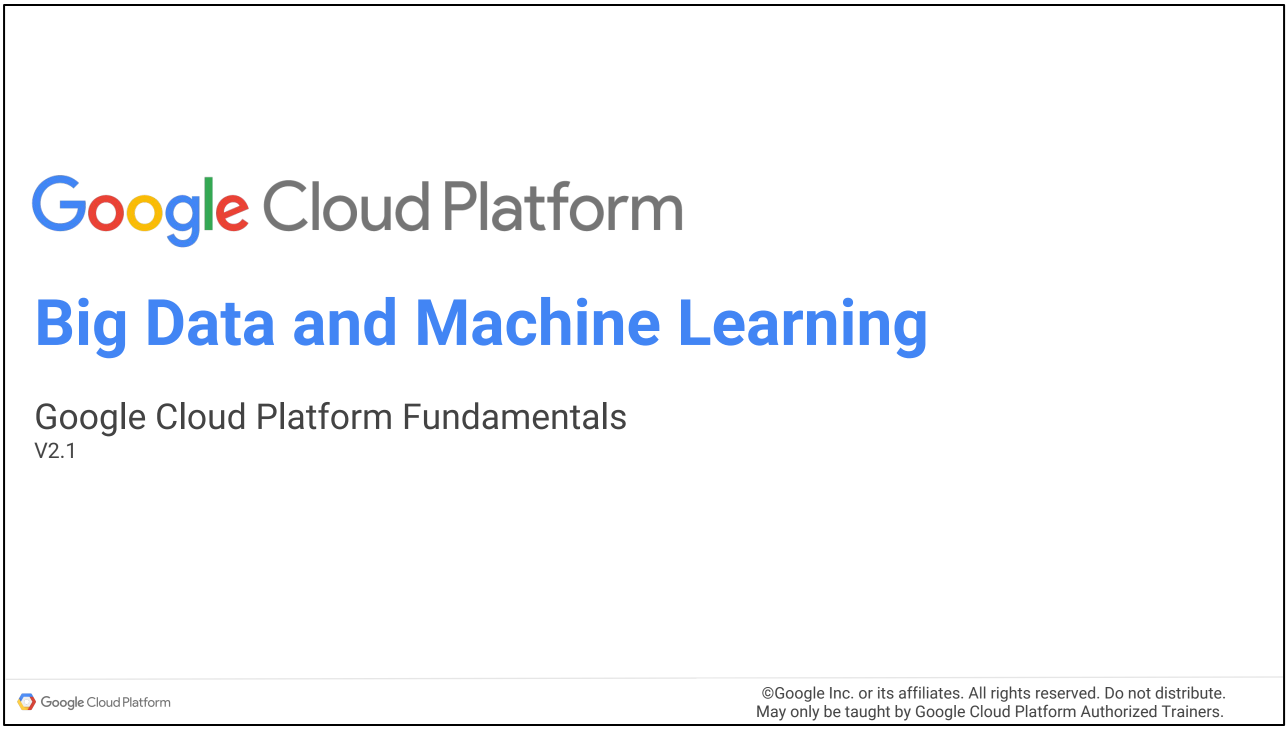


Google Cloud Platform

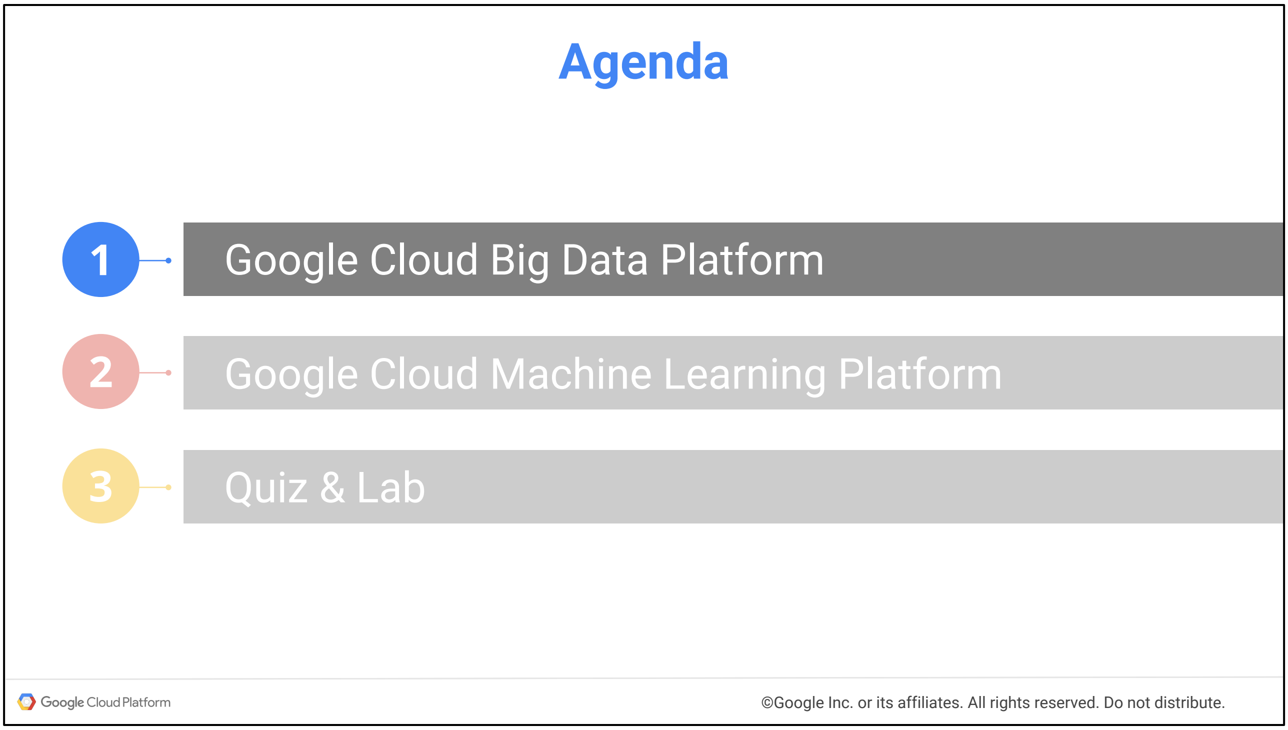
Big Data and Machine Learning

Google Cloud Platform Fundamentals

V2.1

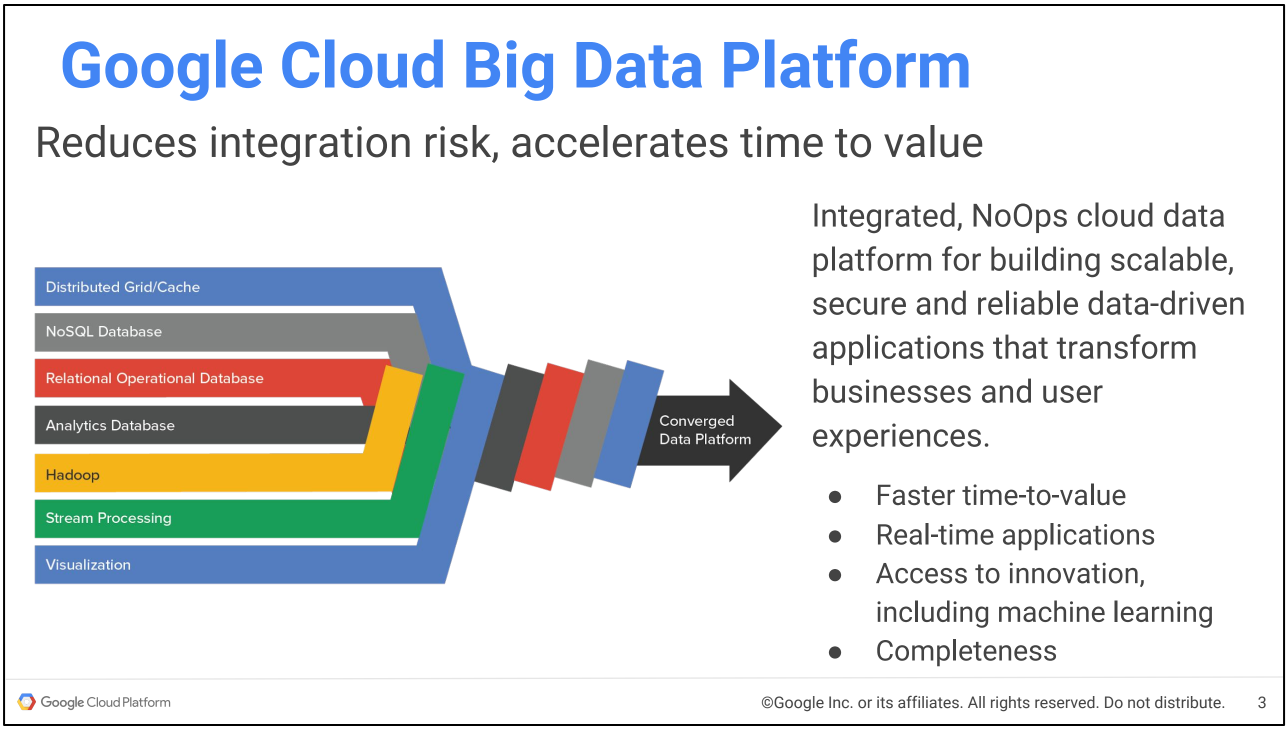
Google Cloud Platform

Timing: Approximately 30 minutes



Agenda

1. Google Cloud Big Data Platform
2. Google Cloud Machine Learning Platform
3. Quiz & Lab



Google Cloud Big Data Platform

Reduces integration risk, accelerates time to value

Distributed Grid/Cache

NoSQL Database

Relational Operational Database

Analytics Database

Hadoop

Stream Processing

Visualization

Converged Data Platform

Integrated, NoOps cloud data platform for building scalable, secure and reliable data-driven applications that transform businesses and user experiences.

