Big Data and Machine Learning in the Cloud

GCP Fundamentals: Core Infrastructure

Getting Started with BigQuery

Last modified 2018-08-24

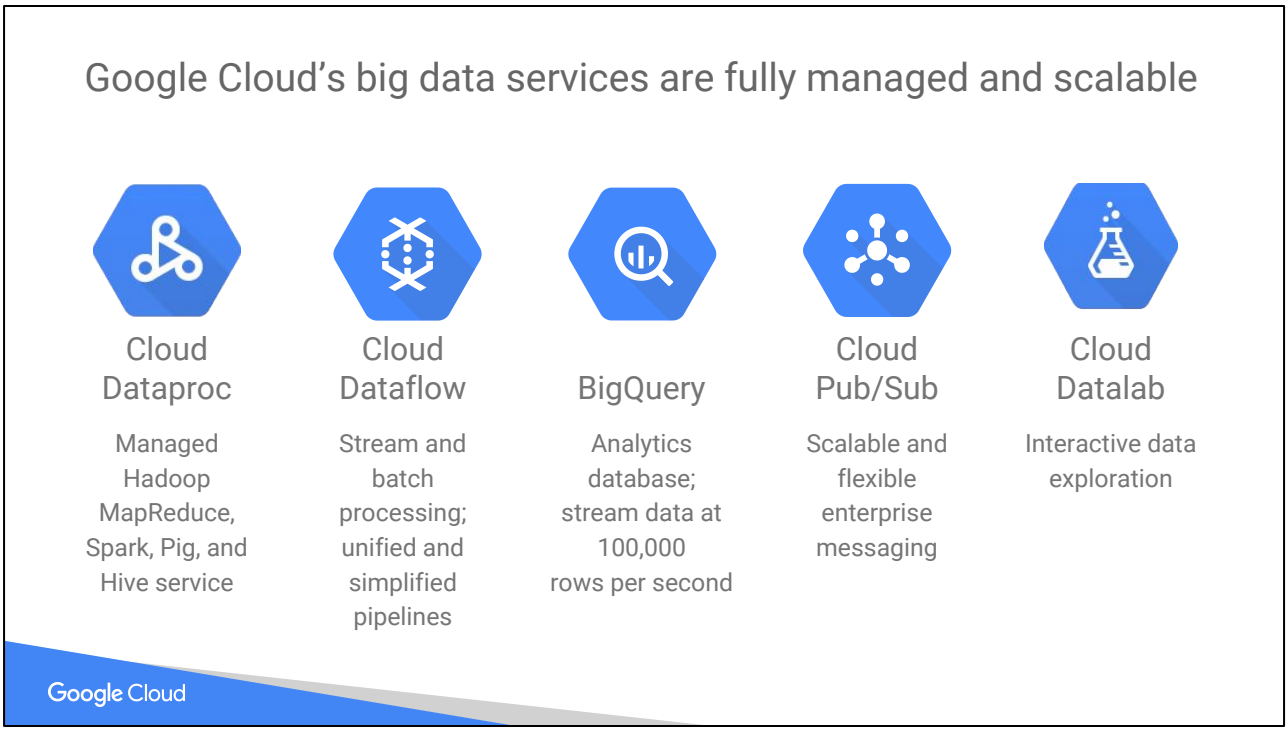
Agenda

Google Cloud Big Data Platform

Google Cloud Machine Learning Platform

Quiz and Lab

Google Cloud



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 GCP data services work together to help you create custom solutions.

Cloud Dataproc is managed Hadoop

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

Google Cloud

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.

Cloud Dataproc is a fast, easy, managed way to run Hadoop, Spark, Hive, and Pig on Google Cloud Platform. 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 Stackdriver.

Why use Cloud 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.

Google Cloud

Running on-premises Hadoop jobs requires a hardware investment. On the other hand, running these jobs in Cloud Dataproc allows you to pay only for hardware resources during the life of the ephemeral customer you create. You can further save money using preemptible instances for batch processing.

You can also save money by telling Cloud Dataproc to use preemptible Compute Engine instances for your 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. At the time this video was made, preemptible instances were around 80% cheaper. 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.

Cloud 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

Google Cloud

Cloud 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 Cloud 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. Cloud Dataflow frees you from operational tasks like resource management and performance optimization.

