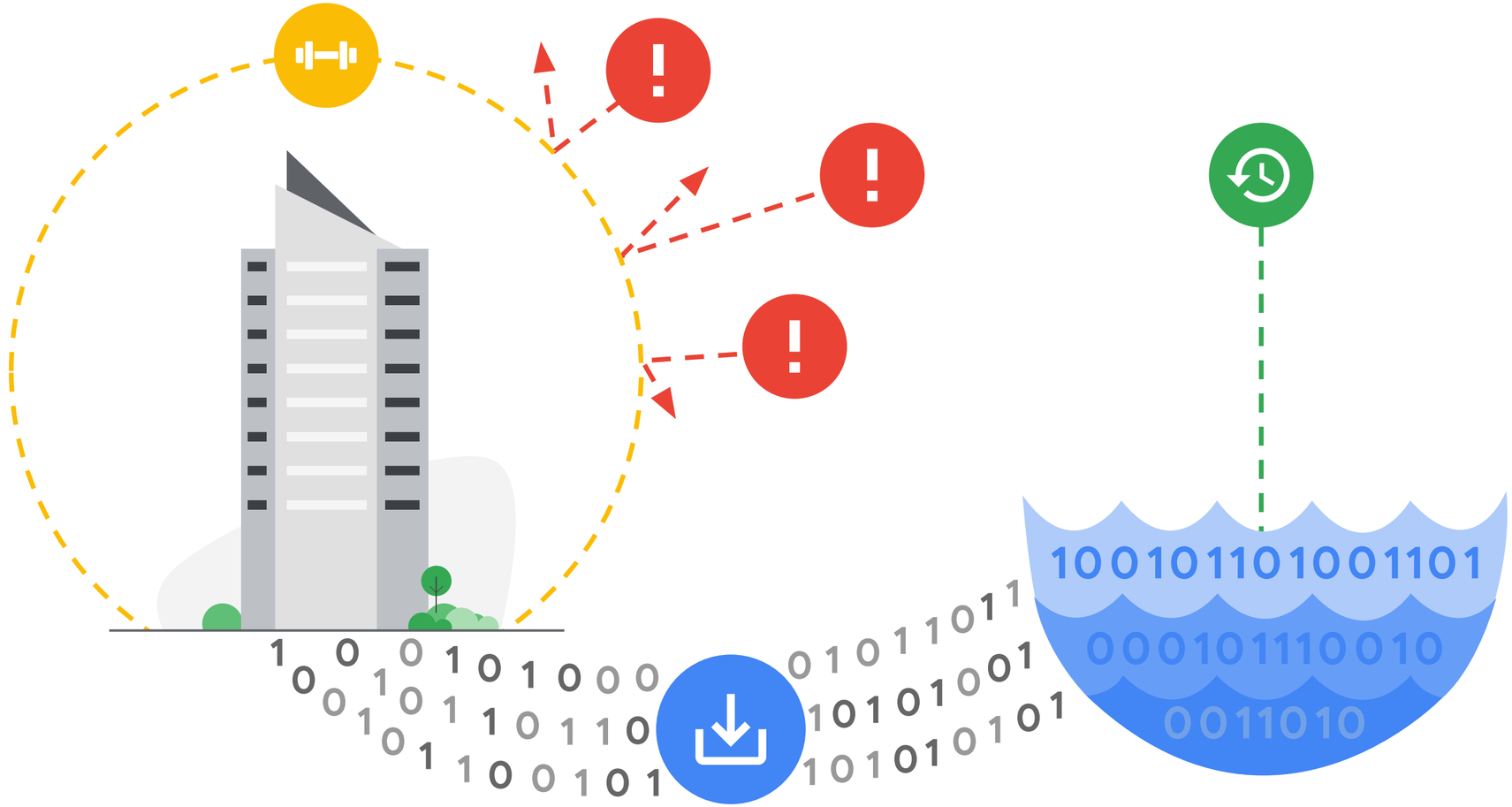
Pub/Sub and DataFlow can work together to bring unstructured data into the cloud and transform it into semi-structured data.



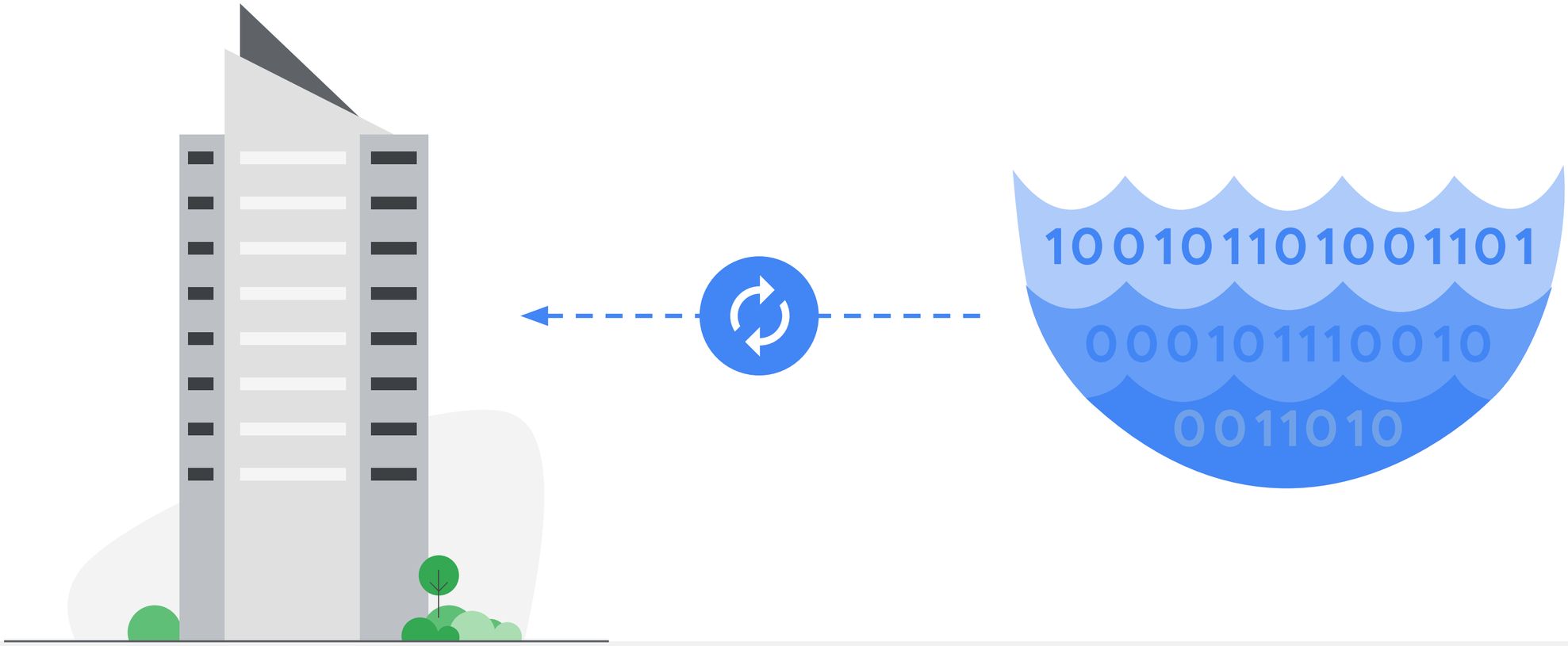
This transformed data can then be sent directly from Dataflow to BigQuery, where it is made immediately available for analysis.