- Faster time-to-value
- Real-time applications
- Access to innovation, including machine learning
- Completeness

Google Cloud Platform

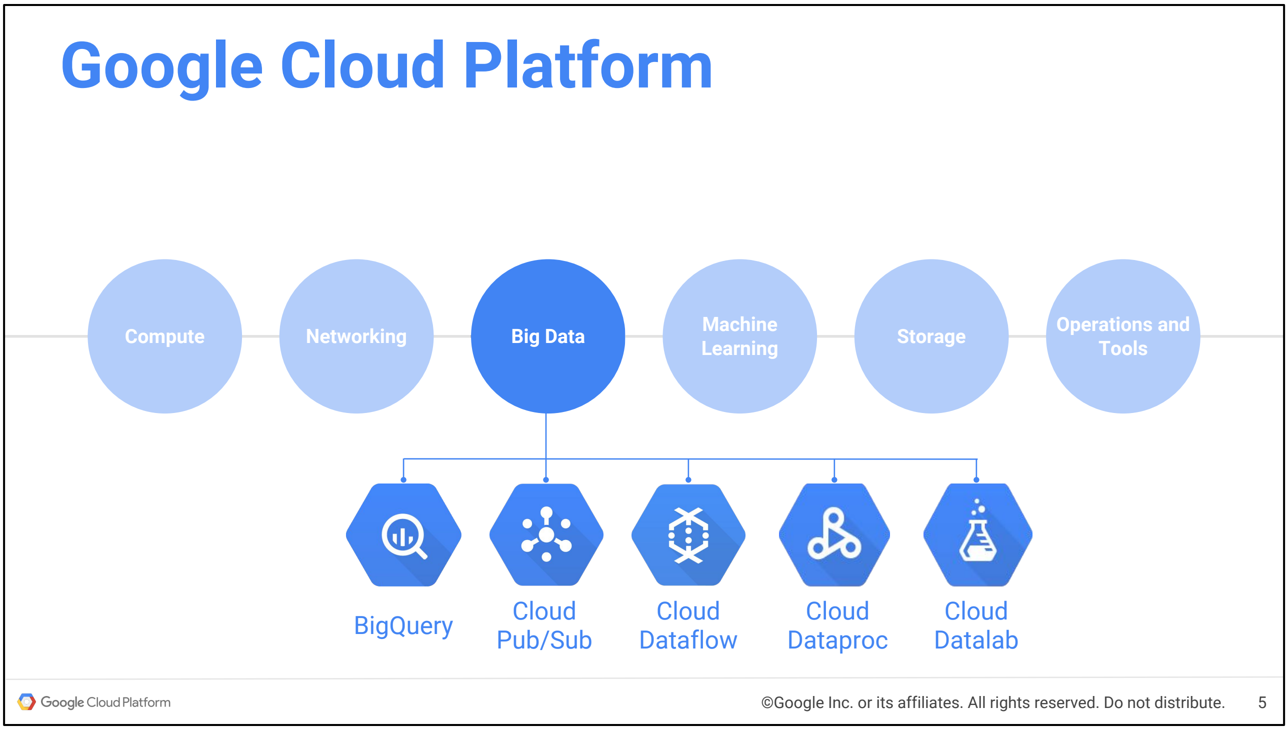
Notes:
Why Google for Big Data?

NoOps
Cloud Platform's cloud-native data processing services require no provisioning, and you only pay for what you consume. Pay for BigQuery storage separately from queries. Pay for queries only when they are actually running. Submit Cloud Dataflow jobs that run on dynamically provisioned and autoscaled resources, not a fixed cluster.

Innovation
Google has led the industry with innovations in data processing technologies such as MapReduce, BigTable, and Dremel. Now, Google is making the latest generation of its data processing tools available to everyone, including industry leading programming tools and programming models. For example, Cloud Dataflow enables the same pipeline to run in streaming or batch mode, and Cloud Dataflow programs are portable in and out of Cloud Platform.

Scale
Cloud Platform is designed to scale like Google’s own products, even when you experience a huge traffic spike. Managed services such as BigQuery, Google Cloud Pub/Sub and Cloud Dataflow give you autoscaling that enables

your application to grow with your users, without you needing to over-provision and pay for unused capacity.



Google Cloud Platform

Compute

Networking

Big Data

Machine Learning

Storage

Operations and Tools

BigQuery

Cloud Pub/Sub

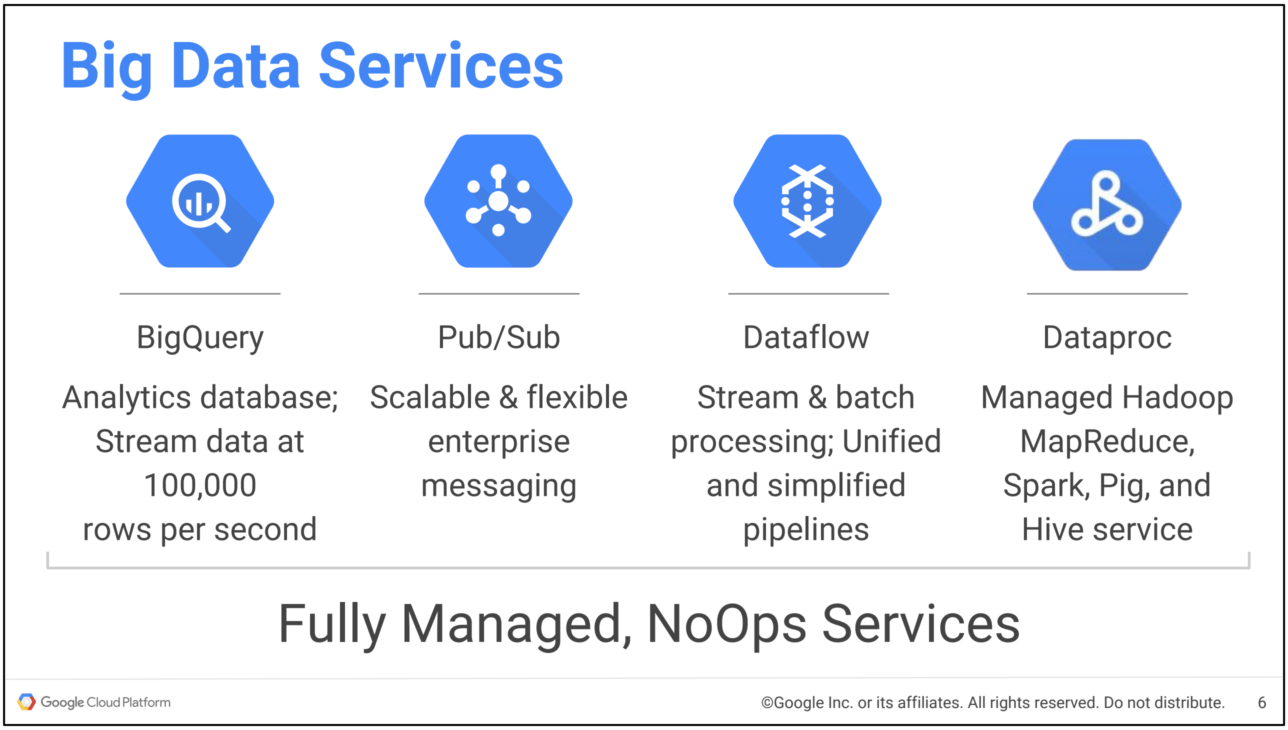
Cloud Dataflow

Cloud Dataproc

Cloud Datalab

Google Cloud Platform

Notes:
The Google Cloud Big Data platform includes several services to help you collect, ingest, and analyze your data including: Google BigQuery, Google Cloud Pub/Sub, Google Cloud Dataflow, Google Cloud Dataproc, and Google Cloud Datalab. Each of these services is discussed in this module.



Big Data Services

BigQuery

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

Pub/Sub

Scalable & flexible enterprise messaging

Dataflow

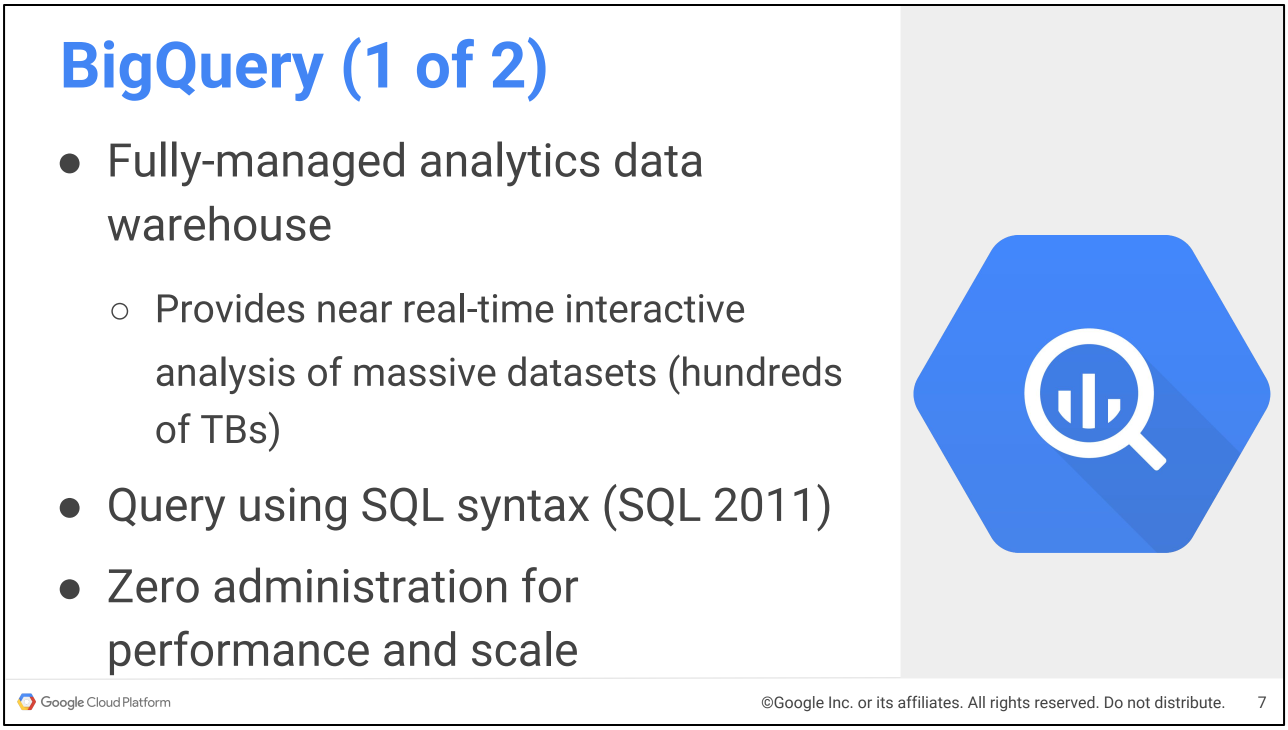
Stream & batch processing; Unified and simplified pipelines

Dataproc

Managed Hadoop MapReduce, Spark, Pig, and Hive service

Fully Managed, NoOps Services

Google Cloud Platform



BigQuery (1 of 2)

- Fully-managed analytics data warehouse
  - ○ Provides near real-time interactive analysis of massive datasets (hundreds of TBs)
- Query using SQL syntax (SQL 2011)
- Zero administration for performance and scale

Google Cloud Platform

Notes:
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 Google Cloud Storage or Google 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 headaches.