Cloud Dataflow features:

Resource Management
Cloud Dataflow fully automates management of required processing resources. No more spinning up instances by hand.

On Demand

All resources are provided on demand, enabling you to scale to meet your business needs. No need to buy reserved compute instances.

Intelligent Work Scheduling
Automated and optimized work partitioning which can dynamically rebalance lagging work. No more chasing down “hot keys” or pre-processing your input data.

Auto Scaling
Horizontal auto scaling of worker resources to meet optimum throughput requirements results in better overall price-to-performance.

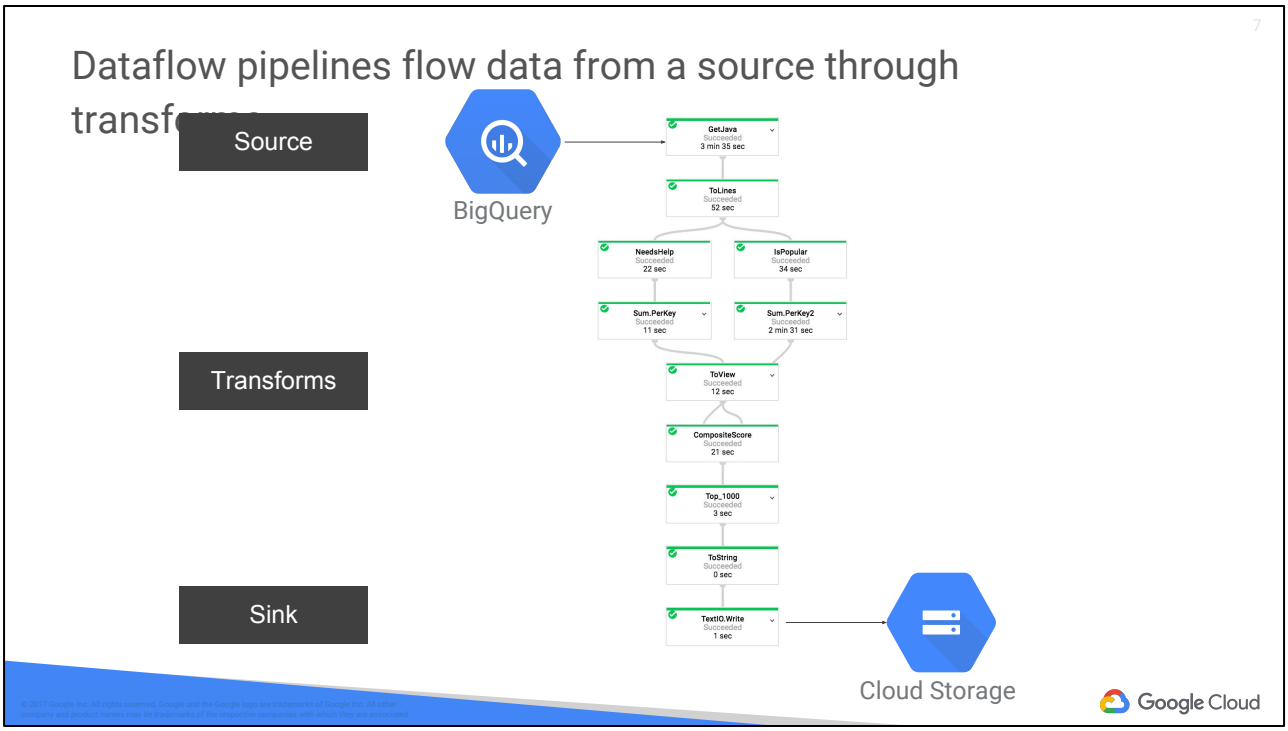
Unified Programming Model
The Dataflow API enables you to express MapReduce like operations, powerful data windowing, and fine grained correctness control regardless of data source.

Open Source
Developers wishing to extend the Dataflow programming model can fork and or submit pull requests on the Java-based Cloud Dataflow SDK. Dataflow pipelines can also run on alternate runtimes like Spark and Flink.

Monitoring
Integrated into the Google Cloud Platform Console, Cloud Dataflow provides statistics such as pipeline throughput and lag, as well as consolidated worker log inspection—all in near-real time.