What is a data lake?

Data lakes are a repository for raw data and tend to serve many purposes.



They often hold 'back-up' data, which helps businesses build resilience against unexpected harm affecting their data. Businesses are protected against data loss. They also hold data that is historic and not relevant to day-to-day business operations.



One way to classify an organization's requirements for storage is by how often they need to access the data.

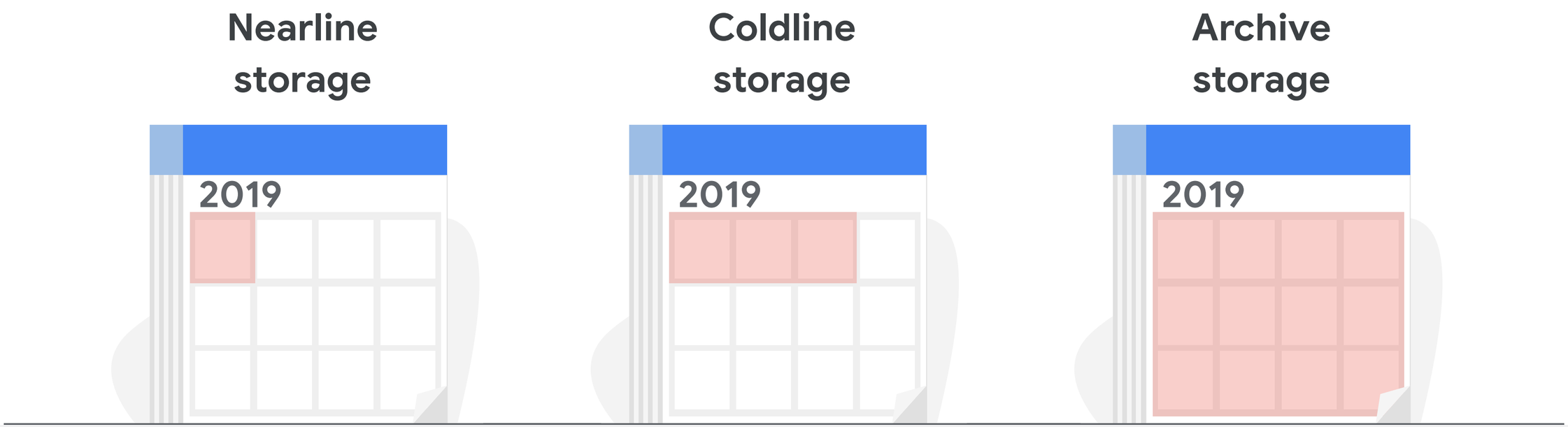
Cloud Storage Benefits

- Any amount of data
- Low latency
