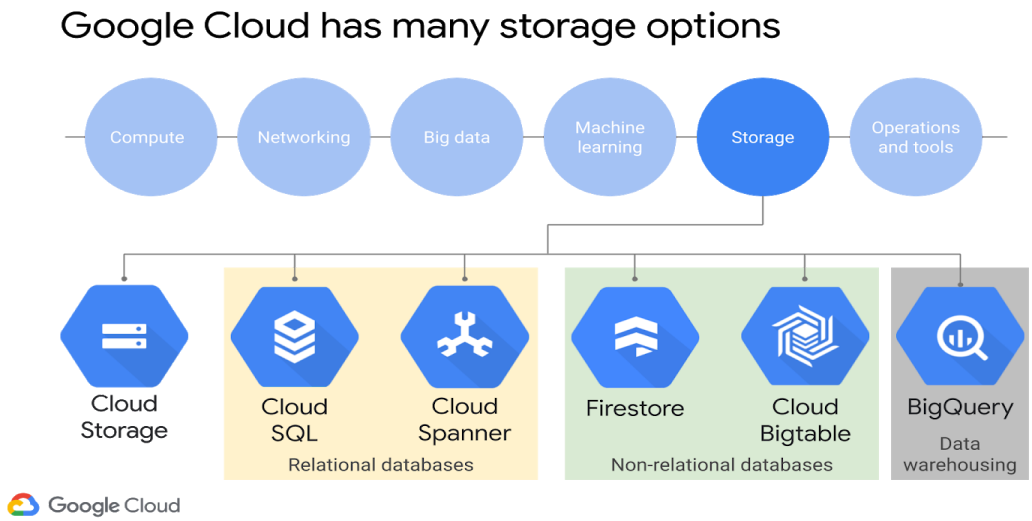
Google Cloud

Storage in the Cloud



Google Cloud has many storage options that satisfy nearly every customer use case. In this module, we turn our attention to the core storage options: Cloud Storage, Cloud SQL, Cloud Spanner, Firestore, and Cloud Bigtable.

Agenda

- Cloud Storage
- Cloud Bigtable
- Cloud SQL and Cloud Spanner
- Firestore
- Comparing Storage Options
- Quiz and Lab
- Resources

Google Cloud

Cloud Storage is binary large-object storage

- High performance, internet-scale.
  - Simple administration.
- Does not require capacity management.
- Data encryption at rest.
- Data encryption in transit by default from Google to endpoint.
- Online and offline import services are available.

Google Cloud

Cloud Storage offers developers and IT organizations durable and highly available object storage. It assesses no minimum fee; you pay only for what you use. Prior provisioning of capacity isn't necessary.

What's object storage? It's not the same as file storage, in which you manage your data as a hierarchy of folders. It's not the same as block storage, in which your operating system manages your data as chunks of disk. Instead, object storage means this: you say to your

storage, “Here, keep this arbitrary sequence of bytes,,” and the storage lets you address it with a unique key. In Cloud Storage and in other systems, these unique keys are in the form of URLs, which means object storage interacts well with web technologies.

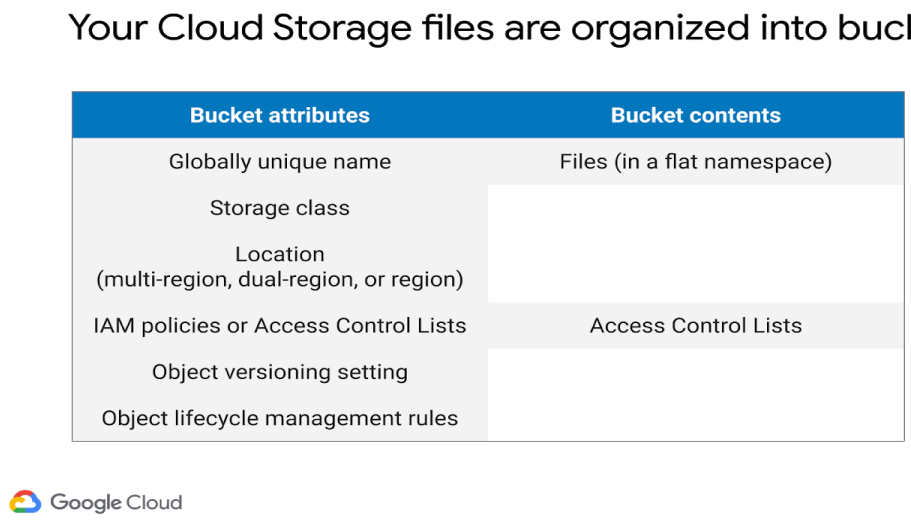
Cloud Storage always encrypts your data on the server side, before it is written to disk, at no additional charge. Data traveling between a customer’s device and Google is encrypted by default using HTTPS/TLS (Transport Layer Security). In fact, Google was the first major cloud provider to enable HTTPS/TLS by default.