Integrated
Integrates with Cloud Storage, Cloud Pub/Sub, Cloud Datastore, Cloud Bigtable, and BigQuery for seamless data processing. And can be extended to interact with others sources and sinks like Apache Kafka and HDFS.

Reliable & Consistent Processing
Cloud Dataflow provides built-in support for fault-tolerant execution that is consistent and correct regardless of data size, cluster size, processing pattern or pipeline complexity.



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 custer.

Why use Cloud 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 GCP services like Cloud Storage, Cloud Pub/Sub, BigQuery, and Bigtable
  - Open source Java and Python SDKs

Google Cloud

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 realtime applications such as personalizing gaming user experiences.

BigQuery is a fully managed data warehouse

- Provides near real-time interactive analysis of massive datasets (hundreds of TBs)
- Query using SQL syntax (SQL 2011)
- No cluster maintenance is required.

Google Cloud

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.

BigQuery is Google's fully managed, petabyte scale, low cost analytics data warehouse. BigQuery is NoOps: there is no infrastructure to manage and you don't need a database administrator, so you can focus on analyzing data to find meaningful insights, use familiar SQL, and take advantage of our pay-as-you-go model. BigQuery is a powerful big data analytics platform used by all types of organizations, from startups to Fortune 500 companies.

BigQuery's features:

Flexible Data Ingestion
Load your data from Cloud Storage or Cloud Datastore, or stream it into BigQuery at 100,000 rows per second to enable real-time analysis of your data.

Global Availability
You have the option to store your BigQuery data in European locations while continuing to benefit from a fully managed service, now with the option of geographic data control, without low-level cluster maintenance.

Security and Permissions
You have full control over who has access to the data stored in BigQuery. If you share datasets, doing so will not impact your cost or performance; those you share with pay for their own queries.

Cost Controls
BigQuery provides cost control mechanisms that enable you to cap your daily costs at an amount that you choose. For more information, see Cost Controls.

Highly Available
Transparent data replication in multiple geographies means that your data is available and durable even in the case of extreme failure modes.

Super Fast Performance
Run super-fast SQL queries against multiple terabytes of data in seconds, using the processing power of Google's infrastructure.

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

Connect with Google Products
You can automatically export your data from Google Analytics Premium into BigQuery and analyze datasets stored in Google Cloud Storage, Google Drive, and Google Sheets.

BigQuery can make Create, Replace, Update, and Delete changes to databases, subject to some limitations and with certain known issues.

BigQuery runs on Google's high-performance infrastructure

- Compute and storage are separated with a terabit network in between
- You only pay for storage and processing used
- Automatic discount for long-term data storage

Google Cloud

It's easy to get data into BigQuery. You can load from Cloud Storage or Cloud Datastore, or stream it into BigQuery at up to 100,000 rows per second.

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 its available 99.9% service level agreement.

Long term storage pricing is an automatic discount for data residing in BigQuery for extended periods of time. When the age of your data reaches 90 days in BigQuery, Google will automatically drop the price of storage from $0.02 per GB per month down to $0.01 per GB per month.

For more information on the architecture of BigQuery, see: https://cloud.google.com/blog/big-data/2016/01/bigquery-under-the-hood

Cloud Pub/Sub is scalable, reliable messaging

- Supports many-to-many asynchronous messaging
  - Application components make push/pull subscriptions to topics
- Includes support for offline consumers
- Based on proven Google technologies
- Integrates with Cloud Dataflow for data processing pipelines

Google Cloud

Cloud Pub/Sub is a fully managed real-time messaging service that allows you to send and receive messages between independent applications. You can leverage Cloud Pub/Sub's flexibility to decouple systems and components hosted on Google Cloud Platform or elsewhere on the internet. By building on the same technology Google uses, Cloud Pub/Sub is designed to provide "at least once" delivery at low latency with on-demand scalability to 1 million messages per second (and beyond).

Cloud Pub/Sub features:

Highly Scalable
Any customer can send up to 10,000 messages per second, by default—and millions per second and beyond, upon request.

Push and Pull Delivery
Subscribers have flexible delivery options, whether they are accessible from the internet or behind a firewall.

Encryption
Encryption of all message data on the wire and at rest provides data security and protection.

Replicated Storage
Designed to provide “at least once” message delivery by storing every message on multiple servers in multiple zones.