- Accessible from anywhere

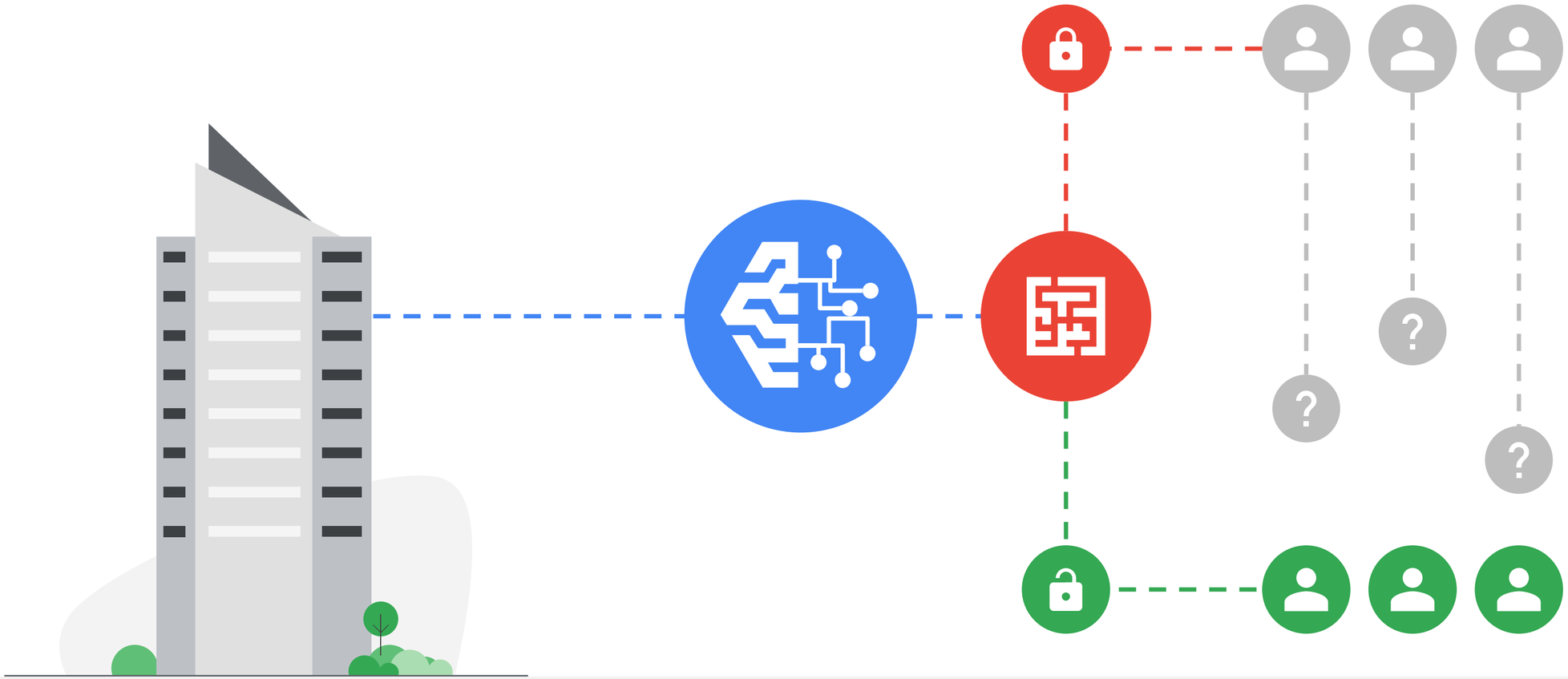
Multi-regional storage

Cloud Storage offers multi-regional storage. It’s ideal for serving content to users worldwide.

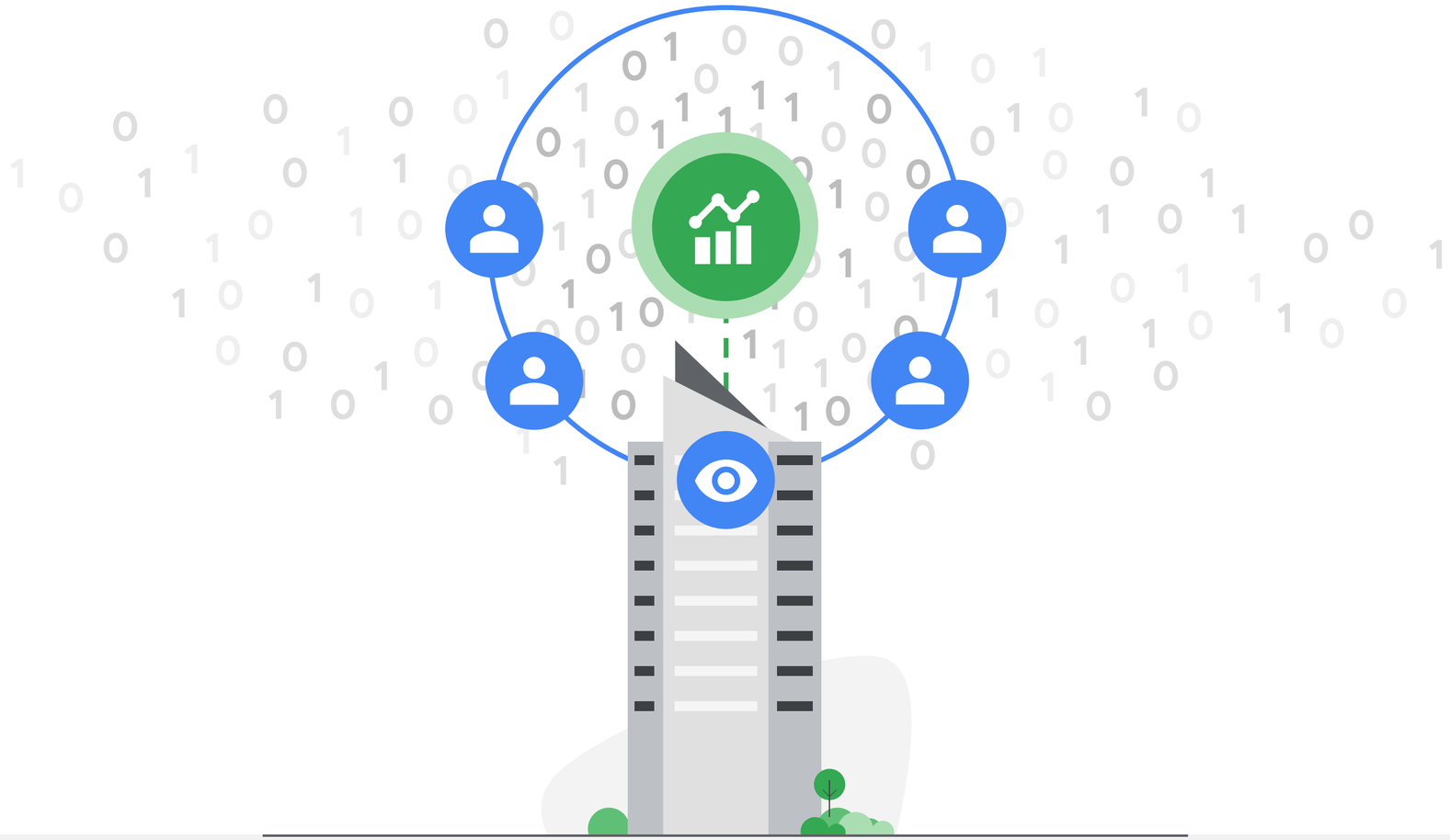
Regional storage offered by Cloud Storage is ideal when an organization wants to use the data locally; it gives added throughput and performance by storing data in the same region as your compute infrastructure.



For data that will be accessed less often, Cloud Storage offers Nearline, Coldline and Archive storage classes.



