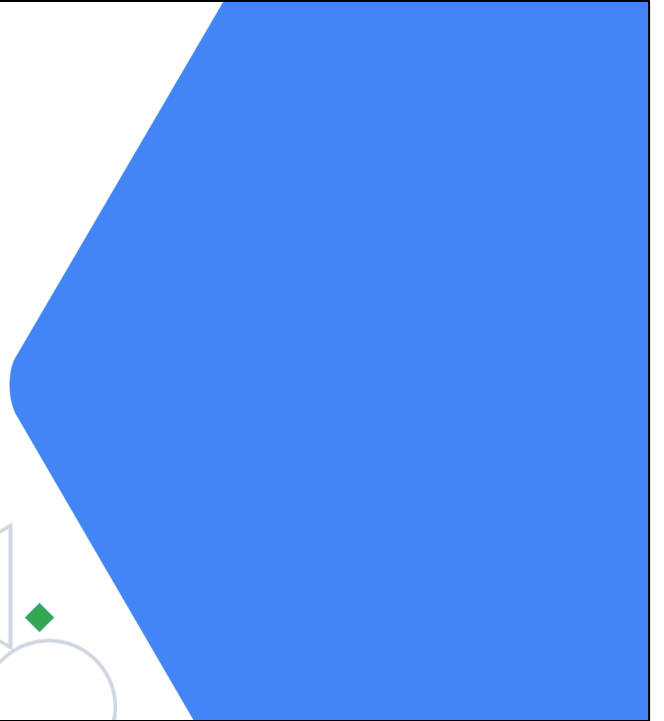




# Big Data and Machine Learning in the Cloud



---

# Agenda

Google Cloud Big Data Platform

Machine Learning in the Cloud

Quiz and Lab

Resources



## Google Cloud's big data services are fully managed and scalable



Dataproc

Managed Hadoop MapReduce, Spark, Pig, and Hive service



Dataflow

Stream & batch processing; unified and simplified pipelines



BigQuery

Analytics database; stream data at 100,000 rows per second



Pub/Sub

Scalable & flexible enterprise messaging



AI Platform Notebooks

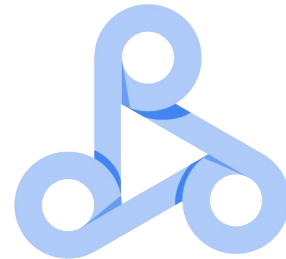
Interactive data exploration



Google Cloud Big Data solutions are designed to help you transform your business and user experiences with meaningful data insights. It is an integrated, serverless platform. “Serverless” means you don’t have to provision compute instances to run your jobs. The services are fully managed, and you pay only for the resources you consume. The platform is “integrated” so Google Cloud data services work together to help you create custom solutions.

## Dataproc is managed Hadoop

- Fast, easy, managed way to run Hadoop and Spark/Hive/Pig on Google Cloud.
- Create clusters in 90 seconds or less on average.
- Scale clusters up and down even when jobs are running.

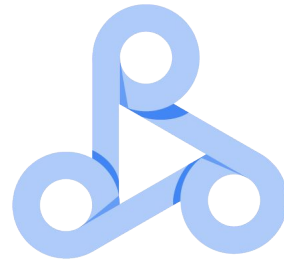


Apache Hadoop is an open-source framework for big data. It is based on the MapReduce programming model, which Google invented and published. The MapReduce model, at its simplest, means that one function -- traditionally called the “map” function -- runs in parallel across a massive dataset to produce intermediate results; and another function -- traditionally called the “reduce” function -- builds a final result set based on all those intermediate results. The term “Hadoop” is often used informally to encompass Apache Hadoop itself and related projects, such as Apache Spark, Apache Pig, and Apache Hive.

Dataproc is a fast, easy, managed way to run Hadoop, Spark, Hive, and Pig on Google Cloud. All you have to do is to request a Hadoop cluster. It will be built for you in 90 seconds or less, on top of Compute Engine virtual machines whose number and type you can control. If you need more or less processing power while your cluster's running, you can scale it up or down. You can use the default configuration for the Hadoop software in your cluster, or you can customize it. And you can monitor your cluster using Cloud Monitoring.

## Why use Dataproc?

- Easily migrate on-premises Hadoop jobs to the cloud.
- Quickly analyze data (like log data) stored in Cloud Storage; create a cluster in 90 seconds or less on average, and then delete it immediately.
- Use Spark/Spark SQL to quickly perform data mining and analysis.
- Use Spark Machine Learning Libraries (MLlib) to run classification algorithms.



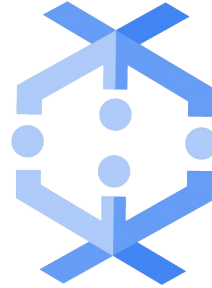
Running on-premises Hadoop jobs requires a hardware investment. On the other hand, running these jobs in Dataproc allows you to pay only for hardware resources during the life of the ephemeral cluster you create.