Cloud Storage is not a file system, although it can be accessed as one via third-party tools such as Cloud Storage FUSE. The storage objects offered by Cloud Storage are “immutable,” which means that you do not edit them in place, but instead create a new version. Cloud Storage’s primary use is whenever binary large-object storage is needed: online content, backup and archiving, storage of intermediate results in processing workflows, and more.

Offline Media Import/Export is a third-party solution that allows you to load data into Cloud Storage by sending your physical media, such as hard disk drives (HDDs), tapes, and USB flash drives, to a third-party service provider who uploads data on your behalf. Offline Media Import/Export is helpful if you’re limited to a slow, unreliable, or expensive internet connection.

Offline import is available through third-party providers:
https://cloud.google.com/storage/docs/offline-media-import-export
Cloud Storage Transfer Service enables you to import large amounts of online data into Cloud Storage quickly and cost-effectively. To use Cloud Storage Transfer Service, you set up a transfer from a data source to data sink. Data sources can be an Amazon Simple Storage Service (Amazon S3) bucket, an HTTP/HTTPS location, or another Cloud Storage bucket. Data sinks are always a Cloud Storage bucket.

Example uses of Cloud Storage Transfer Service include:
Backing up data to a Cloud Storage bucket from other storage providers.
Moving data from a Standard Storage bucket to a Nearline Storage bucket to lower your storage costs.



Your Cloud Storage files are organized into buckets. When you create a bucket: you give it a globally-unique name; you specify a geographic location where the bucket and its contents are stored; and you choose a default storage class. Pick a location that minimizes latency for your users. For example, if most of your users are in Europe, you probably want to pick a European location: a Google Cloud region in Europe, or else the EU multi-region.

There are several ways to control users' access to your objects and buckets. For most purposes, Cloud IAM is sufficient. Roles are inherited from project to bucket to object. If you need finer control, you can create access control lists ("ACLs") that offer finer control, ACLs define who has access to your buckets and objects, as well as what level of access they have. Each ACL consists of two pieces of information: A scope, which defines who can perform the specified actions (for example, a specific user or group of users). And a permission, which defines what actions can be performed (for example, read or write).

Remember that Cloud Storage objects are immutable. You can turn on object versioning on your buckets if you want. If you do, Cloud Storage keeps a history of modifications--that is, overwrites or deletes--of all objects in the bucket. You can list the archived versions of an object, restore an object to an older state, or permanently delete a version, as needed. If you don't turn on object versioning, new always overwrites old.

Cloud Storage also offers lifecycle management policies. For example, you could tell Cloud Storage to delete objects older than 365 days, or to delete objects created before January 1, 2013; or to keep only the 3 most recent versions of each object in a bucket that has versioning enabled.

Choosing among Cloud Storage classes

Storage Class	Characteristics	Availability	Use Cases	Name for APIs and gsutil
Standard Storage	• Optimized performance when co-located with resources • No minimum storage duration	>99.99% availability in multi-regions and dual-regions; 99.99% in regions	Access data frequently ("hot" data) and/or store for brief periods • Serve website content • Stream videos • Interactive workloads • Mobile and gaming apps	STANDARD
Nearline Storage	• Lower at-rest storage costs • Low-cost data retrieval cost • 30-day min storage duration	99.95% availability in multi-regions and dual-regions; 99.9% in regions	Read/modify data ≤ once per month • Data backup • Serve long-tail multimedia content	NEARLINE
Coldline Storage	• Lower storage cost than Nearline • Higher data retrieval cost • 90-day min storage duration		Read/modify data no more than once a quarter	COLDLINE
Archive Storage	• Lowest storage cost • Highest cost for access/ operations • 365-day min storage duration		Access data less than once a year • Cold data storage • Disaster recovery	ARCHIVE

Google Cloud

Google Cloud Storage has four primary storage classes, with different characteristics, use cases, and prices for your needs.

Standard Storage is best for data that is frequently accessed ("hot" data) and/or stored for only brief periods of time. When used in a region, co-locating your resources maximizes the performance for data-intensive computations and can reduce network charges. When used in a dual-region, you still get optimized performance when accessing Google Cloud products that are located in one of the associated regions, but you also get the improved availability that comes from storing data in geographically separate locations. When used in a multi-region, Standard Storage is appropriate for storing data that is accessed around the world, such as serving website content, streaming videos, executing interactive workloads, or serving data supporting mobile and gaming applications.