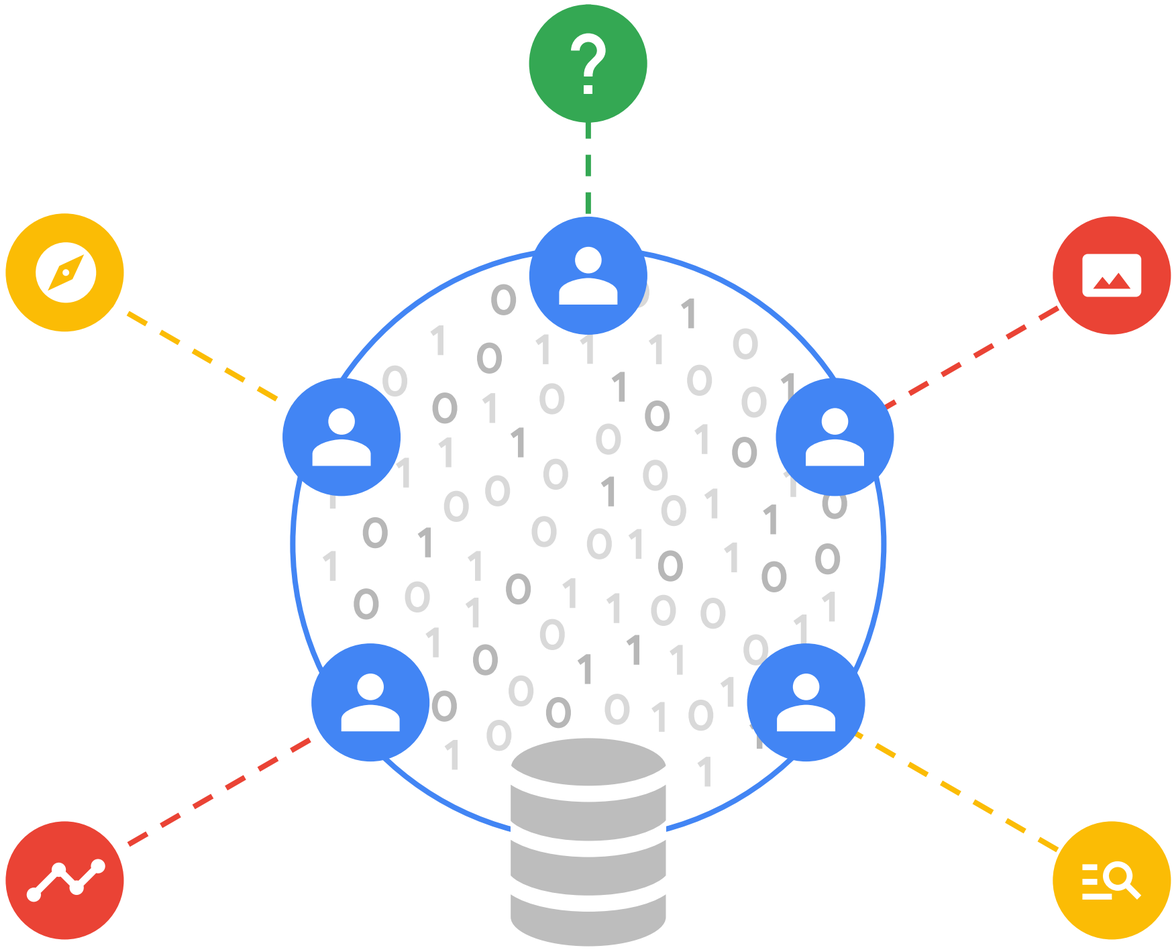
The challenge businesses often face is identifying the right business intelligence solution. Some solutions are too complex and not accessible by anyone outside the data engineering or data analysis teams.



Other solutions let everyone in the business perform their own data analysis. But they can only perform their analysis with portions of the available data. This means that only a few people, or possibly no one, has a full view of the company's business data.

What is Looker?

Looker is a Google Cloud business intelligence solution. It's a data platform that sits on top of any analytics database and makes it simple to describe your data and define business metrics.



What is Looker?

Once you have a reliable source of truth for your business data, anyone on your team can analyse and explore it, ask and answer their own questions, create visualisations, and explore row level details.