Security & Permissions
You have full control over who has access to the data stored in Google

BigQuery. Shared datasets will not impact your cost or performance (those you share with pay for their own queries).

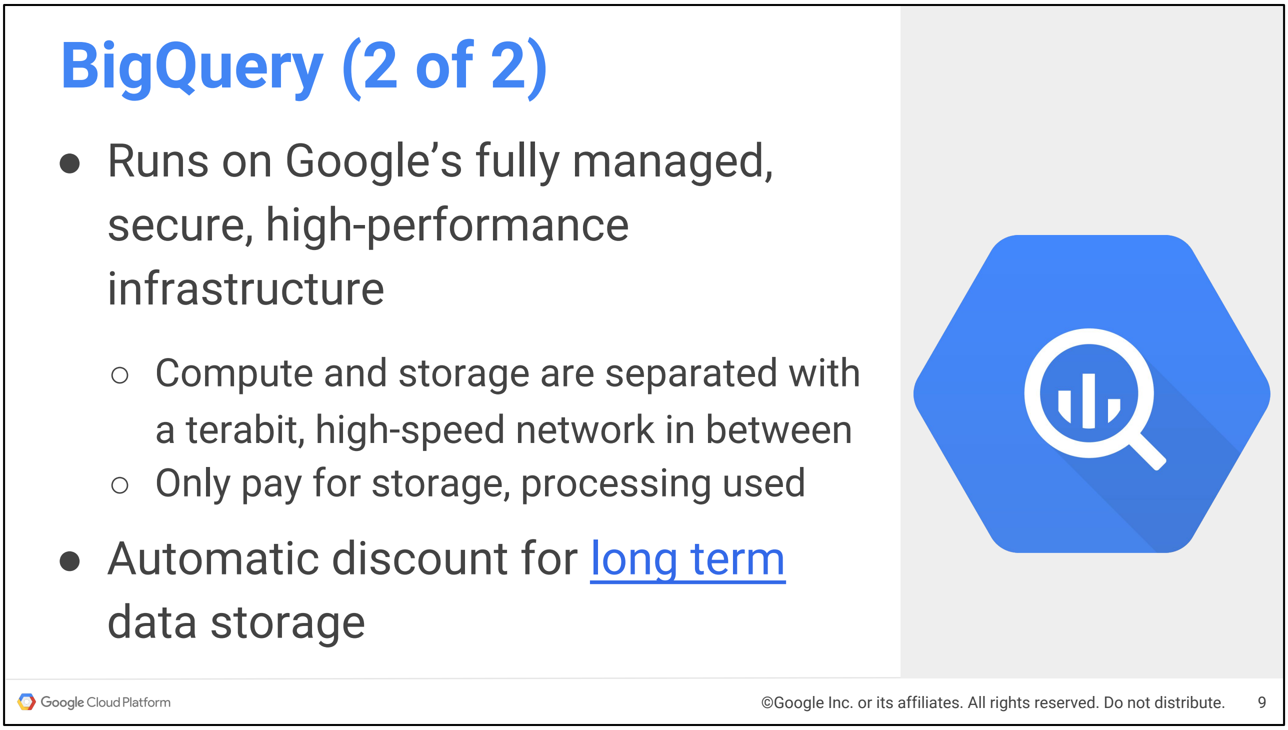
Cost Controls
BigQuery provides cost control mechanisms that enable you to cap your daily costs to an amount that you choose. For more information, see Cost Controls.

Highly Available
Transparent data replication in multiple geographies means 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 (2 of 2)

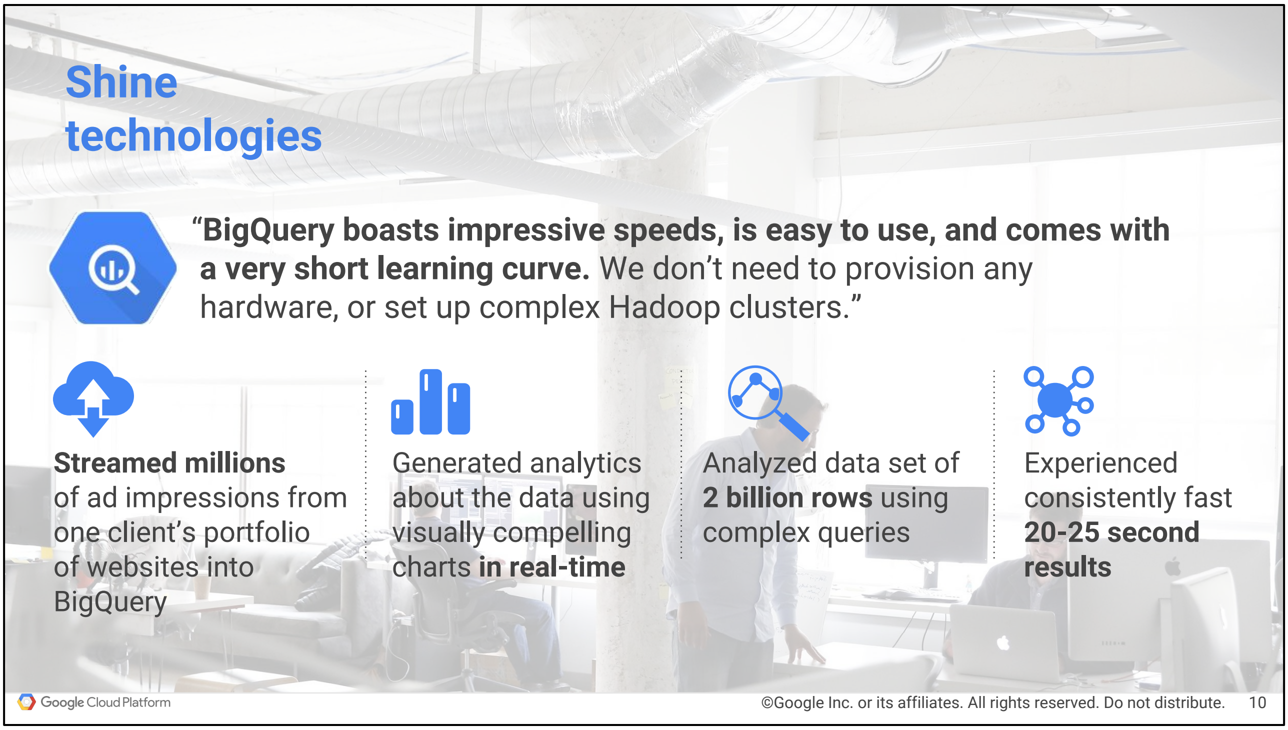
- Runs on Google's fully managed, secure, high-performance infrastructure
    - Compute and storage are separated with a terabit, high-speed network in between
    - Only pay for storage, processing used
- Automatic discount for long term data storage

Google Cloud Platform

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

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

Tracking began with data edited on February 1st, 2016, so you should see the price of your long term storage drop 90 days later, on May 1st, 2016.



Shine technologies

"**BigQuery boasts impressive speeds, is easy to use, and comes with a very short learning curve.** We don't need to provision any hardware, or set up complex Hadoop clusters."

Streamed millions of ad impressions from one client's portfolio of websites into BigQuery

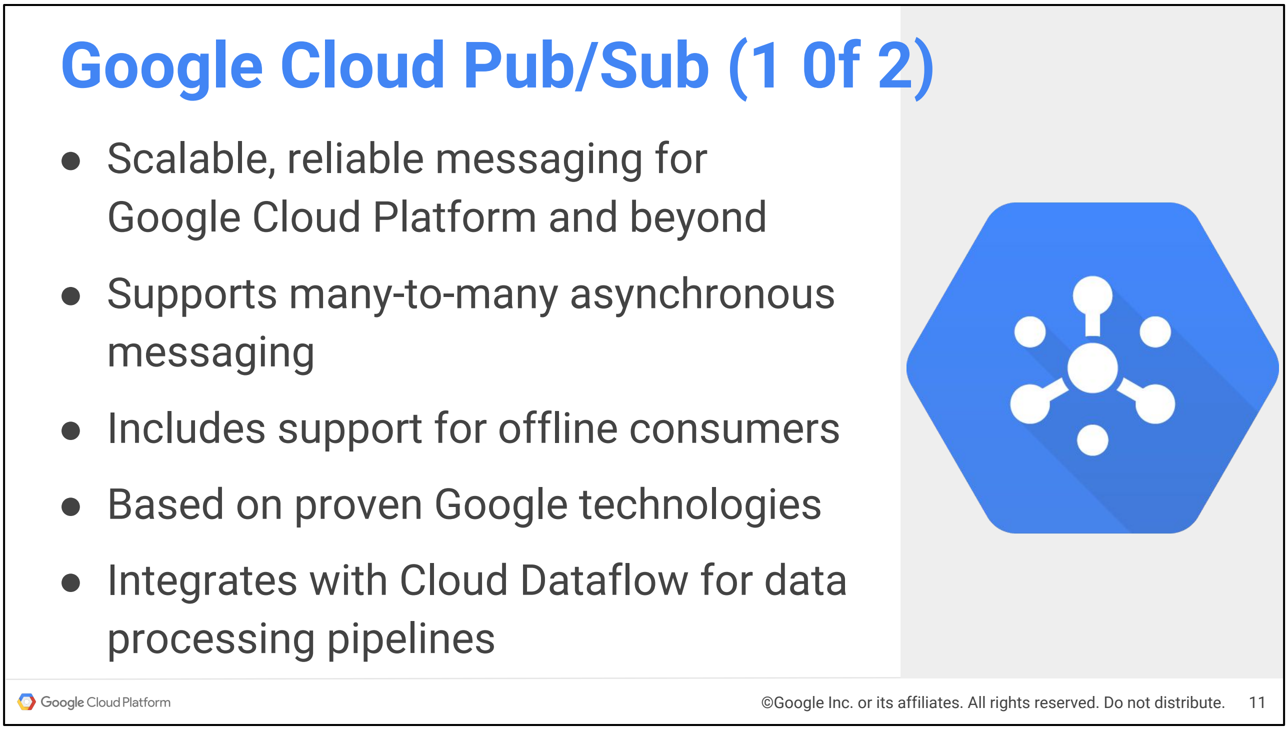
Generated analytics about the data using visually compelling charts **in real-time**

Analyzed data set of **2 billion rows** using complex queries

Experienced consistently fast **20-25 second results**

Google Cloud Platform

Notes:
For more information on the Shine technologies story, go to: https://cloudplatform.googleblog.com/2015/01/shine-technologies-reels-in-big-data.html.