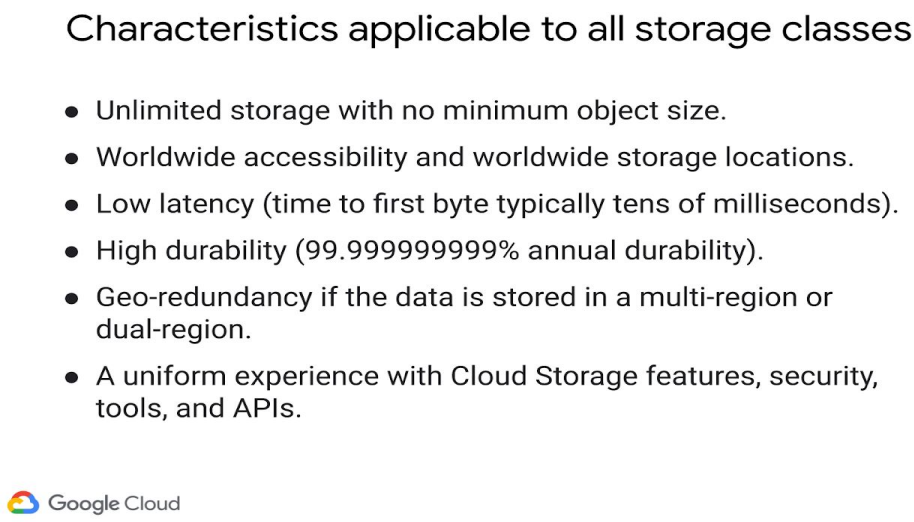
Nearline Storage is a low-cost, highly durable storage service for storing infrequently accessed data. Nearline Storage is a better choice than Standard Storage in scenarios where slightly lower availability, a 30-day minimum storage duration, and costs for data access are acceptable trade-offs for lowered at-rest storage costs. Nearline Storage is ideal for data you plan to read or modify on average once per month or less. Nearline Storage is appropriate for data backup, long-tail multimedia content, and data archiving.

Coldline Storage is a very-low-cost, highly durable storage service for storing infrequently accessed data. Coldline Storage is a better choice than Standard Storage or Nearline Storage in scenarios where slightly lower availability, a 90-day minimum storage duration, and higher costs for data access are acceptable trade-offs for lowered at-rest storage costs. Coldline Storage is ideal for data you plan to read or modify at most once a quarter.

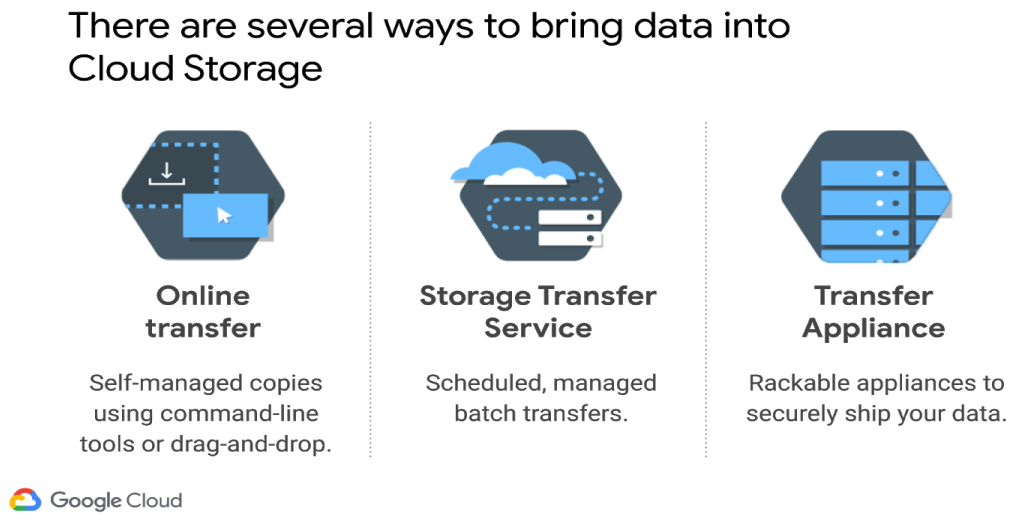
Archive Storage is the lowest-cost, highly durable storage service for data archiving, online

backup, and disaster recovery. Archive Storage has higher costs for data access and operations, as well as a 365-day minimum storage duration. Archive Storage is the best choice for data that you plan to access less than once a year. For example, cold data storage, such as data stored for legal or regulatory reasons, and disaster recovery.

For more information, see: https://cloud.google.com/storage/docs/storage-classes



We've discussed the four primary storage classes and differentiated between them in terms of characteristics, availability and use cases. It is worth noting that there are a number of characteristics that apply across all storage classes. These include:
Unlimited storage with no minimum object size requirement,
Worldwide accessibility and locations,
Low latency and high durability,
Geo-redundancy if data is stored in a multi-region or dual-region, and
A uniform experience, which extends to security, tools, and APIs.