Message Queue
Build a highly scalable queue of messages using a single topic and subscription to support a one-to-one communication pattern.

End-to-End Acknowledgement
Building reliable applications is easier with explicit application-level acknowledgements.

Fan-out
Publish messages to a topic once, and multiple subscribers receive copies to support one-to-many or many-to-many communication patterns.

REST API
Simple, stateless interface using JSON messages with API libraries in many programming languages.

Why use Cloud Pub/Sub?

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

Google Cloud

Cloud 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, Cloud Dataflow is a natural pairing with Pub/Sub.

Cloud Datalab offers interactive data exploration

- Interactive tool for large-scale data exploration, transformation, analysis, and visualization
- Integrated, open source
  - Built on Jupyter (formerly IPython)

Google Cloud

For data science, an online lab notebook metaphor is a useful environment, because it feels natural to intersperse data analyses with comments about their results. A popular open-source system for hosting those is Project Jupyter. It lets you create and maintain web-based notebooks containing Python code, and you can run that code interactively and view the results.

Cloud Datalab lets you use Jupyter notebooks to explore, analyze, and visualize data on the Google Cloud Platform. It runs in a Compute Engine virtual machine. To get started, you specify the virtual machine type you want and what GCP region it should run in. When it launches, it presents an interactive Python environment that's ready to use. And it orchestrates multiple GCP services automatically, so you can focus on exploring your data. You only pay for the resources you use; there's no additional charge for Datalab itself.

Cloud Datalab features:

Integrated
Cloud Datalab handles authentication and cloud computation out of the box and is integrated with BigQuery, Compute Engine, and Cloud Storage.

Multi-Language Support

Cloud Datalab currently supports Python, SQL, and JavaScript (for BigQuery user-defined functions).

Notebook Format
Cloud Datalab combines code, documentation, results, and visualizations together in an intuitive notebook format.

Pay-per-use Pricing
Only pay for the cloud resources you use: the App Engine application, BigQuery, and any additional resources you decide to use, such as Cloud Storage.

Interactive Data Visualization
Use Google Charts or matplotlib for easy visualizations.

Collaborative
Git-based source control of notebooks with the option to sync with non-Google source code repositories like GitHub and Bitbucket.

Open Source
Developers who want to extend Cloud Datalab can fork and/or submit pull requests on the GitHub hosted project.

Custom Deployment
Specify your minimum VM requirements, the network host, and more.

IPython Support
Cloud Datalab is based on Jupyter (formerly IPython) so you can use a large number of existing packages for statistics, machine learning, etc. Learn from published notebooks and swap tips with a vibrant IPython community.

Why use Cloud Datalab?

- Create and manage code, documentation, results, and visualizations in intuitive notebook format.
  - Use Google Charts or matplotlib for easy visualizations.
- Analyze data in BigQuery, Compute Engine, and Cloud Storage using Python, SQL, and JavaScript.
- Easily deploy models to BigQuery.

Google Cloud

Cloud Datalab is integrated with BigQuery, Compute Engine, and Cloud Storage, so accessing your data doesn't run into authentication hassles.

When you're up and running, you can visualize your data with Google Charts or matplotlib. And, because there's a vibrant interactive Python community, you can learn from published notebooks. There are many existing packages for statistics, machine learning, and so on.

You can attach a GPU to a Cloud Datalab instance for faster processing. At the time of this writing, this feature was in beta, which means that no SLA is available and that the feature could change in backwards-incompatible ways.