You can further save a significant amount of money using [preemptible Compute Engine instances for batch processing](#). You have to make sure that your jobs can be restarted cleanly if they're terminated and you get a significant break in the cost of the instances. Be aware that the cost of the Compute Engine instances isn't the only component of the cost of a Dataproc cluster, but it's a significant one.

Once your data is in a cluster, you can use Spark and Spark SQL to do data mining, and you can use MLlib, which is Apache Spark's Machine Learning Libraries, to discover patterns through machine learning.

## Dataflow offers managed data pipelines

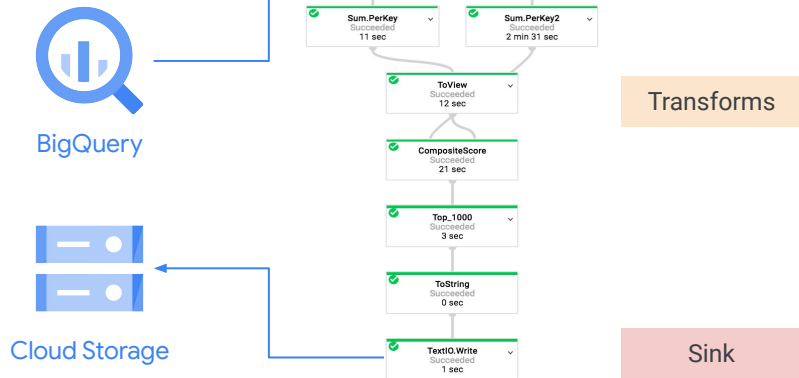
- Processes data using Compute Engine instances.
  - Clusters are sized for you.
  - Automated scaling, no instance provisioning required.
- Write code once and get batch and streaming.
  - Transform-based programming model.



Dataproc is great when you have a dataset of known size, or when you want to manage your cluster size yourself. But what if your data shows up in realtime? Or it's of unpredictable size or rate? That's where Dataflow is a particularly good choice. It's both a unified programming model and a managed service, and it lets you develop and execute a big range of data processing patterns: extract-transform-and-load, batch computation, and continuous computation. You use Dataflow to build data pipelines, and the same pipelines work for both batch and streaming data.

Dataflow is a unified programming model and a managed service for developing and executing a wide range of data processing patterns including ETL, batch computation, and continuous computation. Dataflow frees you from operational tasks like resource management and performance optimization.

Dataflow pipelines flow data from a source through transforms

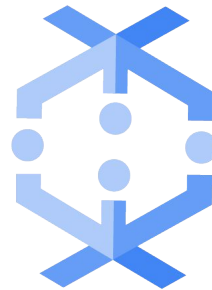


This example Dataflow pipeline reads data from a BigQuery table (the “source”), processes it in various ways (the “transforms”), and writes its output to Cloud Storage (the “sink”). Some of those transforms you see here are map operations, and some are reduce operations. You can build really expressive pipelines.

Each step in the pipeline is elastically scaled. There is no need to launch and manage a cluster. Instead, the service provides all resources on demand. It has automated and optimized work partitioning built in, which can dynamically rebalance lagging work. That reduces the need to worry about “hot keys” -- that is, situations where disproportionately large chunks of your input get mapped to the same cluster.

## Why use Dataflow?

- *ETL* (extract/transform/load) pipelines to move, filter, enrich, shape data.
- *Data analysis*: batch computation or continuous computation using streaming.
- *Orchestration*: create pipelines that coordinate services, including external services.
- Integrates with Google Cloud services like Cloud Storage, Pub/Sub, BigQuery, and Cloud Bigtable.
  - Open source Java and Python SDKs.



People use Dataflow in a variety of use cases. For one, it serves well as a general-purpose ETL tool.

And its use case as a data analysis engine comes in handy in things like these: fraud detection in financial services; IoT analytics in manufacturing, healthcare, and logistics; and clickstream, Point-of-Sale, and segmentation analysis in retail.

And, because those pipelines we saw can orchestrate multiple services, even external services, it can be used in real time applications such as personalizing gaming user experiences.



## BigQuery is a fully managed data warehouse

- Provides near real-time interactive analysis of massive datasets (petabyte-scale).
- No cluster maintenance is required.
- Query using SQL syntax (SQL 2011).
- Flexible data ingest options.
- Wide range of use cases.



If, instead of a dynamic pipeline, you want to do ad-hoc SQL queries on a massive dataset. That is what BigQuery is for. BigQuery is Google's fully-managed, petabyte-scale, low-cost analytics data warehouse.

Because there's no infrastructure to manage, you can focus on analyzing data to find meaningful insights, use familiar SQL and take advantage of a pay-as-you-go model.

It's easy to get data into BigQuery. You can load it from Cloud Storage or Datastore, or stream it into BigQuery at up to 100,000 rows per second. Once it's in there, you can run super-fast SQL queries against multiple terabytes of data in seconds using the processing power of Google's infrastructure.