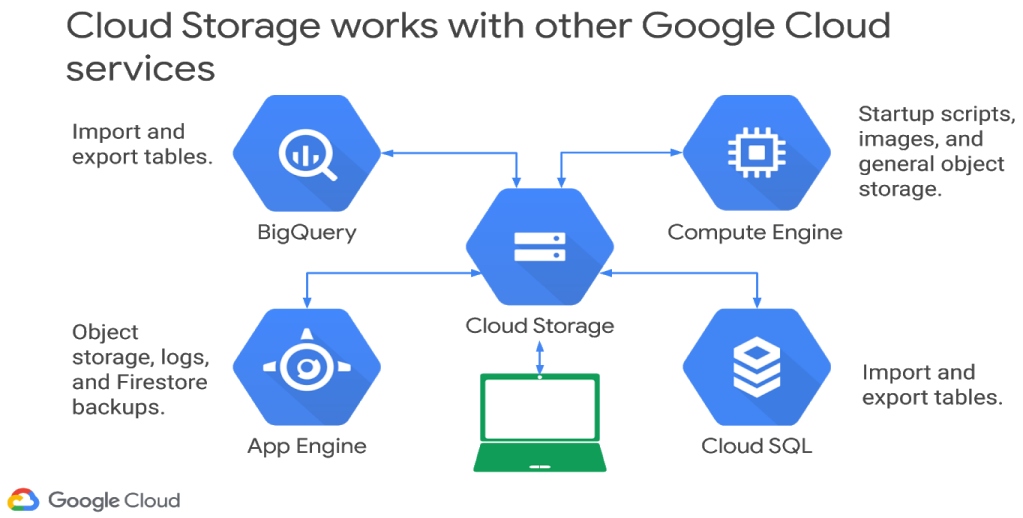


Regardless of which storage class you choose, there are several ways to bring data into Cloud Storage.

Many customers simply use gsutil, which is the Cloud Storage command from the Cloud SDK. You can also move data in with a drag and drop in the Cloud Console, if you use the Google Chrome browser. But what if you have to upload terabytes or even petabytes of data? Google Cloud offers the online Storage Transfer Service and the offline Transfer Appliance to help.

The Storage Transfer Service lets you schedule and manage batch transfers to Cloud Storage from another cloud provider, from a different Cloud Storage region, or from an HTTP(S) endpoint.

The Transfer Appliance is a rackable, high-capacity storage server that you lease from Google Cloud. You simply connect it to your network, load it with data, and then ship it to an upload facility where the data is uploaded to Cloud Storage. The service enables you to securely transfer up to a petabyte of data on a single appliance. As of this recording, it's still beta, and it's not available everywhere, so check the website for details.



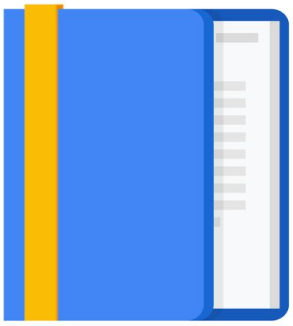
There are other ways of getting your data into Cloud Storage, as this storage option is tightly integrated with many of the Google Cloud products and services.

For example, you can import and export tables from and to BigQuery, as well as Cloud SQL. You can also store App Engine logs, Firestore backups, and objects used by App Engine applications like images. Cloud Storage can also store instance startup scripts, Compute Engine images, and objects used by Compute Engine applications.

In short, Cloud Storage is often the ingestion point for data being moved into the cloud, and is frequently the long-term storage location for data.

Agenda

- Cloud Storage
- **Cloud Bigtable**
- Cloud SQL and Cloud Spanner
- Firestore
- Comparing Storage Options
- Quiz and Lab
- Resources



Google Cloud

Cloud Bigtable is managed NoSQL

- Fully managed NoSQL, wide-column database service for terabyte applications.
- Integrated
  - Accessed using HBase API
  - Native compatibility with big data, Hadoop ecosystems

Google Cloud

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

Why choose Cloud Bigtable?

- Replicated storage.
- Data encryption in-flight and at rest.
- Role-based ACLs.
- Drives major applications such as Google Analytics and Gmail.

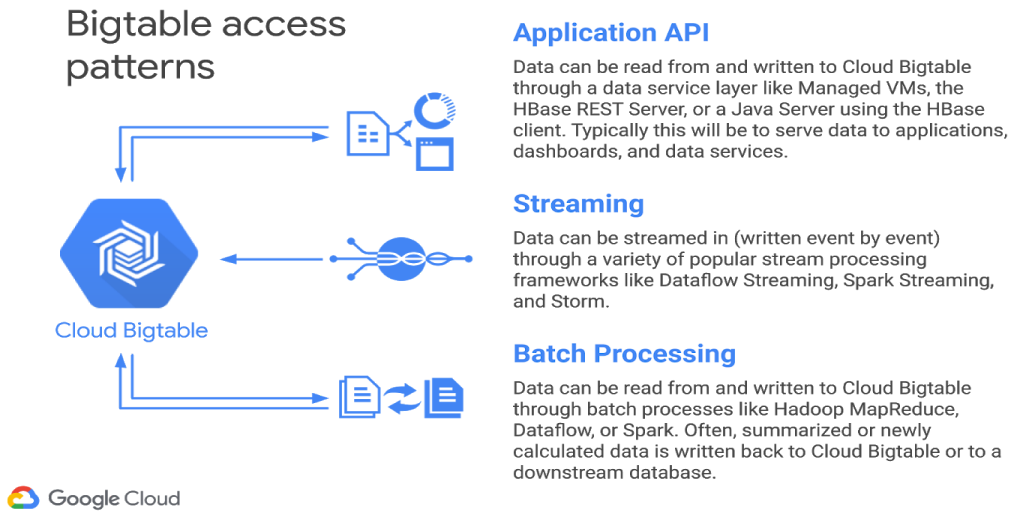
Google Cloud

Customers frequently choose Bigtable if the data is:
Big
Large quantities (>1 TB) of semi-structured or structured data
Fast
Data is high throughput or rapidly changing
NoSQL
Transactions, strong relational semantics not required