Agenda

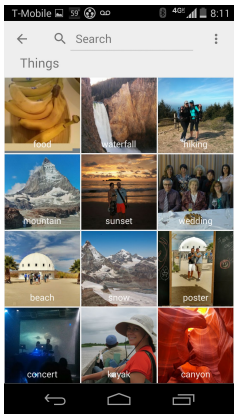
Google Cloud Big Data Platform

Google Cloud Machine Learning Platform

Quiz and Lab

Google Cloud

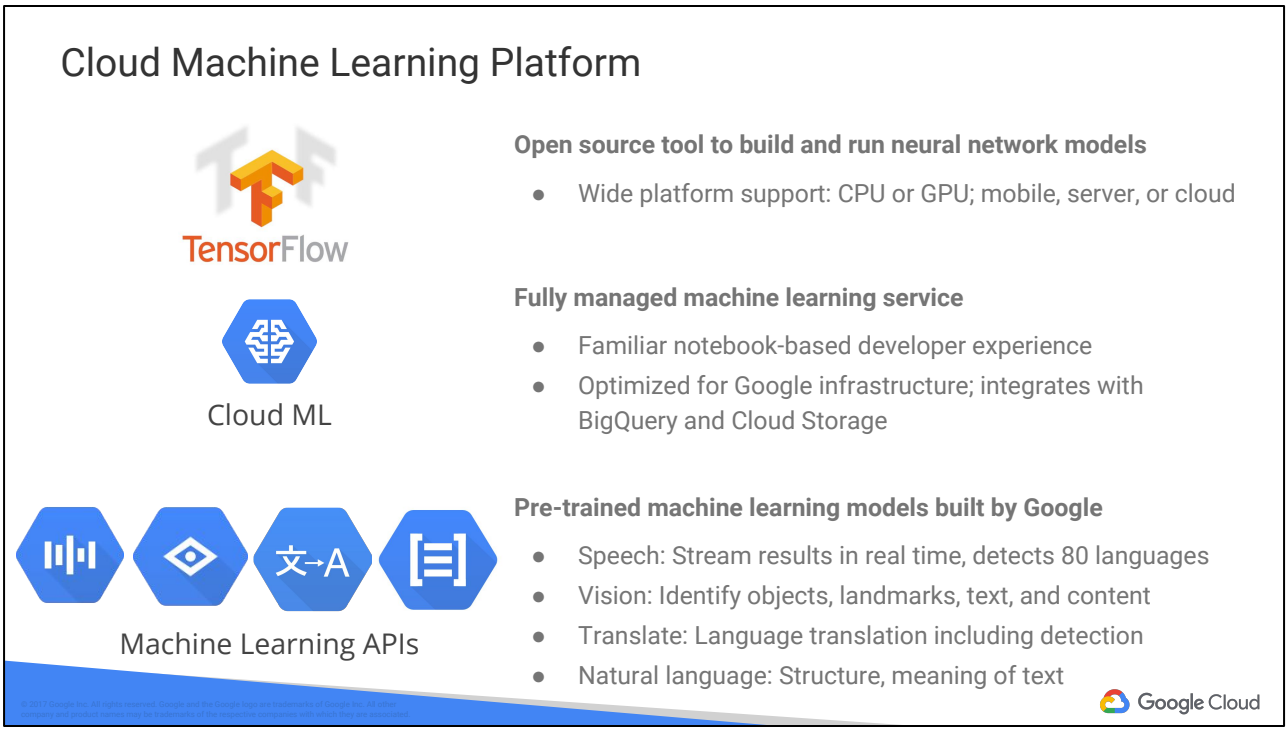
Machine Learning APIs enable apps that see, hear, and understand



Google Cloud

Machine learning is one branch of the field of artificial intelligence. It's a way of solving problems without explicitly coding the solution. Instead, human coders build systems that improve themselves over time, through repeated exposure to sample data, which we call "training data."

Major Google applications use Machine Learning, like YouTube, Photos, the Google mobile app, and Google Translate. The Google machine learning platform is now available as a cloud service, so that you can add innovative capabilities to your own applications.