In addition to SQL queries, you can easily read and write data in BigQuery via Dataflow, Hadoop, and Spark.

BigQuery is used by all types of organizations from startups to Fortune 500 companies - smaller organizations like BigQuery's free monthly quotas, bigger organizations like its seamless scale, and it's available 99.9 percent service level agreement.

## BigQuery runs on Google's high-performance infrastructure

- Global availability.
- Compute and storage are separated with a terabit network in between. You only pay for storage and processing used.
- You have full control over who has access to the data.
- Automatic discount for long-term data storage.



Google's infrastructure is global and so is BigQuery. BigQuery lets you specify the region where your data will be kept. So, for example, if you want to keep data in Europe, you don't have to go set up a cluster in Europe. Just specify the EU location where you create your data set. US and Asia locations are also available. Data replication in multiple geographies also means that your data is available and durable even in the case of extreme failure modes.

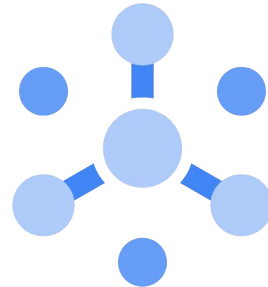
Because BigQuery separates storage and computation, you pay for your data storage separately from queries. That means, you pay for queries only when they are actually running.

You have full control over who has access to the data stored in BigQuery, including sharing datasets with people in different projects. Sharing datasets won't impact your cost or performance with people you share with paying for their own queries.

Long-term storage pricing has an automatic discount for data residing in BigQuery for extended periods of time. When the age of your data reaches 90 days in BigQuery, the price of storage automatically drops.

## Pub/Sub is scalable, reliable messaging

- Foundation for stream analytics.
- Supports many-to-many asynchronous messaging.
- Replicated storage.
- Push and pull delivery.
- Message data is encrypted.



Whenever you're working with events in real time, it helps to have a messaging service. That's what Pub/Sub is. It's meant to serve as a simple, reliable, scalable foundation for stream analytics. The Pub in Pub/Sub is short for publishers and Sub is short for subscribers. Applications can publish messages in Pub/Sub and one or more subscribers receive them.

You can use Pub/Sub to let independent applications you build send and receive messages. That way they're decoupled, so they scale independently. Up to 10,000 messages can be sent per second by default, and millions per second and beyond upon request. Receiving messages doesn't have to be synchronous. That's what makes Pub/Sub great for decoupling systems.

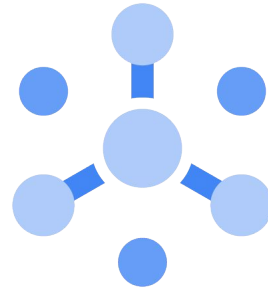
Pub/Sub is designed to provide "at least once" delivery at low latency by storing every message on multiple servers in multiple zones. When we say "at least once delivery," we mean that there is a small chance some messages might be delivered more than once. So, keep this in mind when you write your application.

You can configure your subscribers to receive messages on a push or pull basis, whether they are accessible from the internet or behind a firewall. In other words, subscribers can get notified when new messages arrive for them or they can check for new messages at intervals.

To ensure data security and protection, all message data on the wire and at rest is encrypted.

## Why use Pub/Sub?

- Building block for data ingestion in Dataflow, Internet of Things (IoT), Marketing Analytics.
- Foundation for Dataflow streaming.
- Connect applications across Google Cloud (push/pull between Compute Engine and App Engine).



Pub/Sub builds on the same technology Google uses internally. It's an important building block for applications where data arrives at high and unpredictable rates, like Internet of Things systems.

If you're analyzing streaming data, Dataflow is a natural pairing with Pub/Sub.

Pub/Sub also works well with applications built on Google Cloud's compute platforms.

## AI Platform Notebooks is a notebook service to get your projects up and running in minutes

- Managed JupyterLab experience.
- Secure development and controlled user access.
- Advanced networking.
- Support for data science frameworks and optimized for machine learning.
- Git support.
- Bring your own container.



AI Platform Notebooks is a managed service that offers an integrated and secure JupyterLab environment for data scientists and machine learning developers to experiment, develop, and deploy models into production. You can create instances running JupyterLab that come pre-installed with the latest data science and machine learning frameworks in a single click.

AI Platform Notebooks is built on the industry standard JupyterLab, so you can use it with the RPython and R data science community and customize your environment by installing JupyterLab plugins.

AI Platform Notebooks supports popular enterprise security architectures through VPC-SC, shared VPC, and private IP controls. You can also encrypt your data on disk with CMEK. You can choose between two predefined user access modes: restrict AI Platform Notebooks to a single-user or use a service account. You can also customize access based on your enterprise security architecture based on Cloud Identity and Access Management.

You can select any virtual private cloud for your AI Platform Notebook instances, provided that they have access either through Google Private Access or the internet to Cloud Storage. You can also turn off public IP address and access your instance via proxy.