Google Cloud Pub/Sub (1 0f 2)

- Scalable, reliable messaging for Google Cloud Platform and beyond
- Supports many-to-many asynchronous messaging
- Includes support for offline consumers
- Based on proven Google technologies
- Integrates with Cloud Dataflow for data processing pipelines

Google Cloud Platform

Notes:
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.

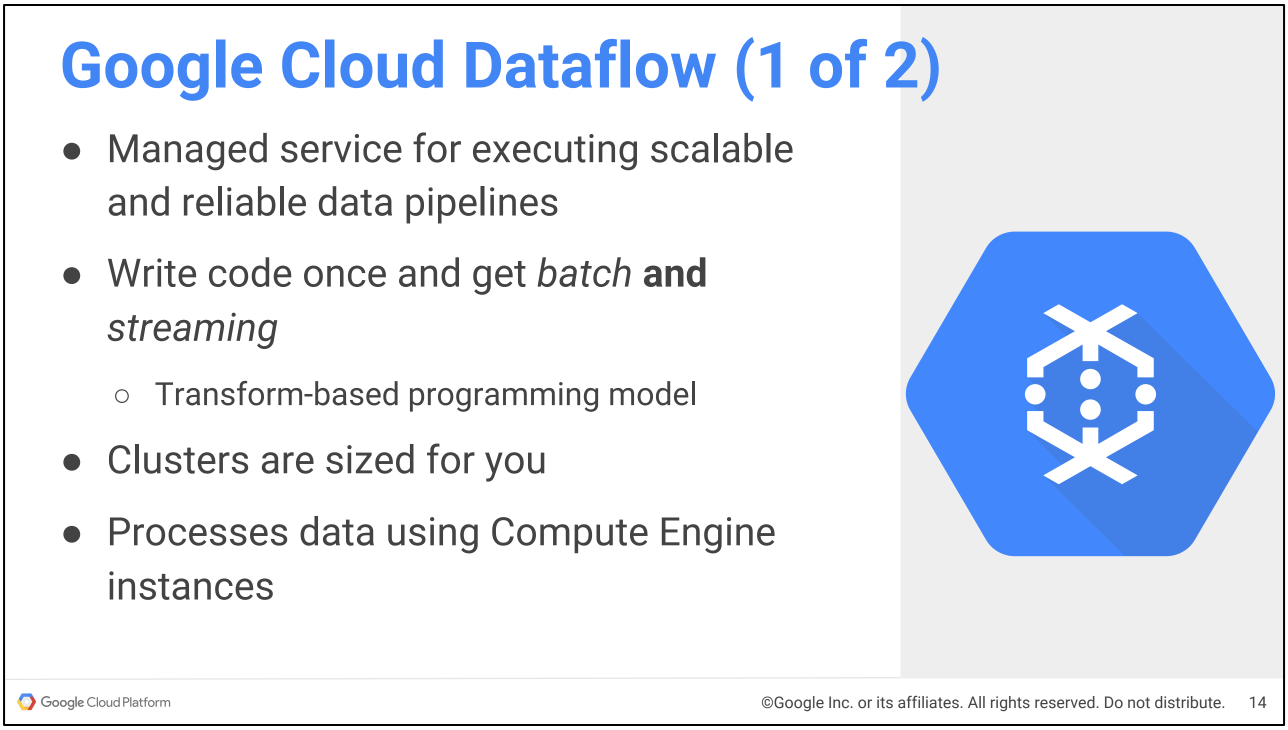
End-to-End Acknowledgement
Building reliable applications is easier with explicit application-level acknowledgements.

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.

Google Cloud Pub/Sub (2 0f 2)

- Uses push/pull subscriptions to topics
- Use cases:
    - 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 Dataflow (1 of 2)

- Managed service for executing scalable and reliable data pipelines
- Write code once and get *batch* **and** *streaming*
  - Transform-based programming model
- Clusters are sized for you
- Processes data using Compute Engine instances

Google Cloud Platform

Notes:
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.

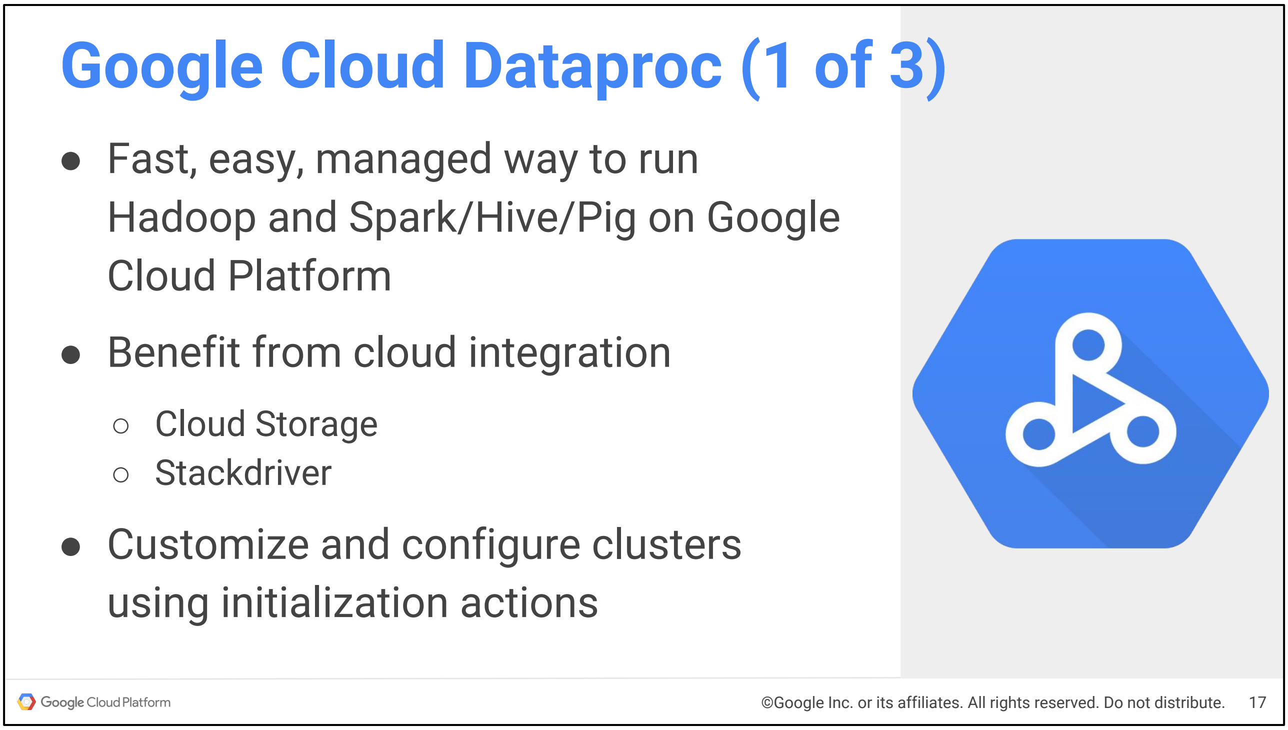
Monitoring
Integrated into the Google Cloud Platform Console, Cloud Dataflow provides statistics such as pipeline throughput and lag, as well as consolidated worker log inspection—all in near-real time.

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.

Google Cloud Dataflow (2 of 2)

- Integrates with GCP services like Cloud Storage, Cloud Pub/Sub, BigQuery, Bigtable
- Open source Java and Python SDKs
- Use cases:
  - *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



Google Cloud Dataproc (1 of 3)

- Fast, easy, managed way to run Hadoop and Spark/Hive/Pig on Google Cloud Platform
- Benefit from cloud integration
  - Cloud Storage
  - Stackdriver
- Customize and configure clusters using initialization actions

Google Cloud Platform

Notes:
Use Google Cloud Dataproc, an Apache Hadoop, Apache Spark, Apache Pig, and Apache Hive service, to easily process big datasets at low cost. Control your costs by quickly creating managed clusters of any size and turning them off when you're done. Cloud Dataproc integrates across Google Cloud Platform products, giving you a powerful and complete data processing platform.

Cloud Dataproc features:

Automated Cluster Management
Managed deployment, logging, and monitoring let you focus on your data, not on your cluster. Your clusters will be stable, scalable, and speedy.

Resizable Clusters
Clusters can be created and scaled quickly with a variety of virtual machine types, disk sizes, number of nodes, and networking options.

Integrated
Built-in integration with Cloud Storage, BigQuery, Bigtable, Cloud Logging, and Cloud Monitoring, giving you a complete and robust data platform.

Versioning
Image versioning allows you to switch between different versions of Apache Spark, Apache Hadoop, and other tools.

Developer Tools
Multiple ways to manage a cluster, including an easy-to-use Web UI, the Google Cloud SDK, RESTful APIs, and SSH access.