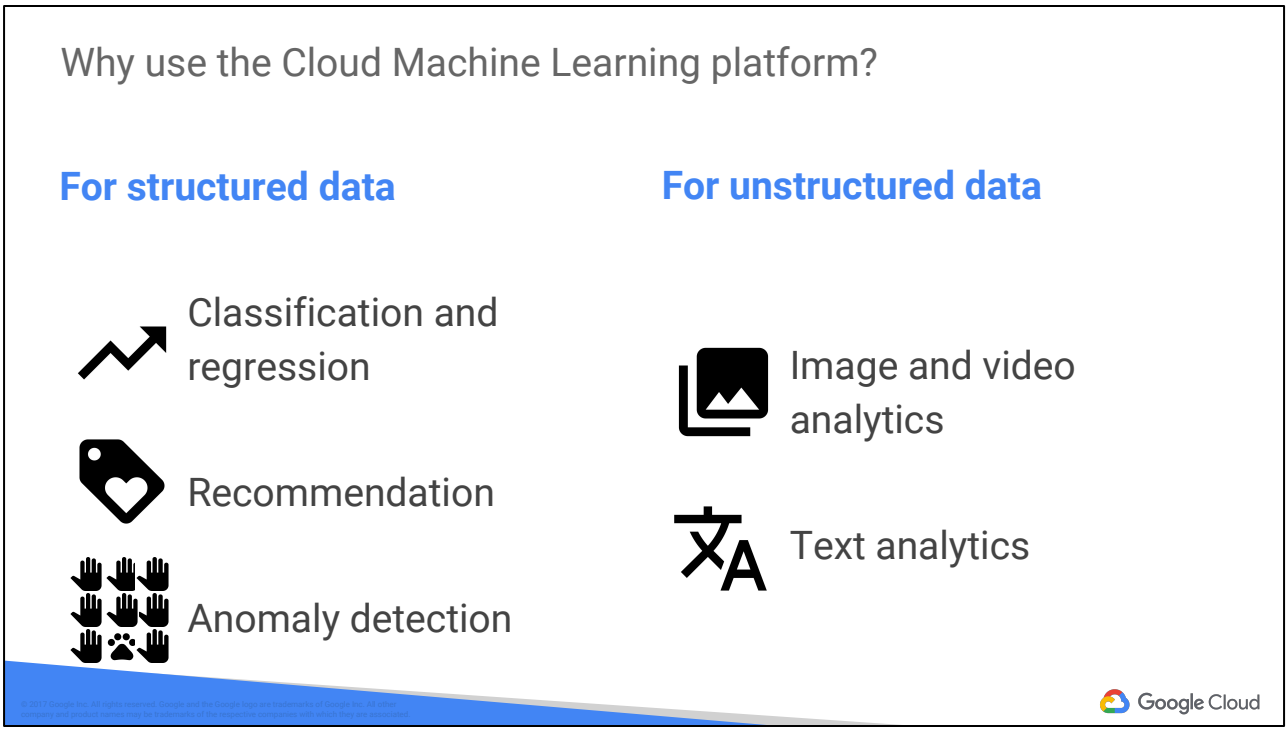
Cloud Machine Learning Platform provides modern machine learning services, with pre-trained models and a platform to generate your own tailored models. As with other GCP products, there's a range of services that stretches from the highly general to the pre-customized.

TensorFlow is an open-source software library that's exceptionally well suited for machine learning applications like neural networks. It was developed by Google Brain for Google's internal use and then open-sourced so that the world could benefit. You can run TensorFlow wherever you like, but GCP is an ideal place for it, because machine learning models need lots of on-demand compute resources and lots of training data. TensorFlow can also take advantage of Tensor Processing Units, which are hardware devices designed to accelerate machine learning workloads with TensorFlow. GCP makes them available in the cloud with Compute Engine virtual machines. Each Cloud TPU provides up to 180 teraflops of performance, and, because you pay only for what you use, there's no up-front capital investment required.

Suppose you want a more managed service. Google Cloud Machine Learning Engine lets you easily build machine learning models that work on any type of

data, of any size. It can take any TensorFlow model and perform large scale training on a managed cluster.

Finally, suppose you just want to add various machine-learning capabilities to your applications, without having to worry about the details of how they are provided. Google Cloud also offers a range of machine-learning APIs suited for specific purposes, and I'll discuss them in a moment.



People use the Cloud Machine Learning platform for lots of applications. Generally, they fall into two categories, depending on whether the data they work on is structured or unstructured.

Based on structured data, you can use ML for various kinds of classification and regression tasks, like customer churn analysis, product diagnostics, and forecasting. It can be the heart of a recommendation engine, for content personalization and cross-sells and up-sells. You can use ML to detect anomalies, as in fraud detection, sensor diagnostics, or log metrics.

Based on unstructured data, you can use ML for image analytics, such as identifying damaged shipment, identifying styles, and flagging content. You can do text analytics too, like call center log analysis, language identification, topic classification, and sentiment analysis.

In many of the most innovative applications for machine learning, several of these kinds of applications are combined. What if, whenever one of your customers posted praise for one of your products on social media, your application could automatically reach out to them with a customized discount

on another product they'll probably like? The Google Cloud Machine Learning platform makes that kind of interactivity well within your grasp.