Google provides a pre-configured environment that supports the most popular data science libraries. Optimized versions of TensorFlow and PyTorch enable you to get

the most out of Google Cloud hardware and seamlessly add and remove GPUs from your instance.

It's easy to pull and push notebooks from your Git repository, making it also easy to share your notebooks with colleagues.

You can also run an AI Platform Notebook instance on a container of your choice. This provides you the flexibility to install specific libraries mandated by your organization or preconfigure the environment running JupyterLab to your preference.

## Why use AI Platform Notebooks?

- Get up and running fast. Deploy new JupyterLab instances with one click.
  - Instances are preconfigured with optimized versions of popular data science and ML libraries.
- Scale on demand.
- Seamless experience.



You can deploy new JupyterLab instances with one click and start analyzing your data immediately. Each instance comes pre-configured with optimized versions of the most popular data science and machine learning libraries including TensorFlow, Keras, PyTorch, fast.ai, RAPIDS, NumPy, scikit-learn, pandas, and Matplotlib.

You can start small and scale up by adding CPUs, RAM, and GPUs. When your data gets too big for one machine, you can seamlessly switch to distributed services like BigQuery, Dataproc, Dataflow, and AI Platform Training and Prediction. You pay for the instances only while they are running.

You can go from data to a deployed machine learning model without leaving AI Platform Notebooks. You can pull data from BigQuery, use Dataproc to transform it, and leverage AI Platform services or Kubeflow for distributed training and online prediction.

---

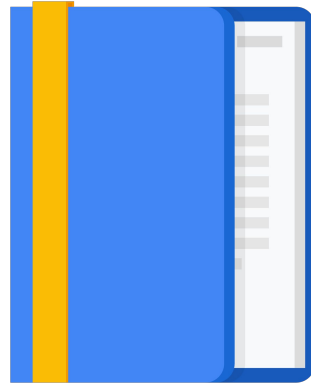
# Agenda

Google Cloud Big Data Platform

Machine Learning in the Cloud

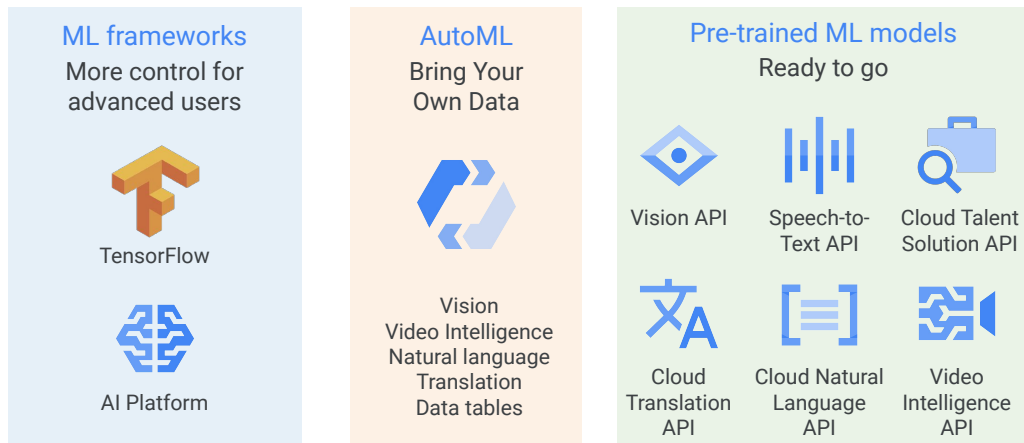
Quiz and Lab

Resources





## The Google Cloud machine learning spectrum



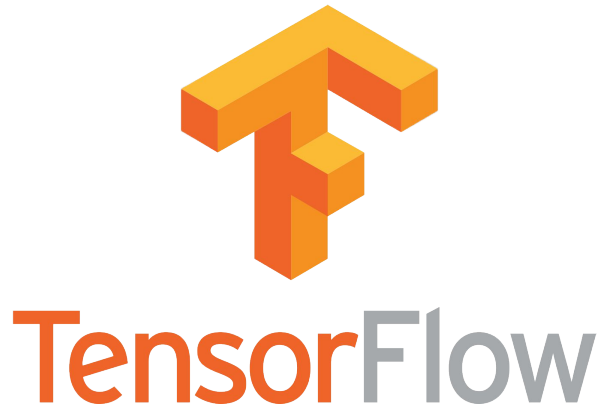
Different options exist when it comes to leveraging machine learning. Advanced users, who want more control over the building and training of ML models, will use tools that offer the levels of flexibility they are looking for. This would involve developing custom models through an ML library like TensorFlow, that's supported on AI Platform. This option works for data scientists with the skills and the need to create a TensorFlow model.

But increasingly, you don't have to do that. Google makes the power of ML available to you even if you have a limited knowledge of ML. You can use AutoML to build on Google's ML capabilities to create your own custom ML models that are tailored to specific business needs, and then integrate those models into applications and web sites.