And especially if it is:
Time series
Data is time-series or has natural semantic ordering
Big data
You run asynchronous batch or real-time processing on the data
Machine learning
You run machine learning algorithms on the data

Bigtable is designed to handle massive workloads at consistent low latency and high throughput, so it's a great choice for both operational and analytical applications, including IoT, user analytics, and financial data analysis.

For more information on Cloud Bigtable, see https://www.google.com/url?q=https://cloudplatform.googleblog.com/2015/05/introducing-Google-Cloud-Bigtable.html



As Cloud Bigtable is part of the Google Cloud ecosystem, it can interact with other Google Cloud services and third-party clients.

From an application API perspective, data can be read from and written to Cloud Bigtable through a data service layer like Managed VMs, the HBase REST Server, or a Java Server using the HBase client. Typically this will be to serve data to applications, dashboards, and data services.

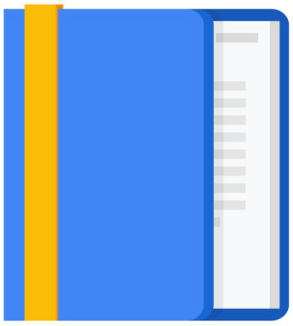
Data can also be streamed in through a variety of popular stream processing frameworks like Dataflow Streaming, Spark Streaming, and Storm.

If streaming is not an option, data can also be read from and written to Cloud Bigtable through batch processes like Hadoop MapReduce, Dataflow, or Spark. Often, summarized or newly calculated data is written back to Cloud Bigtable or to a downstream database.

Agenda

- Cloud Storage
- Cloud Bigtable
- **Cloud SQL and Cloud Spanner**
- Firestore
- Comparing Storage Options
- Quiz and Lab
- Resources

Google Cloud



Cloud SQL is a managed RDBMS

- Offers MySQL, PostgreSQL, and SQL Server databases as a service.
- Automatic replication
- Managed backups
- Vertical scaling (read and write)
- Horizontal scaling (read)
- Google security

Google Cloud

Cloud SQL is an easy-to-use service that delivers fully managed relational databases, including MySQL, PostgreSQL, and SQL Server. Cloud SQL lets you hand off to Google the mundane, but necessary and often time-consuming tasks—like applying patches and updates, managing backups, and configuring replications—so you can put your focus on building great applications.

Every Cloud SQL instance includes a network firewall, allowing you to control network access to your database instance by granting access.

Cloud SQL is easy to use: it doesn't require any software installation or maintenance.

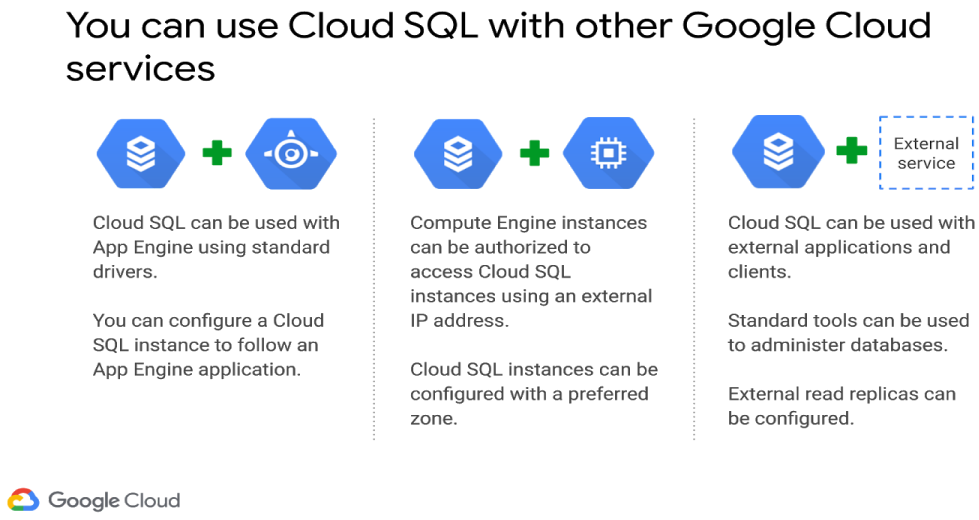
Easily scale up to 64 processor cores and more than 400 GB of RAM and 30 TB of storage. Quickly scale out with read replicas.