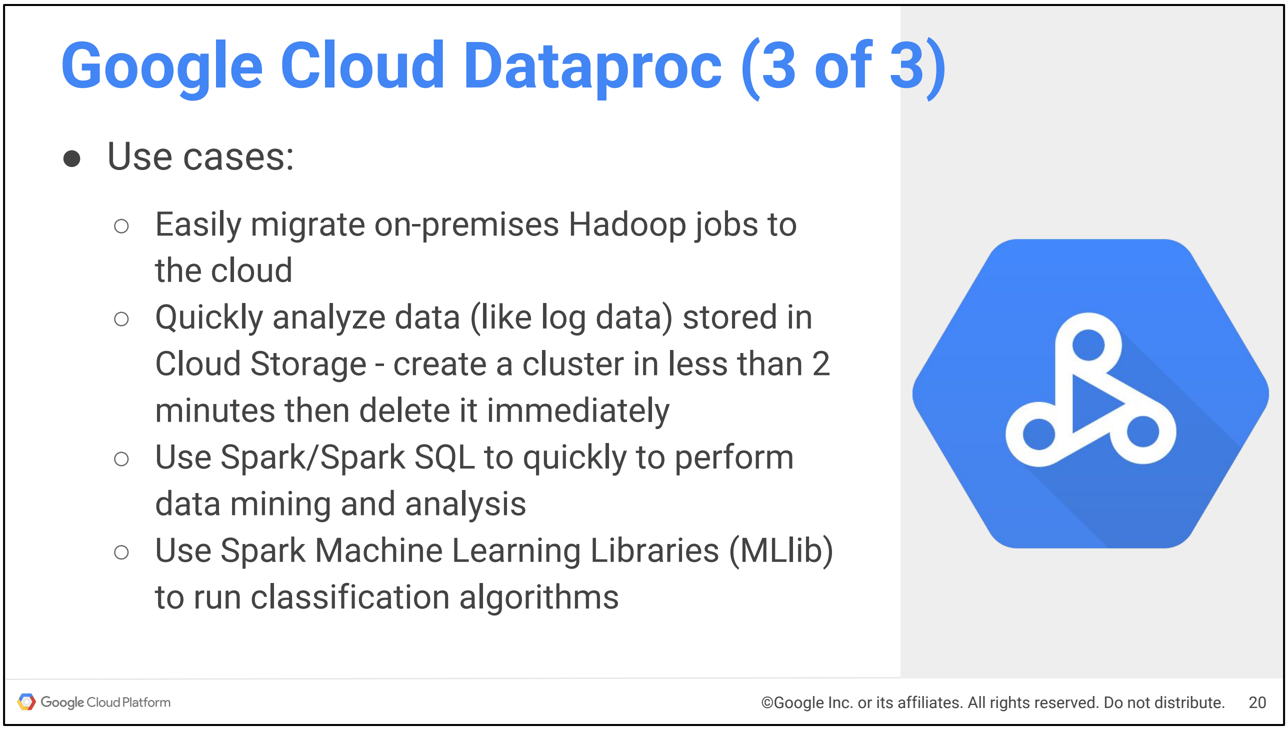
Initialization Actions
Run initialization actions to install or customize the settings and libraries you need when your cluster is created.

Automatic or Manual Configuration
Cloud Dataproc automatically configures hardware and software on clusters for you while also allowing for manual control.

Flexible Virtual Machines
Clusters can use custom machine types and preemptible virtual machines so they are the perfect size for your needs.

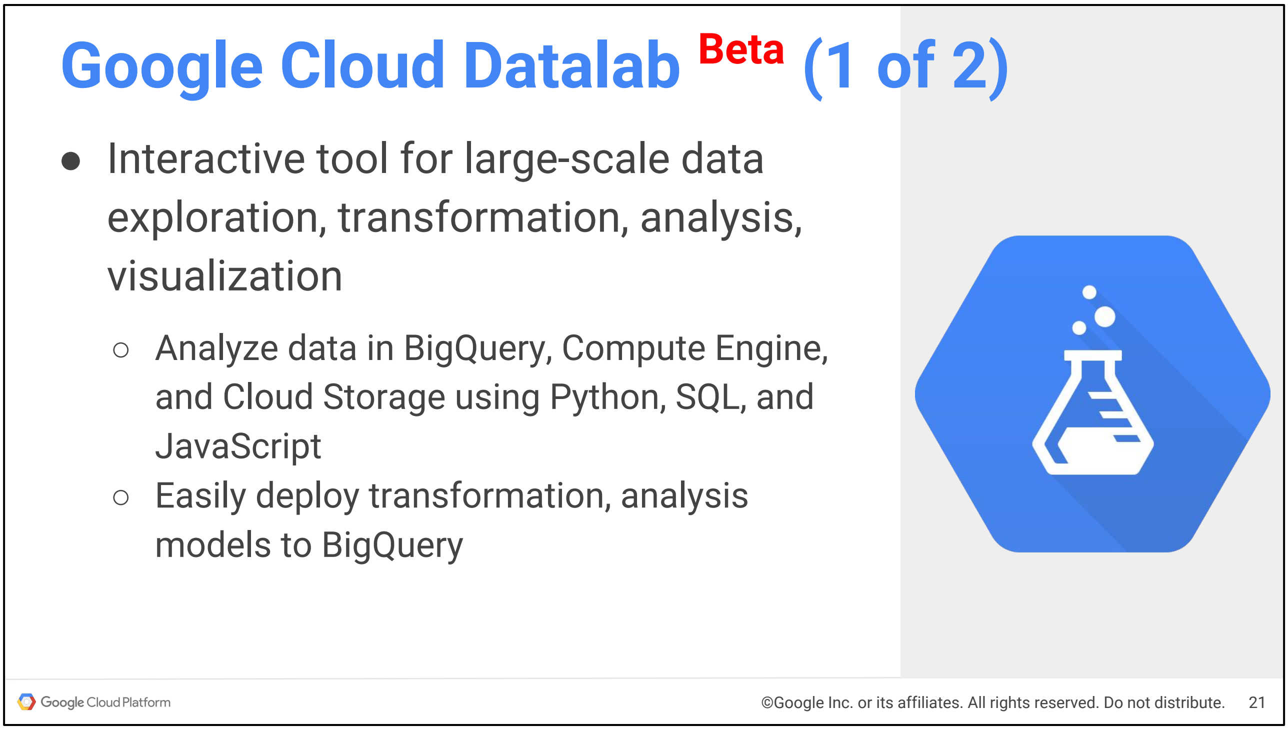
Google Cloud Dataproc (2 of 3)

- Create clusters in 90 sec or less
- Dataproc clusters billed minute-by-minute
  - Save money using preemptible instances for batch processing
- Scale clusters up and down even when jobs are running
- Developer tools
  - RESTful API
  - Integration with Google Cloud SDK



Google Cloud Dataproc (3 of 3)

- Use cases:
  - Easily migrate on-premises Hadoop jobs to the cloud
  - Quickly analyze data (like log data) stored in Cloud Storage - create a cluster in less than 2 minutes then delete it immediately
  - Use Spark/Spark SQL to quickly to perform data mining and analysis
  - Use Spark Machine Learning Libraries (MLlib) to run classification algorithms



Google Cloud Datalab Beta (1 of 2)

- Interactive tool for large-scale data exploration, transformation, analysis, visualization
  - Analyze data in BigQuery, Compute Engine, and Cloud Storage using Python, SQL, and JavaScript
  - Easily deploy transformation, analysis models to BigQuery

Notes:
Cloud Datalab is a powerful interactive tool created to explore, analyze and visualize data with a single click on Google Cloud Platform. It runs on Google App Engine and orchestrates multiple services automatically so you can focus on exploring your data.

Cloud Datalab features:

Integrated
Cloud Datalab handles authentication, 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 wishing 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.

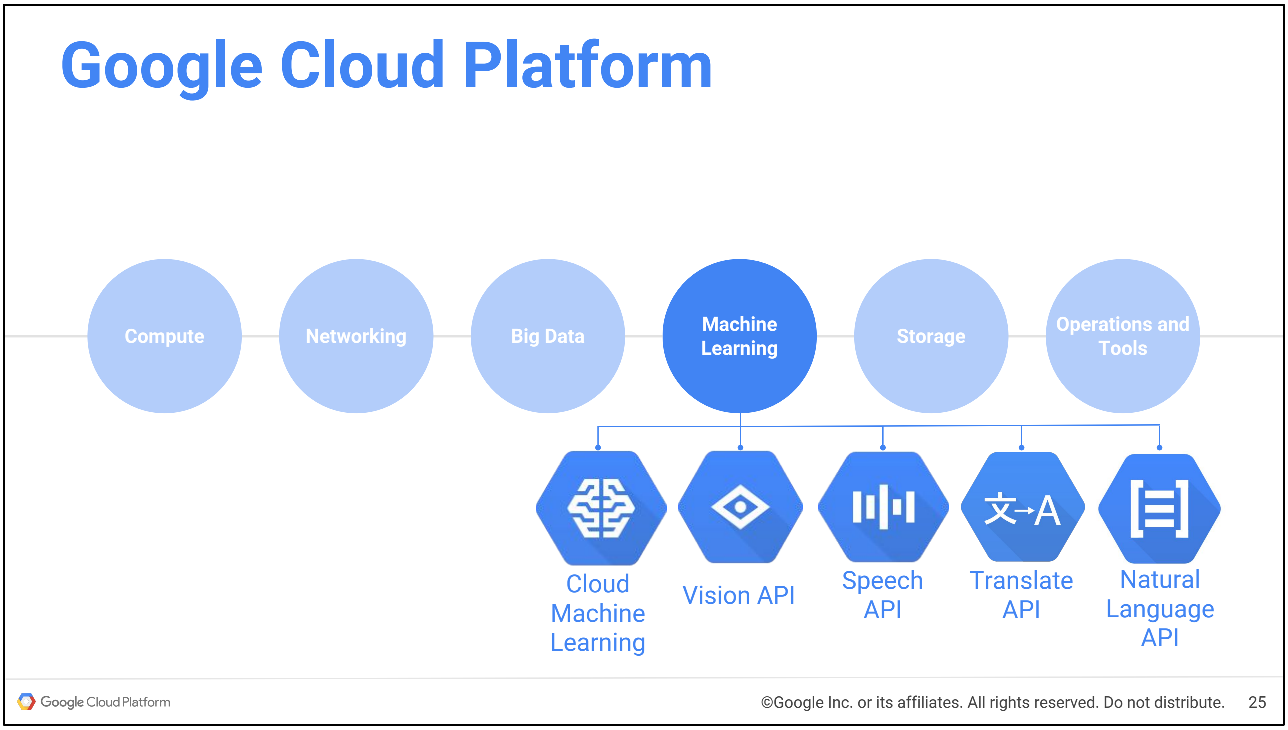
IPython Support
Cloud Datalab is based on Jupyter (formerly IPython) so you can use a large number of existing packages for statistics, machine learning, etc. Learn from published notebooks and swap tips with a vibrant IPython community.