Alternatively, Google has a range of pre-trained ML models that are ready for immediate use within applications in ways that the respective APIs are designed to support. Such pretrained models are excellent ways to replace user input with ML.

---

Create custom ML models with TensorFlow



As a starting point, let's talk a little bit about TensorFlow. TensorFlow is an open-source high-performance library for numerical computation. Not just about machine learning. Any numeric computation. In fact, people have used TensorFlow for all kinds of GPU computing; for example, you can use TensorFlow to solve partial differential equations -- these are useful in domains like fluid dynamics. TensorFlow as a numeric programming library is appealing because you can write your computation code in a high-level language -- Python -- and have it be executed in a fast way.

## Machine learning for the masses!

### ML frameworks

More control for advanced users



TensorFlow



AI Platform

### AutoML

Bring Your Own Data



Vision  
Video Intelligence  
Natural language  
Translation  
Data tables

### Pre-trained ML models

Ready to go



Vision API



Speech-to-Text API



Cloud Talent Solution API



Cloud Translation API



Cloud Natural Language API



Video Intelligence API



AutoML is a suite of ML products that enables users with limited ML expertise to train high-quality models specific to their business needs. AutoML leverages more than 10 years of proprietary Google Research technology to help users' ML models achieve faster performance and more accurate predictions.

## What's required to solve an ML problem?

Training data



Model code



Training & serving  
infrastructure



Time



To put AutoML into context, let's look at what it takes to solve an ML problem. To solve an ML problem without the benefit of a managed service, it's up to you to wrangle your data, code your model, and put together the infrastructure. This can be prohibitively complex and time consuming.

## What's required when using a managed service?

Training data



Model code



Training & serving infrastructure



Time



Earlier, we discussed how AI Platform lets developers and data scientists build and run superior learning models in production. As reflected in this graphic, there's a considerable reduction in the required training and serving infrastructure as well as the amount of model code. However, there's still a requirement to provide extensive training data, and the process is still a time-consuming one.

## What's required when using AutoML?

Training data



Model code



Training & serving infrastructure

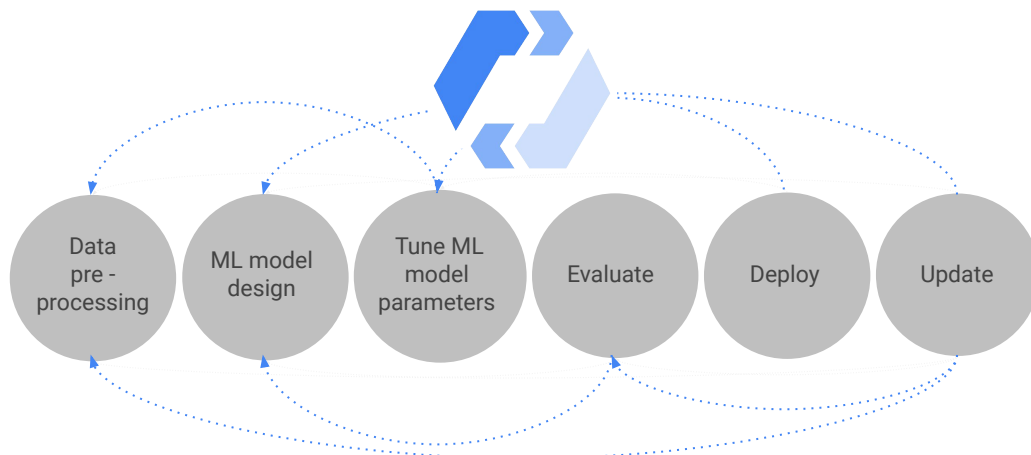


Time



What's immediately notable with AutoML is that there's no requirement on the users' side to develop a model or provide a training and serving infrastructure. In addition, far less training data is required, and results are achieved a lot faster.

## AutoML simplifies the process



 Google Cloud

The ability of AutoML to efficiently solve an ML problem is largely due to how it simplifies these complex steps that are associated with custom ML model building.

## Use AutoML for what you can see



### AutoML Vision

Derive insights from images in the cloud or at the edge.



### AutoML Video Intelligence

Enable powerful content discovery and engaging video experiences.



There are two AutoML products that apply to what you can see.

With AutoML Vision, you simply upload images and train custom image models through an easy-to-use graphical interface. You can optimize your model for accuracy, latency, and size. AutoML Vision Edge allows you to export your custom trained models to an application in the cloud, or to an array of devices at the edge. You can train models to classify images through labels you choose. Alternatively, Google's data labeling service allows you to use their team to help annotate your images, videos, and text. Later, we'll complete a lab where we'll use AutoML Vision to train a custom model to recognize different types of clouds.

AutoML Video Intelligence makes it easy to train custom models to classify and track objects within videos. It's ideal for projects that require custom entity labels to categorize content which aren't covered by the pre-trained Video Intelligence API.