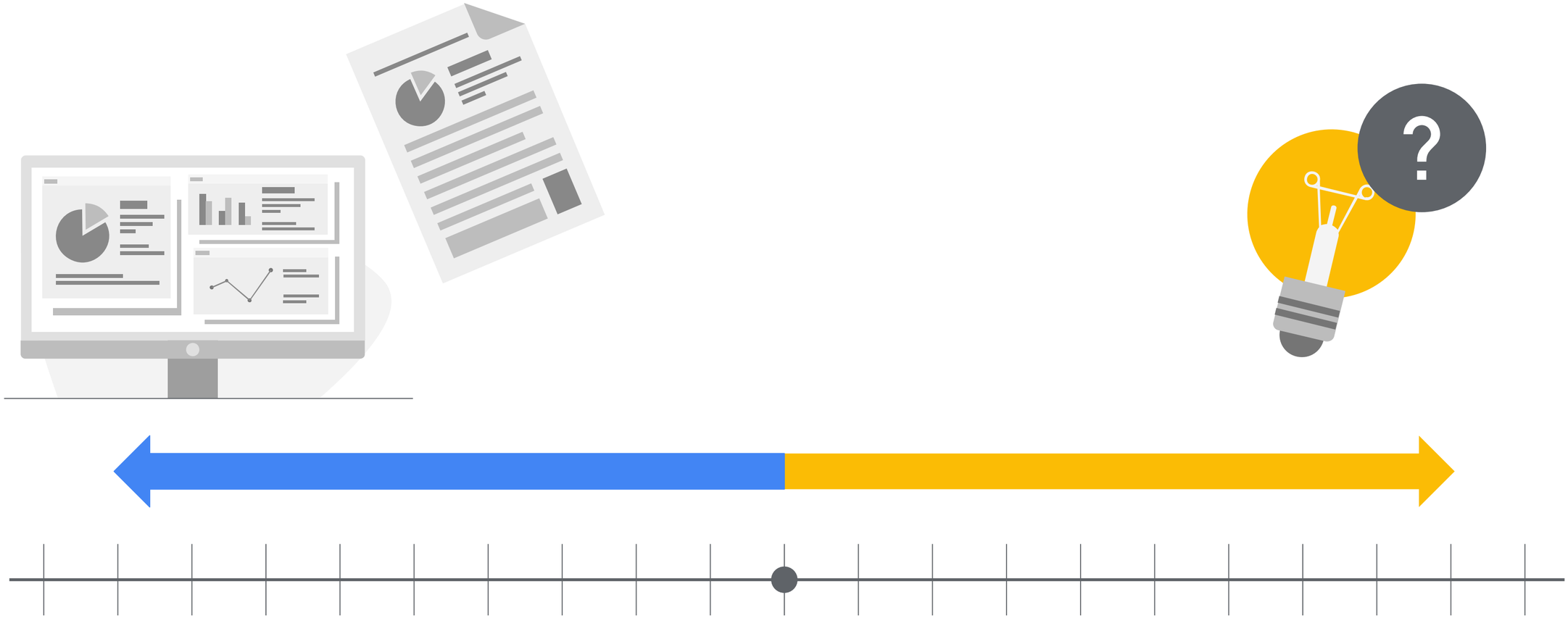
Google Cloud

Module 3: Student Slides

Innovation with Machine Learning

Topics covered

- The definition of ML and AI
- Data quality considerations
- Real world ML use cases



The dashboard and the report are examples of *backward-looking* data. They look at what happened in the past. To create value in your business, you need to use that data to make decisions for **future** business.

Artificial intelligence

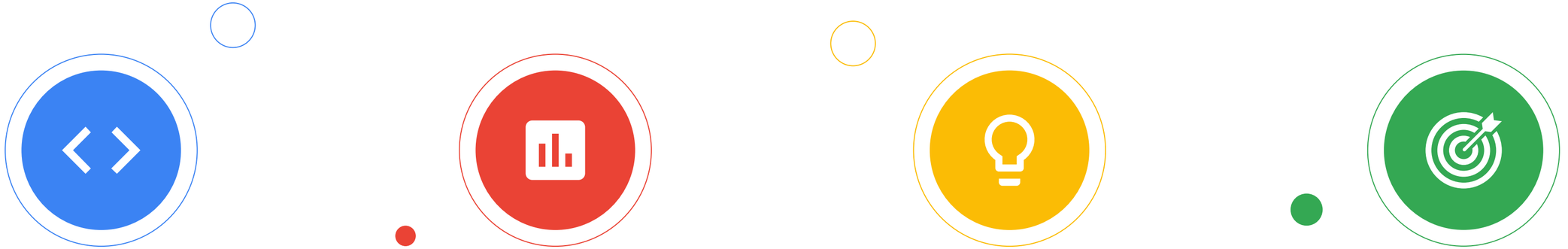
What is AI?

It’s a broad field or term that describes any kind of machine capable of acting autonomously.

Machine Learning

What is ML?

It’s a branch in the field of AI.
Computers that can "learn" from data without using a complex set of rules.



Algorithm

Data

Predictive insight

Decision

Definition used in this course: ML is a way to use **standard algorithms or standard models** to analyze **data** in order to derive **predictive insights** and make repeated **decisions** at scale.

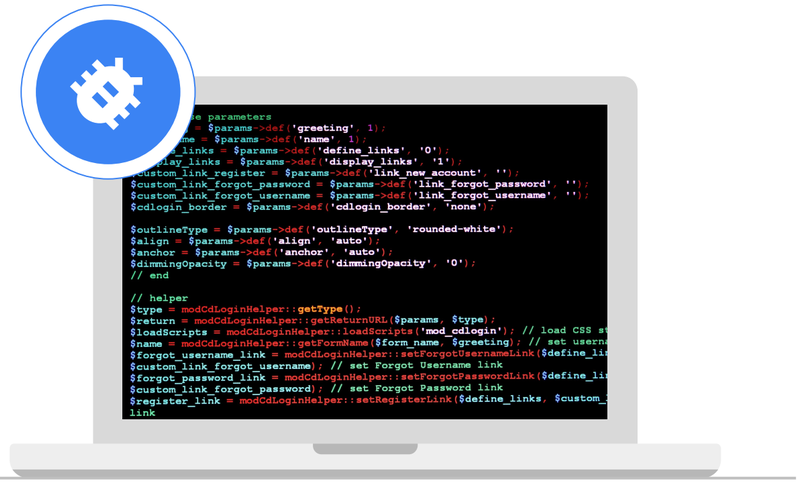
Algorithm

Data

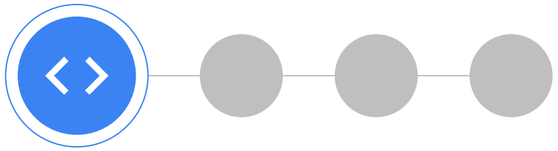
Predictive insight

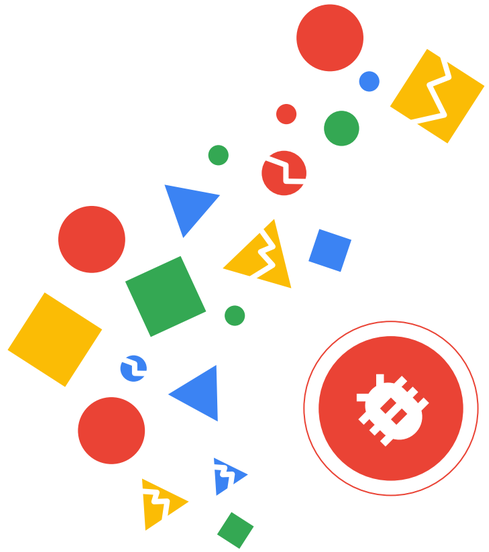
Decision

Analyzed data can be used to derive **predictive insights** and make repeated **decisions**. The accuracy of those predictions depends on large volumes of data that are free of bugs.