Automatic replication
Google Cloud SQL supports the following read replica scenarios:
Cloud SQL instances replicating from a Cloud SQL master instance Replicas are other instances in the same project and location as the master instance.
Cloud SQL instances replicating from an external master instance The master instance is external to Cloud SQL. For example, it can be outside the Google network or in a Compute Engine instance.
External MySQL instances replicating from a Cloud SQL master instance External replicas are in hosting environments, outside of Cloud SQL.

Managed backups
Cloud SQL takes care of securely storing your backed-up data and makes it easy for you to restore from a backup and perform a point-in-time recovery to a specific state of an instance. Cloud SQL retains up to 7 backups for each instance, which is included in the cost of your instance.

Cloud SQL customer data is encrypted when on Google's internal networks and when stored in database tables, temporary files, and backups.

(MySQL instances are available in either First Generation or Second Generation. Google recommends the use of Second Generation instances for most use cases. First Generation instances are recommended primarily when MySQL 5.5 compatibility is required. Also, First Generation instances may be cost-effective for infrequently used or test/dev database instances, because of their available Per-Use billing plan and the available ON DEMAND activation policy, which causes your instance to automatically shut itself off after 15 minutes of inactivity.)



Another benefit of Cloud SQL instances is that they are accessible by other Google Cloud services and even external services. You can use Cloud SQL with App Engine using standard drivers like Connector/J for Java or MySQLdb for Python.

You can authorize Compute Engine instances to access Cloud SQL instances and configure the Cloud SQL instance to be in the same zone as your virtual machine.

Cloud SQL also supports other applications and tools that you might be used to, like SQL Workbench, Toad and other external applications using standard MySQL drivers.

Cloud Spanner is a horizontally scalable RDBMS

Cloud Spanner supports:

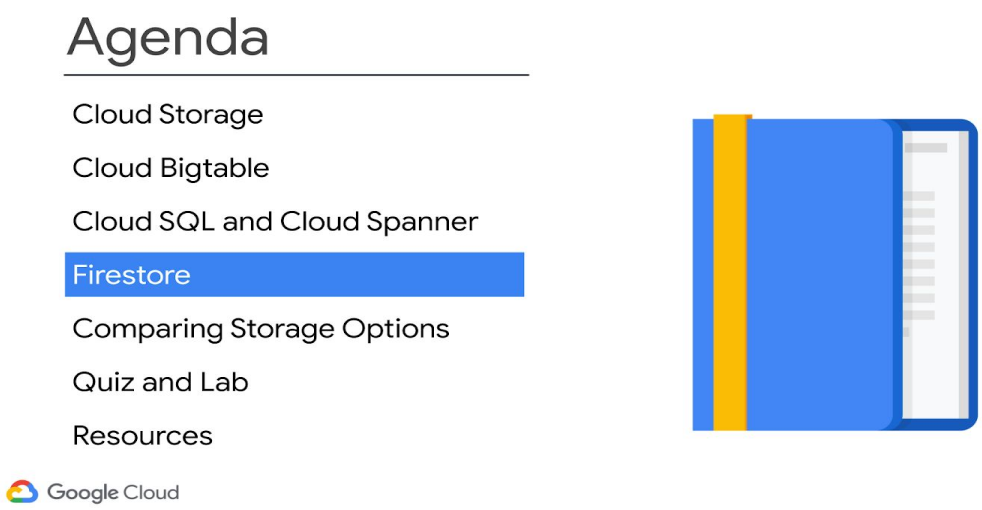
- Automatic replication.
- Strong global consistency.
- Managed instances with high availability.
- SQL (ANSI 2011 with extensions).

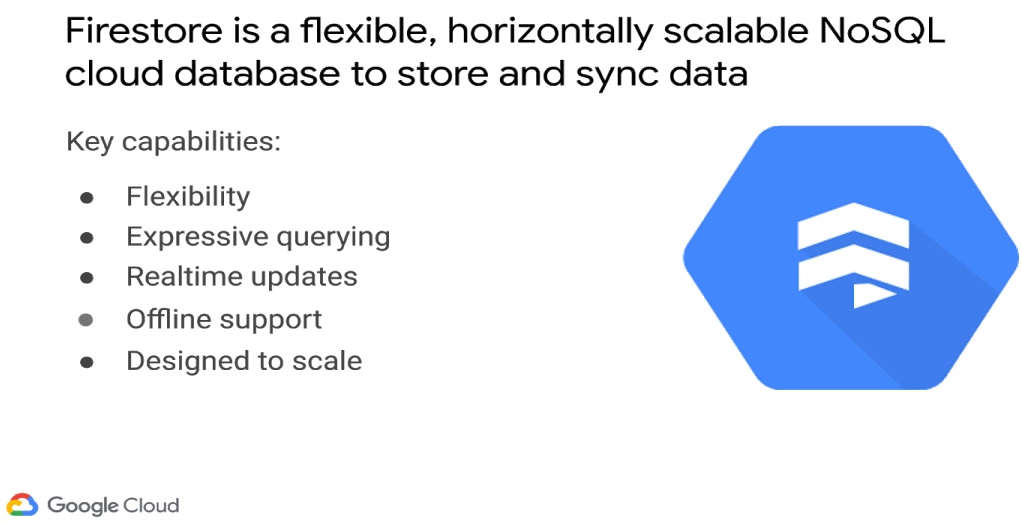
Google Cloud

Cloud Spanner supports strong consistency, including strongly consistent secondary indexes, SQL, and managed instances with high availability through synchronous and built-in data replication. Battle tested by Google's own mission-critical applications and services, Spanner powers Google's $80 billion business.