## Use AutoML for what you can hear



**AutoML Natural Language**  
Reveal the structure and meaning  
of text through machine learning.



**AutoML Translation**  
Dynamically translate between  
languages.



There are also two AutoML products that apply to what you can hear.

With AutoML Natural Language, you can train custom ML models to classify, extract, and detect sentiment. This allows you to identify entities within documents and label them based on your own domain-specific keywords or phrases. The same applies to being able to understand the overall opinion, feeling, or attitude expressed in a block of text that's tuned to domain-specific sentiment scores.

AutoML Translation allows you to upload translated language pairs and it will train a custom model which translation queries return results specific to your domain, and that you can scale and adapt to meet your needs.

## Use AutoML to turn structured data into predictive insights



### AutoML Tables

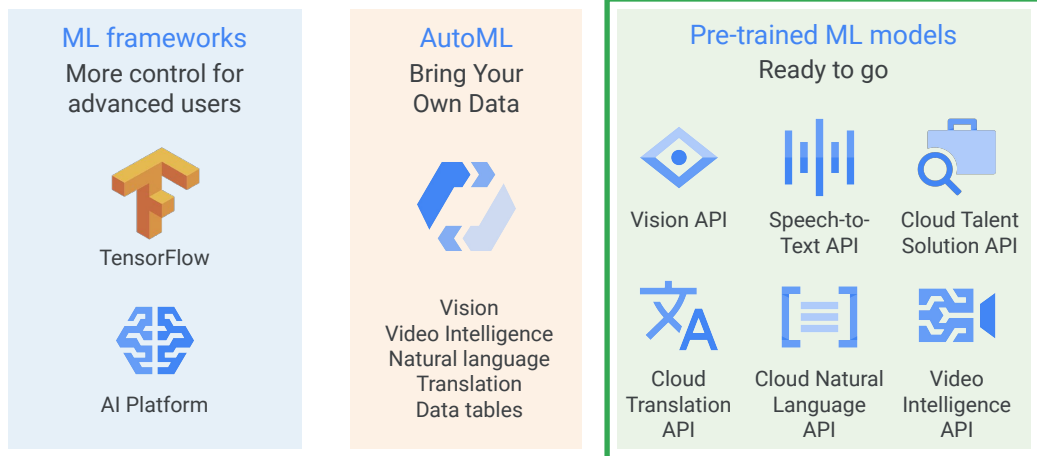
Automatically build and deploy state-of-the-art machine learning models on structured data.



AutoML Tables reduces the time it takes to go from raw data to top-quality, production-ready machine learning models from months to just a few days.

There are many different use cases for AutoML Tables. For example, if you're in retail, you can better predict customer demand so you can preemptively fill gaps and maximize your revenue by optimizing product distribution, promotions, and pricing. In insurance, you could foresee and optimize a policyholder portfolio's risk and return by zeroing in on the potential for large claims and likelihood of fraud. In marketing, you can better understand your customer. For example, What's your average customer's lifetime value? You can make the most of marketing spend by using AutoML Tables to estimate predicted purchasing value, volume, frequency, lead conversion probability, and churn likelihood.

## Access pre-trained ML APIs for common applications



APIs like the Vision API or Natural Language Processing or Translation are already trained for common ML use cases like Image Classification. They save you the time and effort of building, curating, and training a new dataset so you can just jump ahead right to predictions.

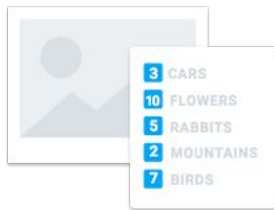
For pre-trained models, Google has already figured out a lot of the hard problems: the Vision API is based on Google's image datasets, the Speech-to-Text API is trained on YouTube captions, and the Cloud Translation API is built on parallel texts for language translations. Remember, how well your model is trained depends on how much data you have. As you would expect, Google has a lot of images and text and ML researchers to train its pre-built models, so you can use those instead of reinventing the wheel.

For example, if you're looking to have captions included in a recent webinar that you've hosted, consider using the Cloud Translation or Speech-to-Text APIs instead of trying to build a language recognition ML model yourself.

Another example, if you have text documents like expense receipts that you need classified by expense type, consider using the Vision API for OCR so you can mine the text from the receipts and drop the data into something like BigQuery.

Let's explore some of these pre-trained machine learning APIs.

## Use the Vision API to understand image content



Detect and label



Extract text



Identify entities



Let's start with the Vision API. There are three major components that all roll up into this REST API, and behind-the-scenes each of these are powered by many ML models and years of research.

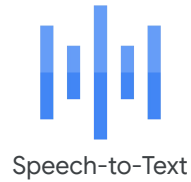
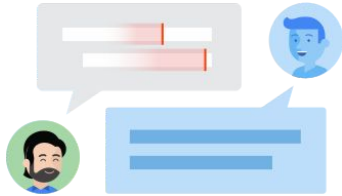
The first is detecting what an image is and classifying it. The Vision API picks out the dominant entity, for example a car or a cat, within an image from a broad set of object categories. This allows you to easily detect broad sets of objects in your images. Facial detection can detect when a face appears in photos, along with associated facial features such as eye, nose and mouth placement, and likelihood of over 8 attributes like joy and sorrow. Facial recognition however, isn't supported and Google doesn't store facial detection information on any Google server. You can use the API to easily build metadata on your image catalog, enabling new scenarios like image based searches or recommendations.