Google Cloud Datalab Beta (2 of 2)

- Integrated, open source
  - Runs on Google App Engine
  - Built on Jupyter (formerly IPython)
  - Use Google Charts or matplotlib for easy visualizations
- Code, documentation, results, visualizations in intuitive notebook format

Agenda

1. Google Cloud Big Data Platform
2. Google Cloud Machine Learning Platform
3. Quiz & Lab



Google Cloud Platform

Compute

Networking

Big Data

Machine Learning

Storage

Operations and Tools

Cloud Machine Learning

Vision API

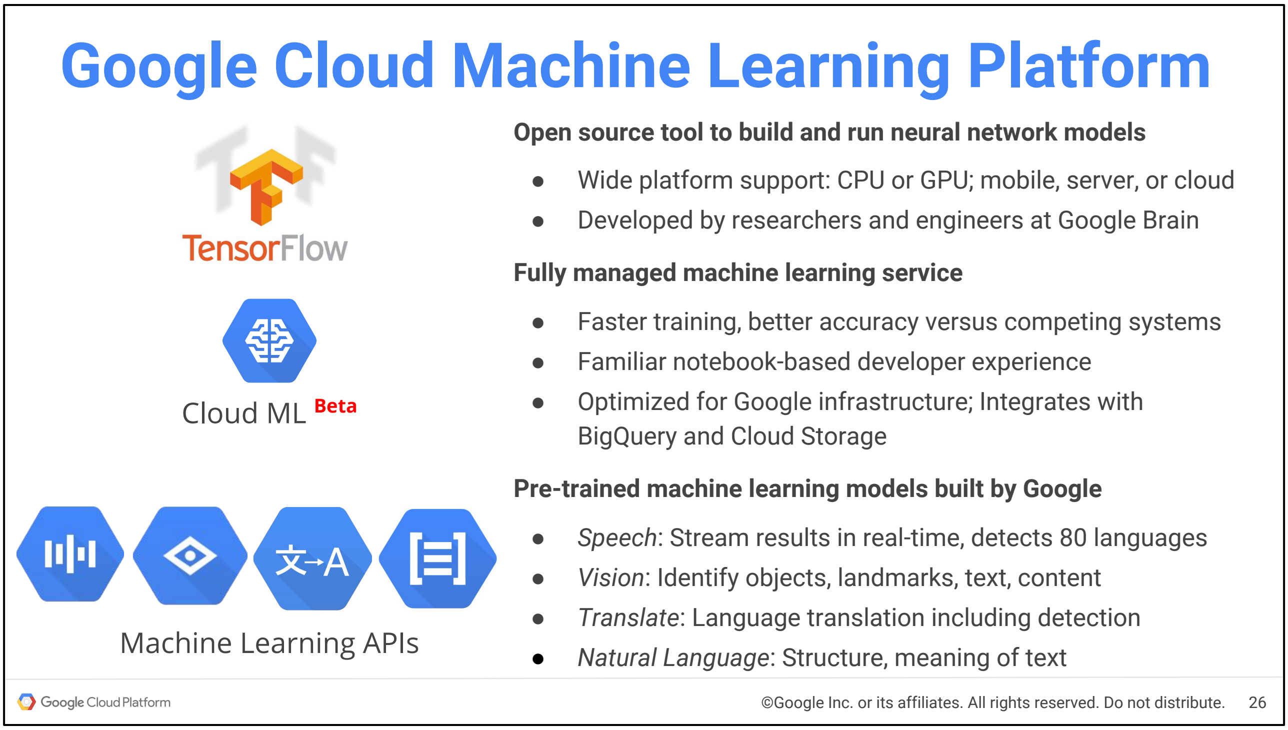
Speech API

Translate API

Natural Language API

Google Cloud Platform

Notes:
The Google Cloud Machine Learning platform includes several services to help your applications see, hear and understand the world around them including: Google Cloud Machine Learning, Vision API, Speech API, and Translate API. Each of these products and services is discussed in this module.



Google Cloud Machine Learning Platform

TensorFlow

Cloud ML Beta

Machine Learning APIs

Open source tool to build and run neural network models

- Wide platform support: CPU or GPU; mobile, server, or cloud
- Developed by researchers and engineers at Google Brain

Fully managed machine learning service

- Faster training, better accuracy versus competing systems
- Familiar notebook-based developer experience
- Optimized for Google infrastructure; Integrates with BigQuery and Cloud Storage

Pre-trained machine learning models built by Google

- *Speech*: Stream results in real-time, detects 80 languages
- *Vision*: Identify objects, landmarks, text, content
- *Translate*: Language translation including detection
- *Natural Language*: Structure, meaning of text

Google Cloud Platform

Notes:
The Google Machine Learning Platform allows you to Use your own data to create fully trained machine learning models. But where Google Cloud Platform is uniquely different is how it is enabling the future. For most customers these differences matter in multiple ways:

1) Google's infrastructure is unique and world class - built-in auto-scaling, global load balancing, built-in performance monitoring, and by far the best price-performance options of any public cloud provider.
2) Google provides a radically different level of scale and performance for data services. Google provides a comprehensive set of offerings for data - from storage to big data analytics to production databases and machine learning. These offerings let you build qualitatively different kinds of applications, and increase your company's competitiveness by reducing time to insight.
3) Google has built no-touch, operations free application platforms like Google Application Engine, and Google Container Engine - these let you scale operate and maintain your application without taking time or resources away from your development teams to configure and manage infrastructure.

Google Cloud Machine Learning provides modern machine learning services, with pre-trained models and a platform to generate your own tailored models.

Our neural net-based ML platform has better training performance and increased accuracy compared to other large scale deep learning systems. Our services are fast, scalable and easy to use. Major Google applications use Cloud Machine Learning, including Photos (image search), the Google app (voice search), Translate, and Inbox (Smart Reply). Our platform is now available as a cloud service to bring unmatched scale and speed to your business applications.

Google Cloud Machine Learning Platform makes it easy for you to build accurate, large scale machine learning models in a short amount of time. It is a portable, fully managed and integrated with other Google Cloud Data platform products such as Google Cloud Storage or Google BigQuery so you can easily train your models.

Google Cloud Vision API enables you to understand the content of an image by encapsulating powerful machine learning models in an easy to use REST API.

Google Cloud Speech API enables you to convert audio to text by applying neural network models in an easy to use API.

Google Cloud Translate API provides a simple programmatic interface for translating an arbitrary string into any supported language.

For more information on the Google Cloud Machine Learning Platform, see: https://cloudplatform.googleblog.com/2016/03/Google-takes-Cloud-Machine-Learning-service-mainstream.html

Google Cloud Machine Learning Use Cases

Structured Data

Classification/ Regression

- Customer churn analysis
- Product diagnostics
- Forecasting

Recommendation

- Content personalization
- Product X-sells/up-sells

Anomaly Detection

- Fraud detection
- Asset sensor diagnostics
- Log metric anomalies

Unstructured Data

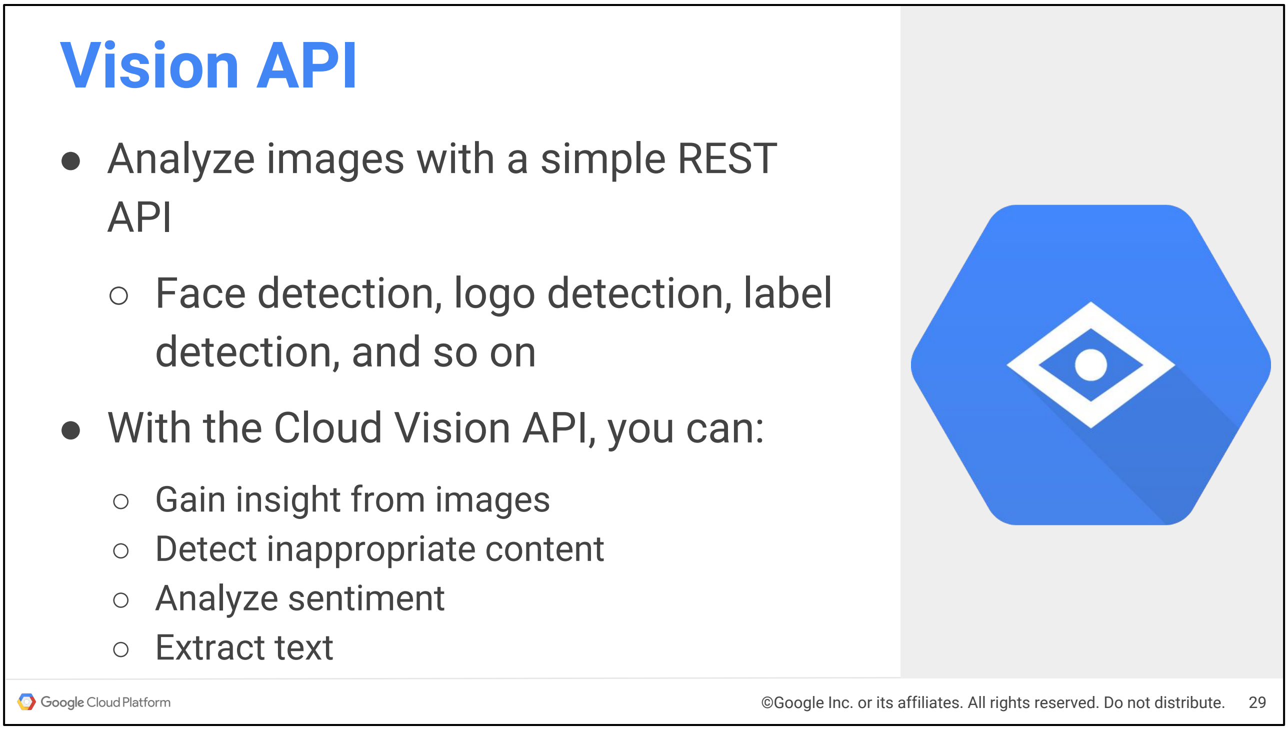
Image Analytics

- Identify damaged shipments
- Explicit content classification
- Identify “styles” in images