Cloud Spanner is especially suited for applications requiring:
A SQL RDBMS, with joins and secondary indexes
Built-in high availability
Strong global consistency
Database sizes exceeding ~2 TB
Many IOPS (Tens of thousands of reads/writes per second or more)

For a technical overview of Cloud Spanner, see https://cloudplatform.googleblog.com/2017/02/inside-Cloud-Spanner-and-the-CAP-Theorem.html.





Firestore is a flexible, horizontally scalable database in the cloud for mobile, web, and server development.

The Firestore data model supports flexible, hierarchical data structures. You store data in documents, organized into collections. Documents can contain complex nested objects in addition to subcollections.

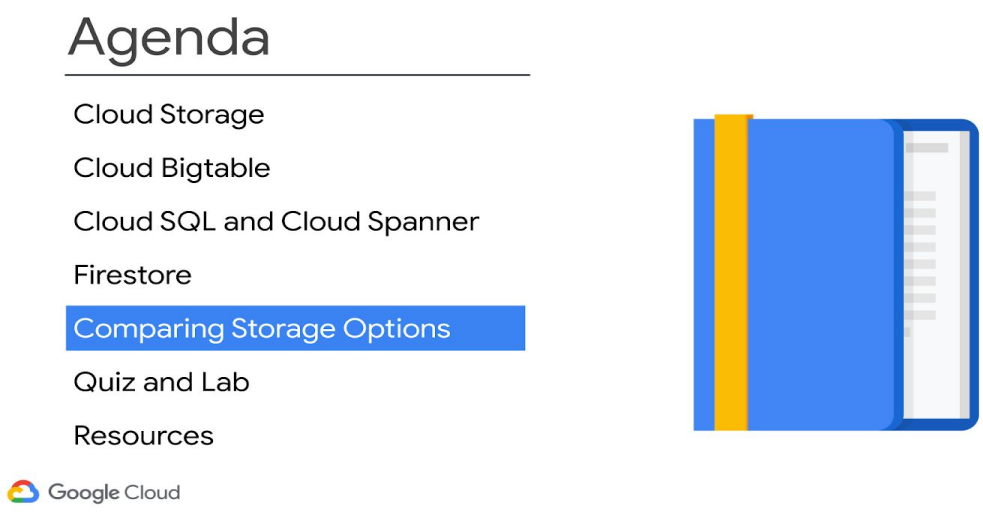
You can use queries to retrieve individual, specific documents or to retrieve all the documents in

a collection that match your query parameters. Queries can include multiple, chained filters and combine filtering and sorting. They're also indexed by default, so query performance is proportional to the size of the result set, not the dataset.

Firestore uses data synchronization to update data on any connected device. However, it's also designed to make simple, one-time fetch queries efficiently.

Firestore caches data that an app is actively using, so the app can write, read, listen to, and query data even if the device is offline. When the device comes back online, Firestore synchronizes any local changes back to Firestore.

Firestore leverages Google Cloud's powerful infrastructure: automatic multi-region data replication, strong consistency guarantees, atomic batch operations, and real transaction support.



Comparing storage options: technical details

	Firestore	Cloud Bigtable	Cloud Storage	Cloud SQL	Cloud Spanner	BigQuery
Type	NoSQL document	NoSQL wide column	Blobstore	Relational SQL for OLTP	Relational SQL for OLTP	Relational SQL for OLAP
Transactions	Yes	Single-row	No	Yes	Yes	No
Complex queries	No	No	No	Yes	Yes	Yes
Capacity	Terabytes+	Petabytes+	Petabytes+	Up to ~10 TB	Petabytes	Petabytes+
Unit size	1 MB/entity	~10 MB/cell ~100 MB/row	5 TB/object	Determined by DB engine	10,240 MiB/ row	10 MB/row

Google Cloud

Now that we covered Google Cloud's core storage options, let's compare them to help you choose the right service for your application or workflow.

This table focuses on the technical differentiators of the storage services. Each row is a technical specification and each column is a service. Let me cover each service from left to right.

Consider Firestore if you need massive scaling and predictability together with real time query results and offline query support. This storage services provides terabytes of capacity with a maximum unit size of 1 MB per entity.