Next, are images with text, like scanned documents or signs. The Vision API uses optical character recognition, or OCR, to extract the text of a wide range of languages into a selectable, searchable format.

Lastly is a bit of intuition from the web and uses the power of Google Image Search. Does the image contain entities we know, like the Eiffel tower or a famous person? Landmark detection allows you to identify popular natural and manmade structures, along with the associated latitude and longitude of the landmark, and logo detection allows you to identify product logos within an image.

You can build metadata on your image catalog, extract text, moderate offensive content, or enable new marketing scenarios through image sentiment analysis. You can also analyze images uploaded in the request or integrate with an image storage on Cloud Storage.

## Convert speech to text and vice versa



Speech-to-Text



Text-to-Speech

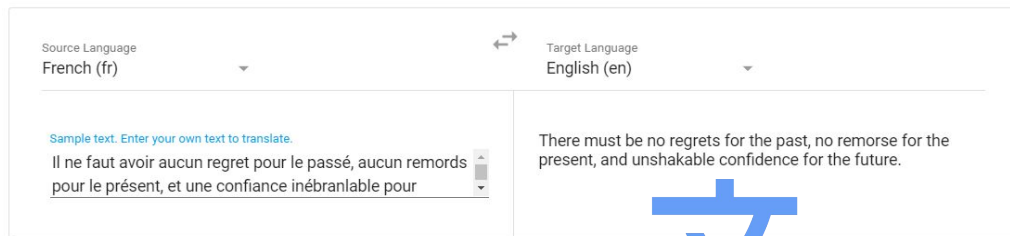


There are two APIs that apply to speech.

The Text-to-Speech API converts text into human-like speech in more than 180 voices across more than 30 languages and variants. It applies research in speech synthesis and Google's powerful neural networks to deliver high-fidelity audio. With this API, you can create lifelike interactions with users that transform customer service, device interaction, and other applications.

The Speech-to-Text API enables you to convert real-time streaming or prerecorded audio to text. The API recognizes 120 languages and variants to support a global user base. You can enable voice command-and-control, transcribe audio from call centers, and so on.

## Dynamically translate between languages using the Cloud Translation API



The screenshot shows the Google Cloud Translation API interface. At the top, there are two dropdown menus: 'Source Language' set to 'French (fr)' and 'Target Language' set to 'English (en)'. Below these, there is a text input area on the left containing the French text: 'Il ne faut avoir aucun regret pour le passé, aucun remords pour le présent, et une confiance inébranlable pour'. To the right of the input area, the translated English text is displayed: 'There must be no regrets for the past, no remorse for the present, and unshakable confidence for the future.' A large blue watermark '文A' is overlaid on the right side of the interface.



The Cloud Translation API provides a simple programmatic interface for translating an arbitrary string into any supported language. The Cloud Translation API is highly responsive, so websites and applications can integrate with the API for fast, dynamic translation of source text from the source language to a target language, for example from French to English. Language detection is also available in cases where the source language is unknown.

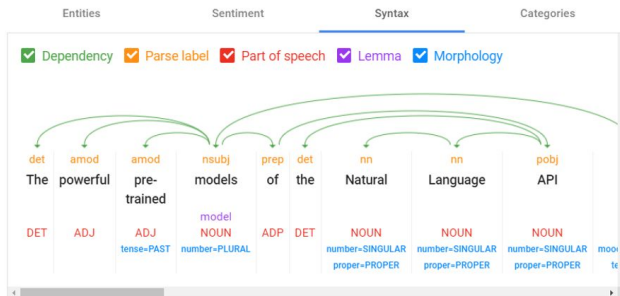
Let's look at a short video that shows how Bloomberg, a global leader in business and financial data, news and insight, applied the Cloud Translation API to reach all of their customers regardless of language.

# Derive insights from unstructured text with the Cloud Natural Language API

The powerful pre-trained models of the Natural Language API let developers work with natural language understanding features including sentiment analysis, entity analysis, entity sentiment analysis, content classification, and syntax analysis.