Text Analytics

- Call center log analysis
- Language identification
- Topic classification

Sentiment analysis

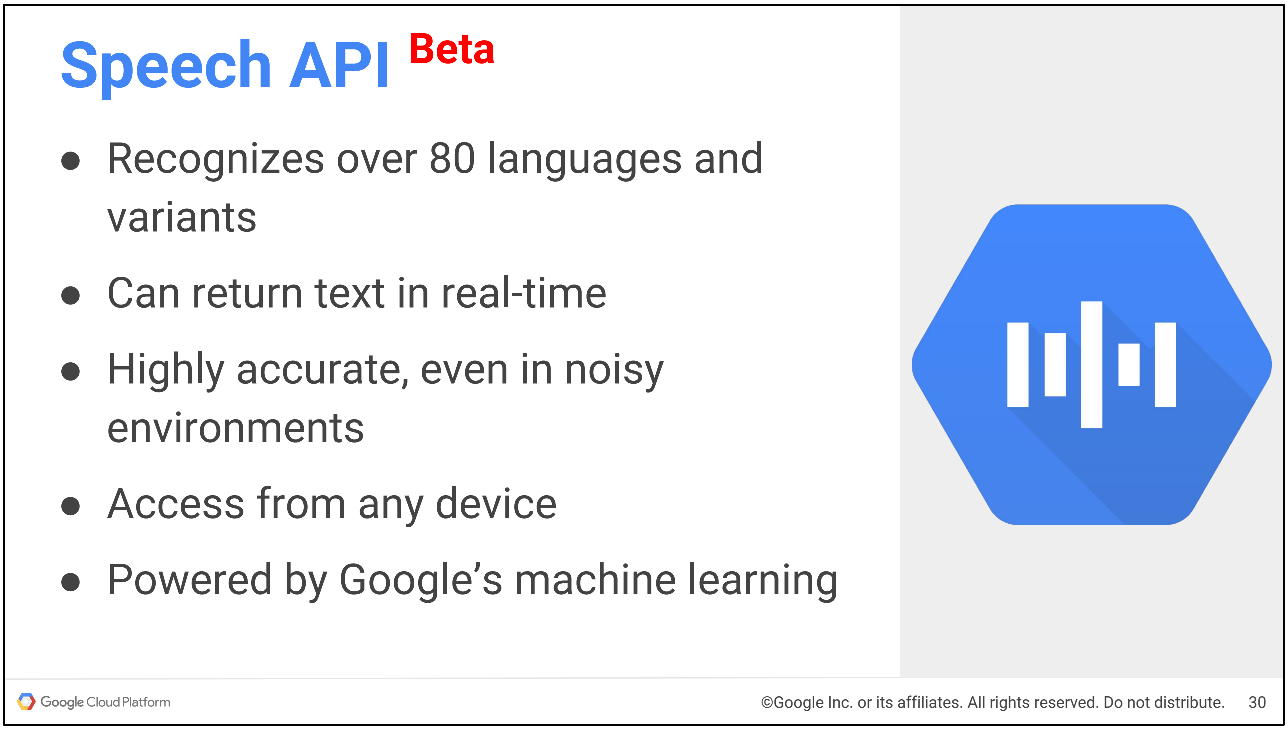


Vision API

- Analyze images with a simple REST API
  - Face detection, logo detection, label detection, and so on
- With the Cloud Vision API, you can:
  - Gain insight from images
  - Detect inappropriate content
  - Analyze sentiment
  - Extract text

Google Cloud Platform

Notes:
Google 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 and faces 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 Google Cloud Storage.

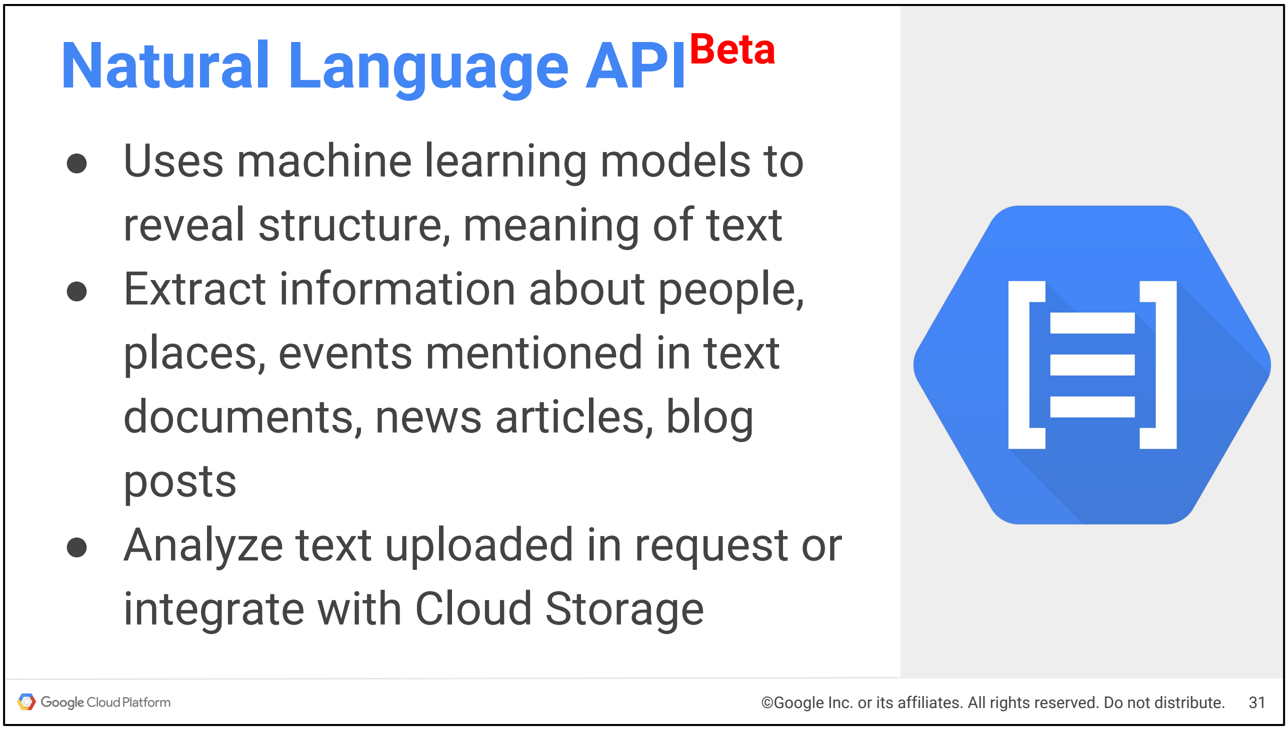


Speech API Beta

- 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 Platform

Notes:
Google Cloud Speech API enables developers to convert audio to text by applying powerful neural network models in an easy to use API. The API recognizes over 80 languages and variants, to support your global user base. You can transcribe the text of users dictating to an application's microphone, enable command-and-control through voice, or transcribe audio files, among many other use cases. Recognize audio uploaded in the request, and in upcoming releases, integrate with your audio storage on Google Cloud Storage.



Natural Language API Beta

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

Google Cloud Platform

Notes:

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.

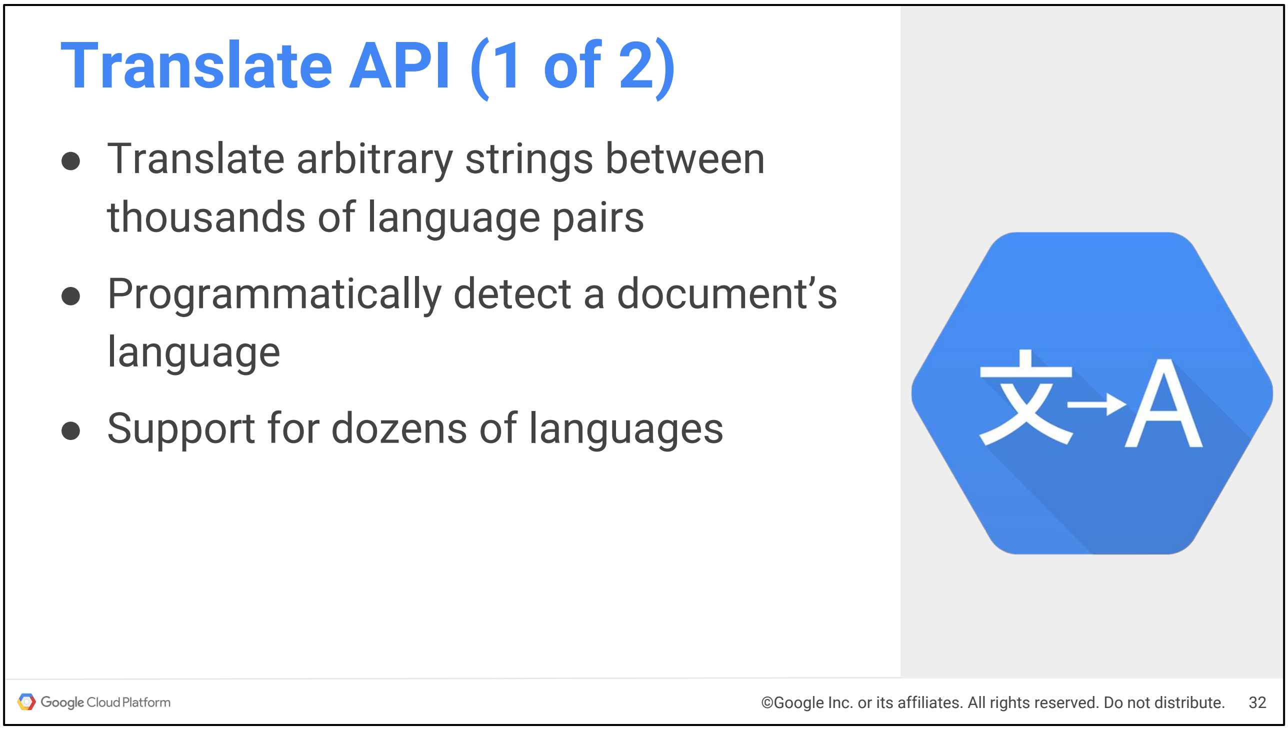
Multi-Language

- 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 Google Cloud Storage.

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



Translate API (1 of 2)

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

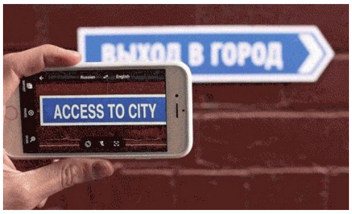
Notes:
Google Translate API provides a simple programmatic interface for translating an arbitrary string into any supported language. Translate API is highly responsive, so websites and applications can integrate with Translate API for fast, dynamic translation of source text from the source language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

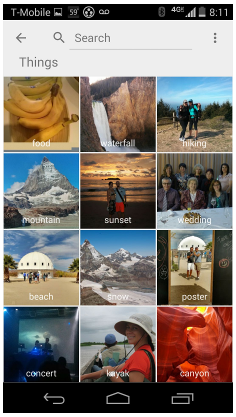
Translate API (2 of 2)