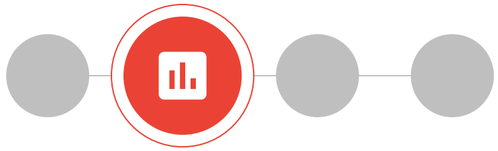


Traditional software





Machine learning



Bugs in ML are often caused by bugs in the data. In traditional software development, a bug is a mistake in the code that causes unexpected or undesired behavior. In ML, even though there can be bugs in the implementation of an algorithm, bugs in data are far more common.

2

Data cleanliness

What is data cleanliness?

Sometimes called “data consistency”

“Dirt” or “inconsistency” in data refers to anything that can prevent the model from making accurate predictions or understanding data behavior.

3

Data completeness

What is data completeness?

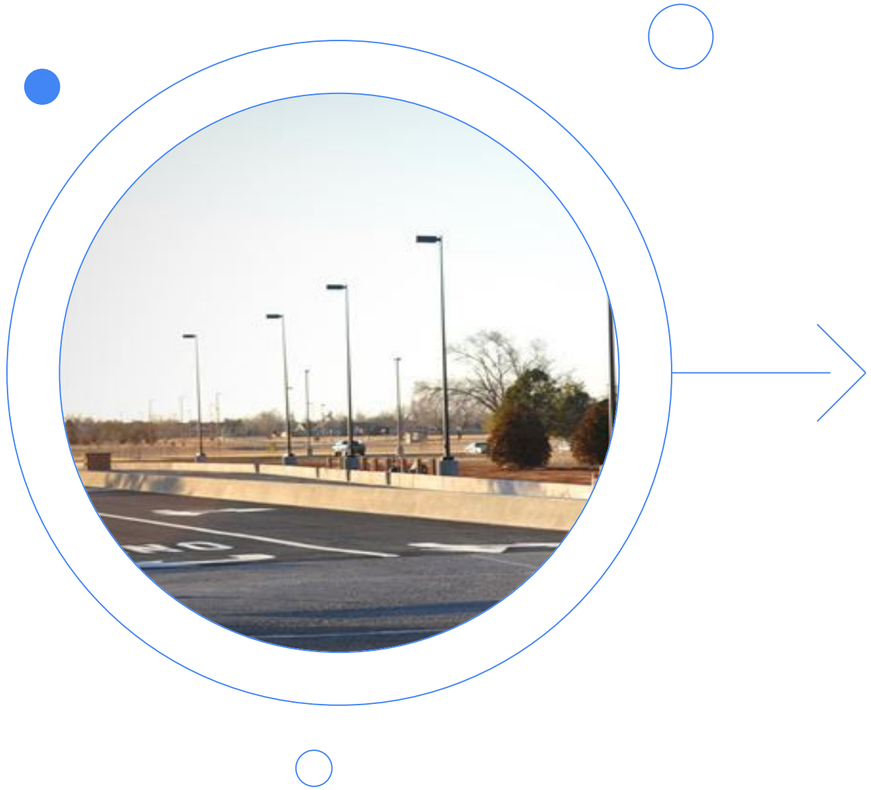
Refers to the **availability** of sufficient data about the world to replace human knowledge.

Lack of availability of better data

Mistaken expectations about how ML works

Poor execution of program design and implementation

Incomplete data can limit the performance of the ML model



Data is the tunnel through which the model views the world.

Improve coverage?

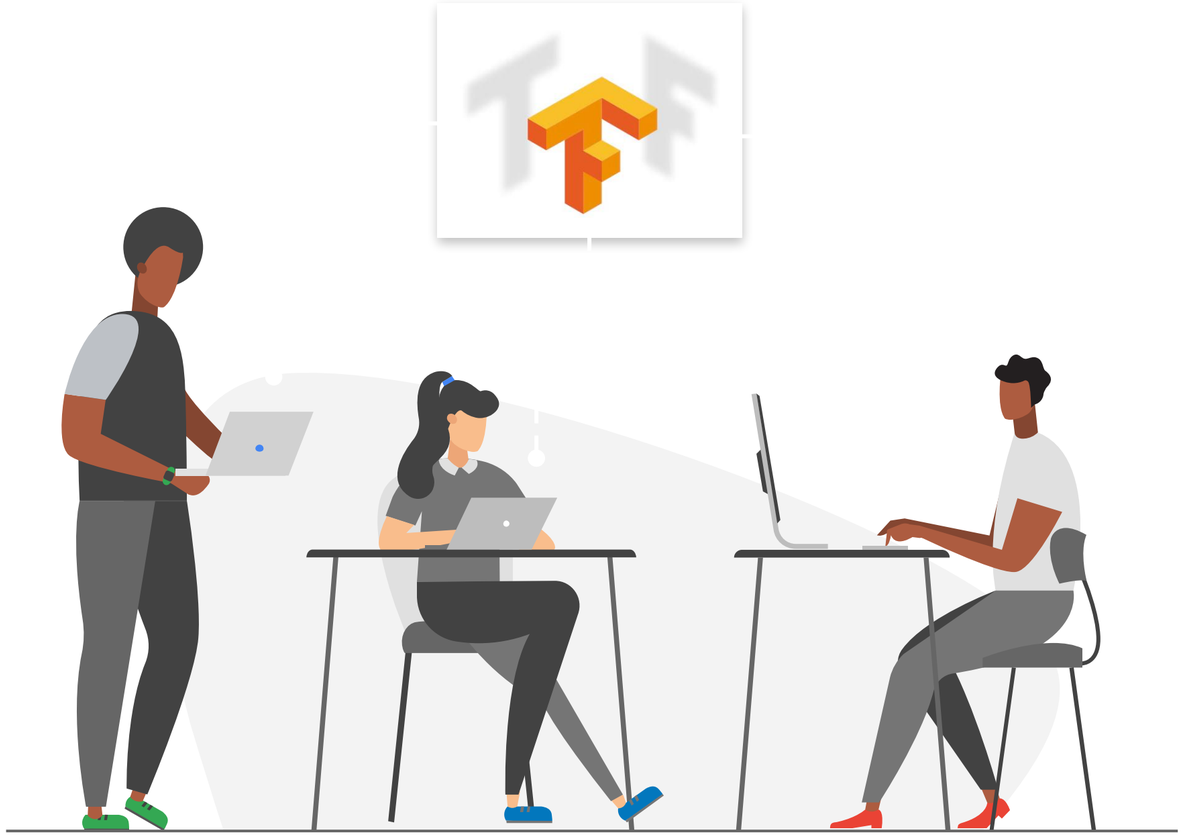
Improve cleanliness?