RESET

[See supported languages](#)



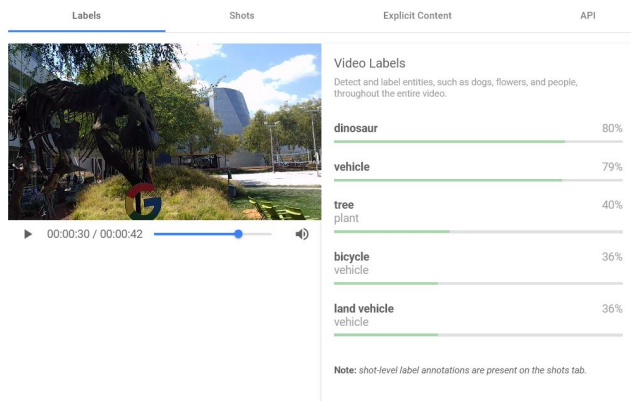
The Cloud Natural Language API offers a variety of natural language understanding technologies. It can do syntax analysis, breaking down sentences into tokens, identify the nouns, verbs, adjectives, and other parts of speech, and figuring out the relationships among the words.

It can also do entity recognition, in other words, it can parse text and flag mentions of people, organizations, locations, events, products and media.

Sentiment analysis allows you to understand customer opinions to find actionable product and UX insights.



# Make your media more discoverable with the Video Intelligence API



The Video Intelligence API allows users to use Google video analysis technology as part of their applications. The REST API enables users to annotate videos stored in Cloud Storage with video and 1 frame-per-second contextual information. It helps you identify key entities -- that is, nouns -- within your video, and when they occur. You can use it to make video content searchable and discoverable.

The API supports the annotation of common video formats, including .MOV, .MPEG4, .MP4, and .AVI.

---

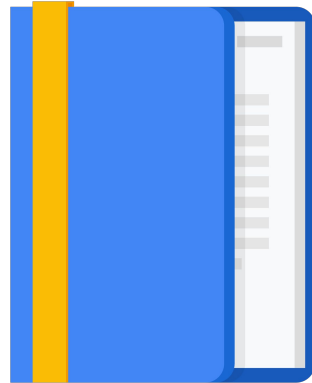
# Agenda

Google Cloud Big Data Platform

Machine Learning in the Cloud

[Quiz and Lab](#)

Resources



---

## Quiz 1

When would you use Dataproc?

---

## Quiz 1

When would you use Dataproc?

You can use it to migrate on-premises Hadoop jobs to the cloud. You can also use it for data mining and analysis of cloud-based data.

---

## Quiz 2

Name two use cases for Dataflow.

---

## Quiz 2

Name two use cases for Dataflow.

1. ETL
2. Orchestration

---

## Quiz 3

Which machine learning tool would be the best option for someone that wants a custom model but has limited application development or data science skills?

- A. AI Platform
- B. AutoML
- C. Tensorflow
- D. Speech API

---

## Quiz 3

Which machine learning tool would be the best option for someone that wants a custom model but has limited application development or data science skills?

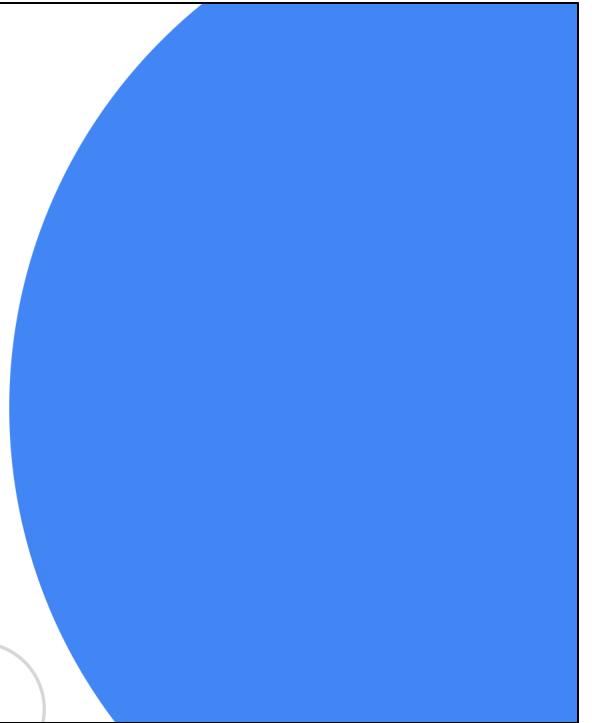
- A. AI Platform
- B. AutoML
- C. Tensorflow
- D. Speech API



# Lab Intro

## Getting Started with BigQuery

Duration: 30 minutes



The objectives for this lab are for you to:

- Load data from Cloud Storage into BigQuery.
- Perform a query on the data in BigQuery.

---

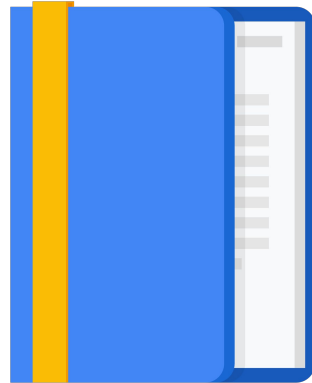
# Agenda

Google Cloud Big Data Platform

Machine Learning in the Cloud

Quiz and Lab

[Resources](#)



---

## Resources

Smart Analytics <https://cloud.google.com/solutions/smart-analytics>

Cloud AI <https://cloud.google.com/products/ai/>