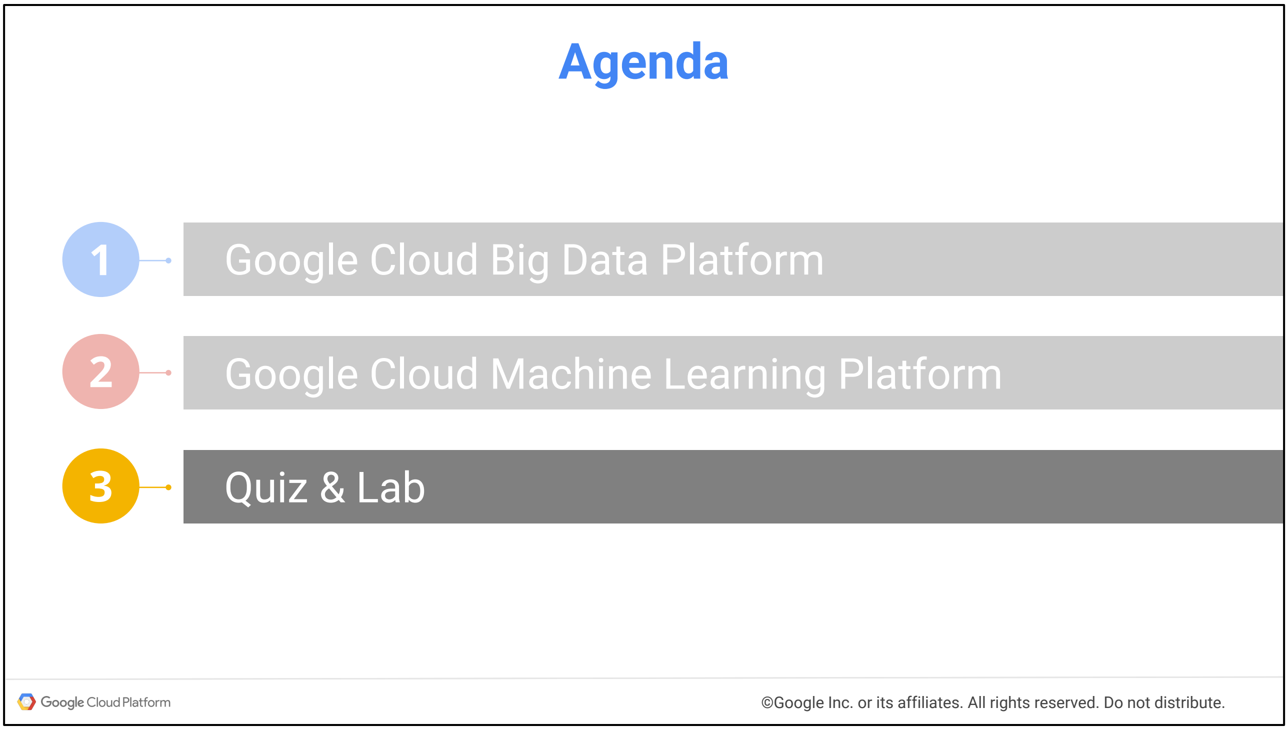
- Supports the standard Google API Client Libraries
  - Python
  - Java
  - Ruby
  - Objective-C
  - And many more
- Try it in your browser

Machine Learning APIs

Enable apps that see, hear, and understand.

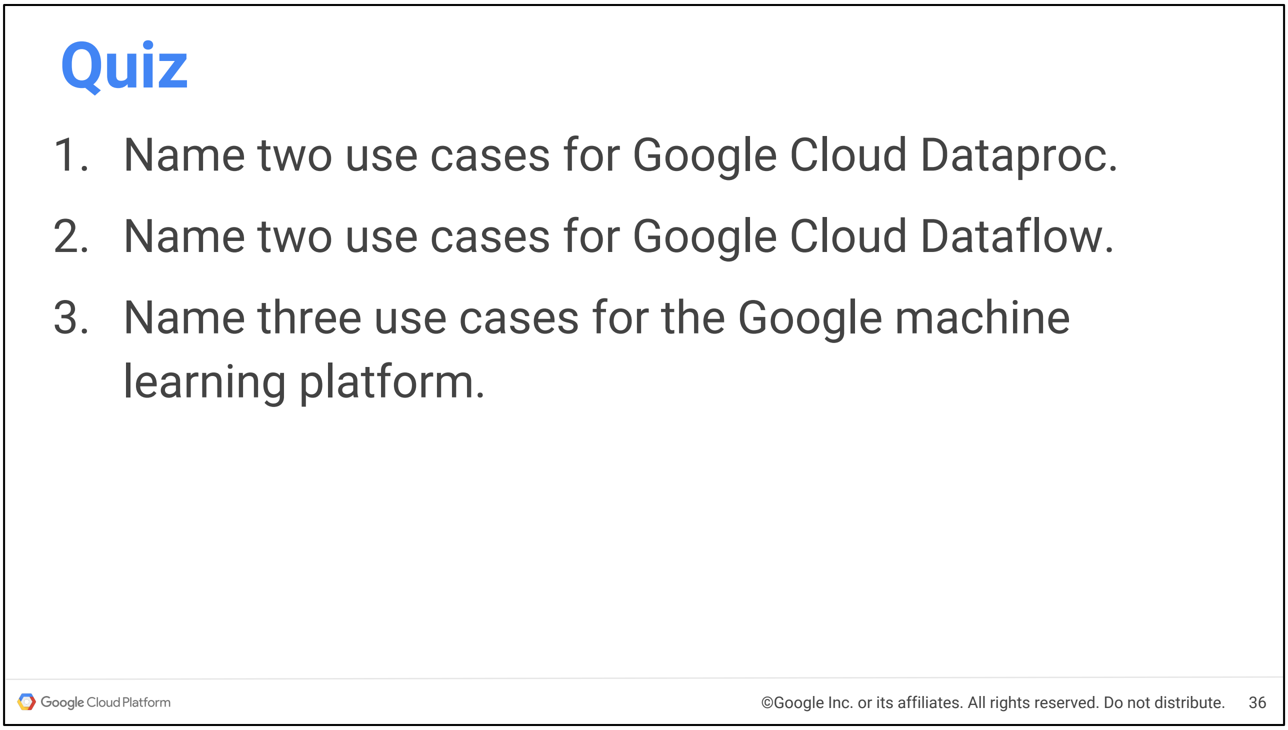






Agenda

1. Google Cloud Big Data Platform
2. Google Cloud Machine Learning Platform
3. Quiz & Lab



Quiz

1. Name two use cases for Google Cloud Dataproc.
2. Name two use cases for Google Cloud Dataflow.
3. Name three use cases for the Google machine learning platform.

Google Cloud Platform

Lab

Load data into BigQuery and analyze it.

1. Load CSV data into a BigQuery table
2. Query the data using the BigQuery web UI and the CLI

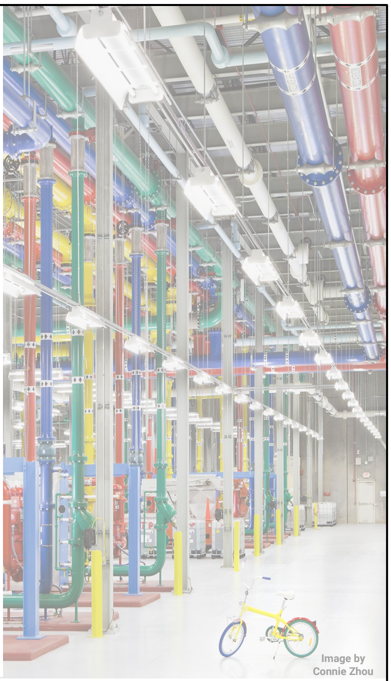
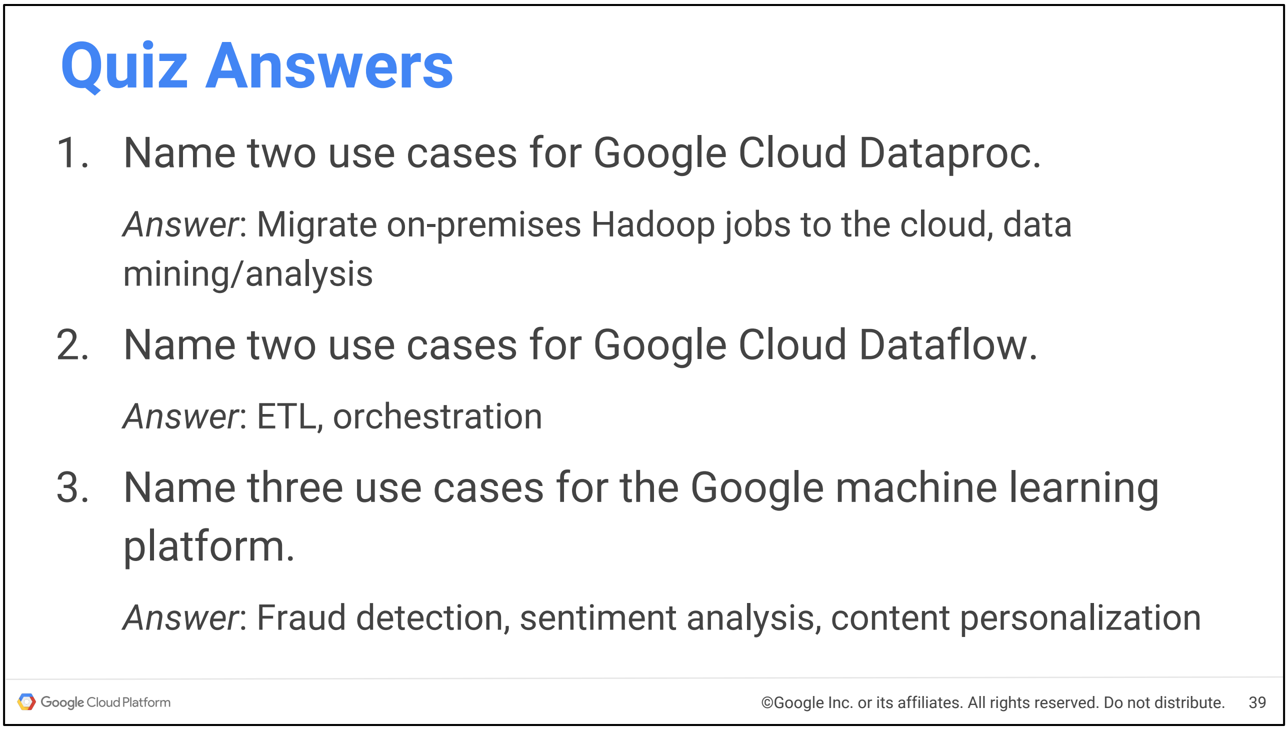


Image by Connie Zhou

Notes:
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.

Resources

- Google Big Data Platform
  https://cloud.google.com/products/#big-data
- Google Machine Learning Platform
  https://cloud.google.com/products/#machine-learning



Quiz Answers

1. Name two use cases for Google Cloud Dataproc.

   Answer: Migrate on-premises Hadoop jobs to the cloud, data mining/analysis

2. Name two use cases for Google Cloud Dataflow.

   Answer: ETL, orchestration

3. Name three use cases for the Google machine learning platform.

   Answer: Fraud detection, sentiment analysis, content personalization

cloud.google.com