Make data complete?

Bugs in data can be improved through purposeful data collection.

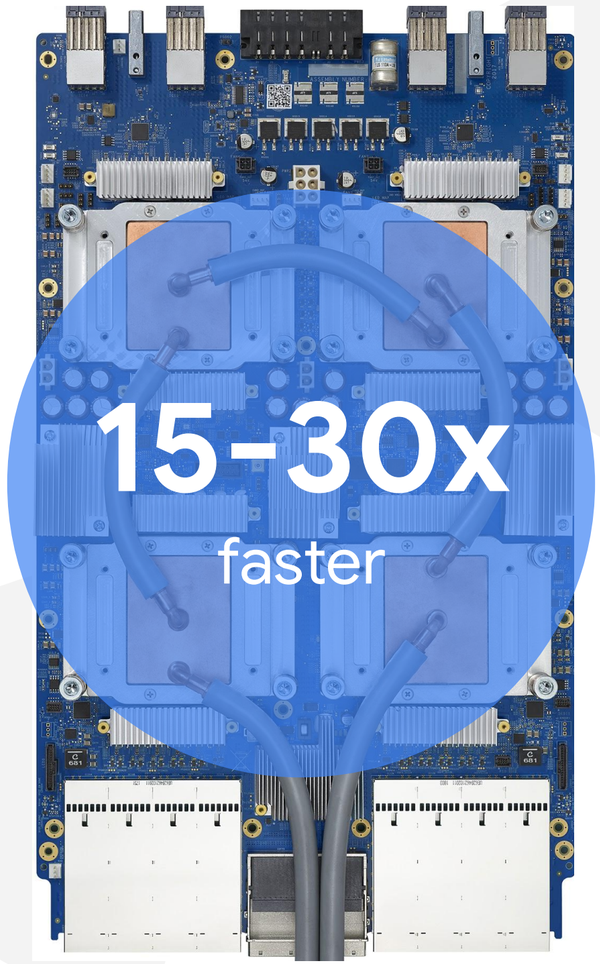
Google Cloud democratizes AI by providing a range of ML and AI solutions that enable businesses to leverage the power of ML and AI, without the traditional costs and efforts.

Google Cloud AI Platform is a unified, simply managed platform that makes machine learning easy to adopt by analysts and developers. It provides modern ML services, with the ability to generate tailored models and use pre-trained models.