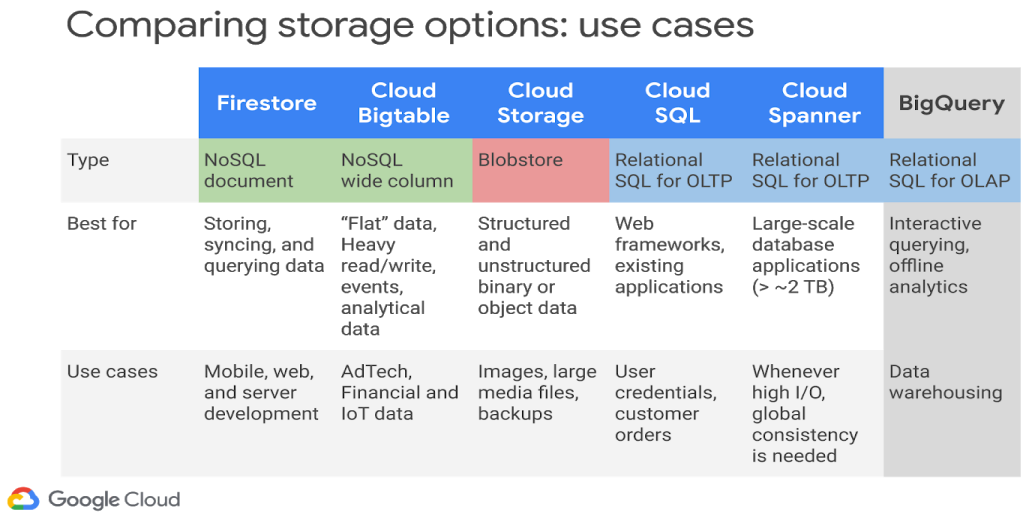
Consider using Cloud Bigtable if you need to store a large amount of structured objects. Cloud Bigtable does not support SQL queries, nor does it support multi-row transactions. This storage service provides petabytes of capacity with a maximum unit size of 10 MB per cell and 100 MB per row.

Consider using Cloud Storage if you need to store immutable blobs larger than 10 MB, such as large images or movies. This storage service provides petabytes of capacity with a maximum unit size of 5 TB per object.

Consider using Cloud SQL or Cloud Spanner if you need full SQL support for an online transaction processing system. Cloud SQL provides up to Up to 10,230 GB, depending on machine type, while Cloud Spanner provides petabytes. If Cloud SQL does not fit your requirements because you need horizontal scalability, not just through read replicas, consider using Cloud Spanner.

We didn't cover BigQuery in this module as it sits on the edge between data storage and data

processing, but you will learn more about it in the “Big Data and Machine Learning in the Cloud” module. The usual reason to store data in BigQuery is to use its big data analysis and interactive querying capabilities. You would not want to use BigQuery, for example, as the backing store for an online application.



Considering the technical differentiators of the different storage services helps some people decide which storage service to choose, others like to consider use cases. Let me go through each service one more time.

Firestore is best for storing, syncing, and querying data for mobile and web apps.
Bigtable is best for analytical data with heavy read and write events, like AdTech, financial or IoT data.
Cloud Storage is best for structured and unstructured binary or object data, like images, large media files and backups.
Cloud SQL is best for web frameworks and existing applications, like storing user credentials and customer orders.
Cloud Spanner is best for large-scale database applications that are larger than 2 TB. For example, for financial trading and e-commerce use cases.

As I mentioned at the beginning of the module, depending on your application you might use one or several of these services to get the job done.

Agenda

- Cloud Storage
- Cloud Bigtable
- Cloud SQL and Cloud Spanner
- Firestore
- Comparing Storage Options
- Quiz and Lab
- Resources

Quiz 1

Your application transcodes large video files. Which storage service should you consider first?

Quiz 1

Your application transcodes large video files. Which storage service should you consider first?

Cloud Storage

Google Cloud

Quiz 2

You stream huge amounts of data from devices with sensors. Which storage service should you consider first?

Google Cloud

Quiz 2

You stream huge amounts of data from devices with sensors. Which storage service should you consider first?

Cloud Bigtable

Google Cloud

Lab

Getting Started With Cloud Storage and Cloud SQL

Duration: 50 minutes

In this lab you will store an image in a Cloud Storage bucket and configure Compute Engine to use a Cloud SQL database to reference the image.

Lab Objectives

Create a Cloud Storage bucket and place an image into it.

Create a Cloud SQL instance and configure it.

Connect to a Cloud SQL instance from a web server.

Use an image stored in a Cloud Storage bucket in a web page.

Google Cloud

In this lab you will create a Google Cloud Storage bucket and place an image in it. You'll also configure an application running in Google Compute Engine to use a database managed by Google Cloud SQL and to reference the image in the Cloud Storage bucket.

Agenda

Cloud Storage

Cloud Bigtable

Cloud SQL and Cloud Spanner

Firestore

Comparing Storage Options

Quiz and Lab

Resources

Google Cloud

Resources

Overview of Cloud Storage https://cloud.google.com/storage/

Getting started with Cloud SQL https://cloud.google.com/sql/docs/quickstart

Cloud Bigtable https://cloud.google.com/stackdriver/docs/

Cloud Spanner https://cloud.google.com/spanner/docs/

Firestore https://firebase.google.com/docs/firestore