Cloud Vision API

- Analyze images with a simple REST API
  - Logo detection, label detection, etc
- With the Cloud Vision API, you can:
  - Gain insight from images
  - Detect inappropriate content
  - Analyze sentiment
  - Extract text

Google Cloud

Cloud Vision API enables developers to understand the content of an image by encapsulating powerful machine learning models in an easy to use REST API. It quickly classifies images into thousands of categories ("sailboat," "lion," "Eiffel Tower"), detects individual objects within images, and finds and reads printed words contained within images. You can build metadata on your image catalog, moderate offensive content, or enable new marketing scenarios through image sentiment analysis. Analyze images uploaded in the request or integrate with your image storage on Cloud Storage.

Cloud Speech API

- Recognizes over 80 languages and variants
- Can return text in real time
- Highly accurate, even in noisy environments
- Access from any device
- Powered by Google's machine learning

Google Cloud

The Cloud Speech API enables developers to convert audio to text. Because you have an increasingly global user base, the API recognizes over 80 languages and variants. You can transcribe the text of users dictating to an application's microphone, enable command-and-control through voice, or transcribe audio files.

Cloud Natural Language API

- Uses machine learning models to reveal structure and meaning of text.
- Extract information about items mentioned in text documents, news articles, and blog posts.
- Analyze text uploaded in request or integrate with Cloud Storage.

Google Cloud

The Cloud Natural Language API offers a variety of natural language understanding technologies to developers.

It can do syntax analysis, breaking down sentences supplied by your users into tokens, identify the nouns, verbs, adjectives, and other parts of speech, and figure out the relationships among the words.

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

It can understand the overall sentiment expressed in a block of text.

And it has these capabilities in multiple languages including English, Spanish, and Japanese.

Cloud Natural Language API features

Syntax Analysis

- Extract tokens and sentences, identify parts of speech (PoS), and create dependency parse trees for each sentence.

Entity Recognition

- Identify entities and label by types such as person, organization, location, events, products and media.

Sentiment Analysis

- Understand the overall sentiment expressed in a block of text.

Multi-Language

- Enables you to easily analyze text in multiple languages including English, Spanish, and Japanese.

Integrated REST API

- Access via REST API. Text can be uploaded in the request or integrated with Cloud Storage.

For more information on the Natural Language API, see: https://cloud.google.com/natural-language/docs/.

Cloud Translation API

- Translate arbitrary strings between thousands of language pairs
- Programmatically detect a document's language
- Support for dozens of languages

Google Cloud

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

The Translation API supports the standard Google API Client Libraries in Python, Java, Ruby, Objective-C, and other languages.

You can try it in your browser:
https://developers.google.com/apis-explorer/#p/translate/v2/

Cloud Video Intelligence API

- Annotate the contents of videos
- Detect scene changes
- Flag inappropriate content
- Support for a variety of video formats

Google Cloud

The Google Cloud Video Intelligence API allows developers to use Google video analysis technology as part of their applications. The REST API enables users to annotate videos stored in Google Cloud Storage with video and frame-level (1 fps) 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

Google Cloud Machine Learning Platform

Quiz and Lab

Google Cloud

Quiz

When would you use Cloud Dataproc?

Name two use cases for Cloud Dataflow.

Name three use cases for the Google machine learning platform.

Quiz Answers

When would you use Cloud 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.
Name two use cases for Cloud Dataflow.	
Name three use cases for the Google machine learning platform.	

Quiz Answers

When would you use Cloud 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.
Name two use cases for Cloud Dataflow.	ETL, orchestration
Name three use cases for the Google machine learning platform.	

Quiz Answers

When would you use Cloud 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.
Name two use cases for Cloud Dataflow.	ETL, orchestration
Name three use cases for the Google machine learning platform.	Fraud detection, sentiment analysis, content personalization

Lab instructions

In this lab, you will load server log data into BigQuery and perform a SQL query on it.

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

Google Cloud

In this lab, you load a CSV file into a BigQuery table. After loading the data, you query it using the BigQuery web user interface, the CLI, and the BigQuery shell.

More resources

Google Big Data Platform
https://cloud.google.com/products/big-data/

Google Machine Learning Platform
https://cloud.google.com/products/machine-learning/