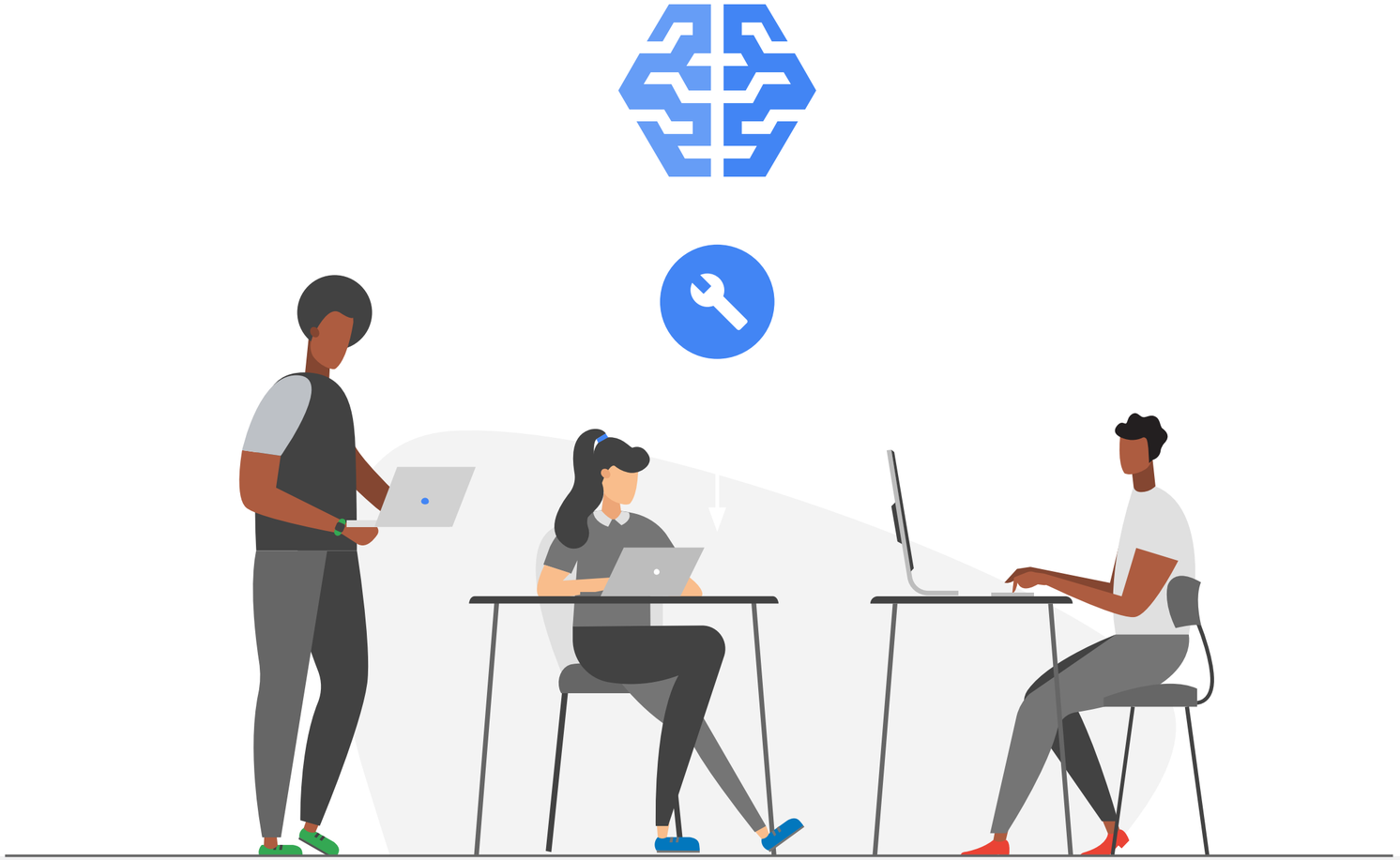


TensorFlow has a comprehensive, flexible ecosystem of tools, libraries and community resources. TensorFlow lets researchers push innovations in ML and developers to easily build and deploy ML powered applications.

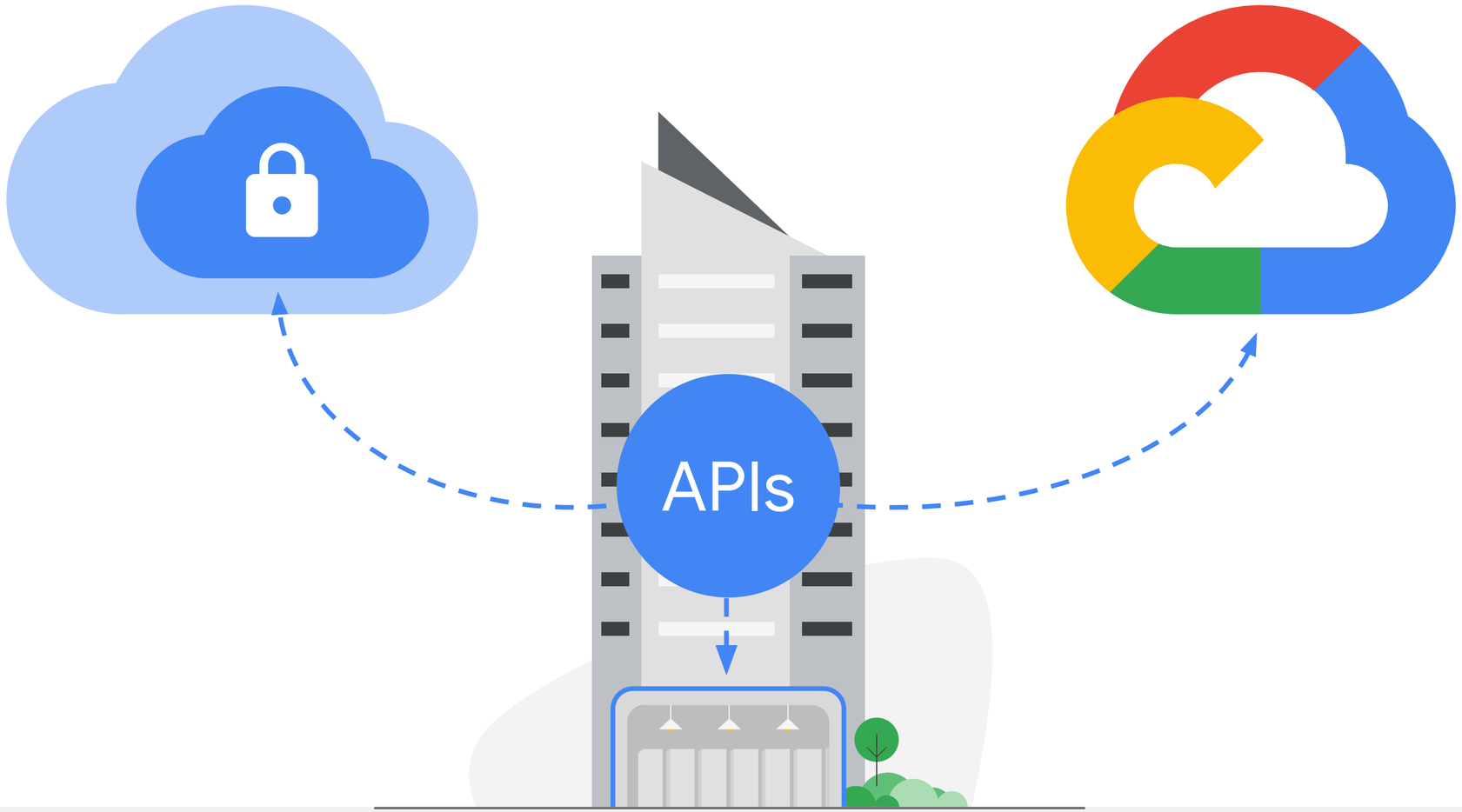
TensorFlow takes advantage of Tensor Processing Units (TPU), hardware devices designed to accelerate ML workloads with TensorFlow by 15-30x. Because you pay only for what you use, there's no up-front capital investment required.



Pay for what you use



The AI Hub is a hosted repository of plug-and-play AI components, including end-to-end AI pipelines and out-of-the-box algorithms.



APIs are simple methods and tools to connect various applications. They can be deployed in a virtual private cloud, on-premises, or in Google's public cloud. They allow developers to quickly and easily train custom models regardless of their level of ML experience.

ML is uniquely placed to create new business value when it can learn from data to automate action and processes, and to customize responses to behavior. The four common business problems that ML is particularly suited to solving are:

1. **Replacing** rule-based systems
2. **Automating** processes
3. **Understanding** unstructured data
4. **Personalizing** applications