Google Cloud Platform

Big Data and Machine Learning

Google Cloud Platform Fundamentals

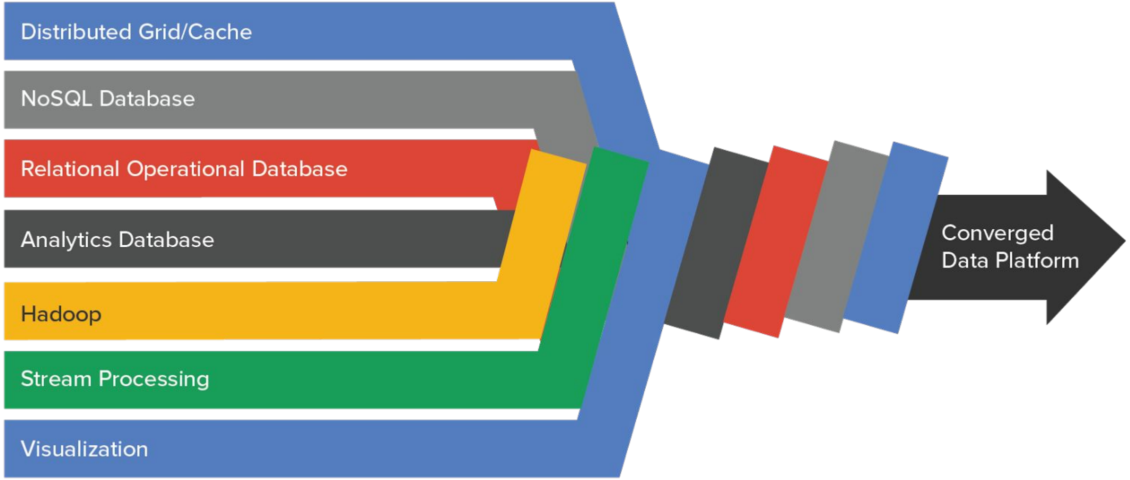
V2.0

Agenda

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

Google Cloud Big Data Platform

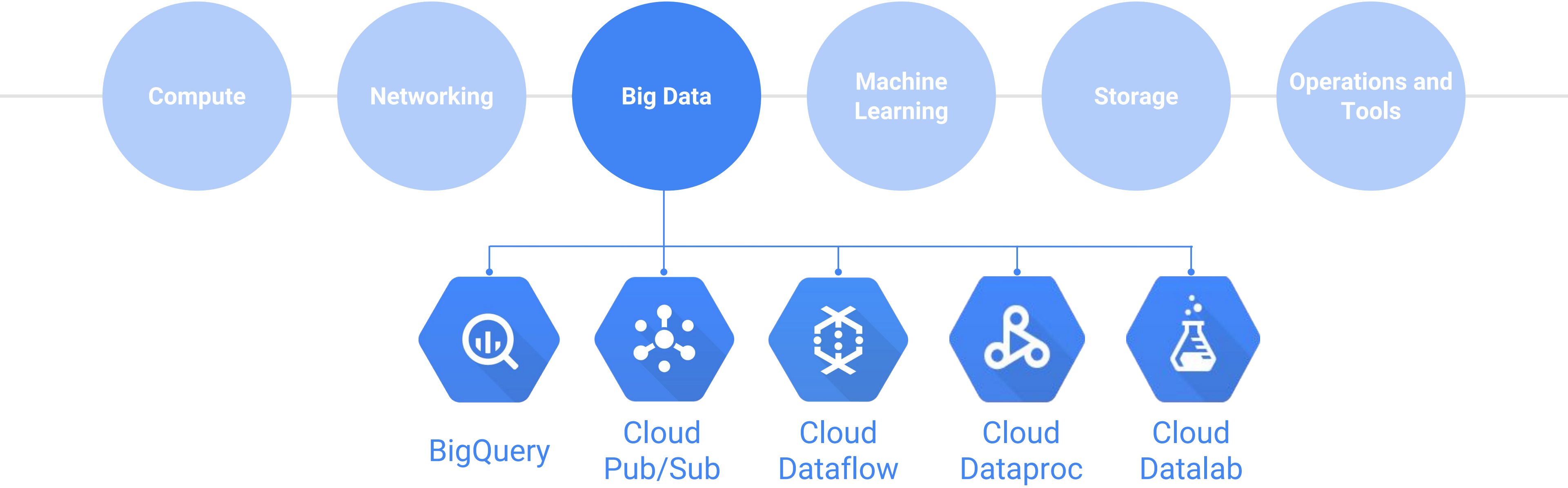
Reduces integration risk, accelerates time to value



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



Compute

Networking

Big Data

Machine Learning

Storage

Operations and Tools

BigQuery

Cloud Pub/Sub

Cloud Dataflow

Cloud Dataproc

Cloud Datalab

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

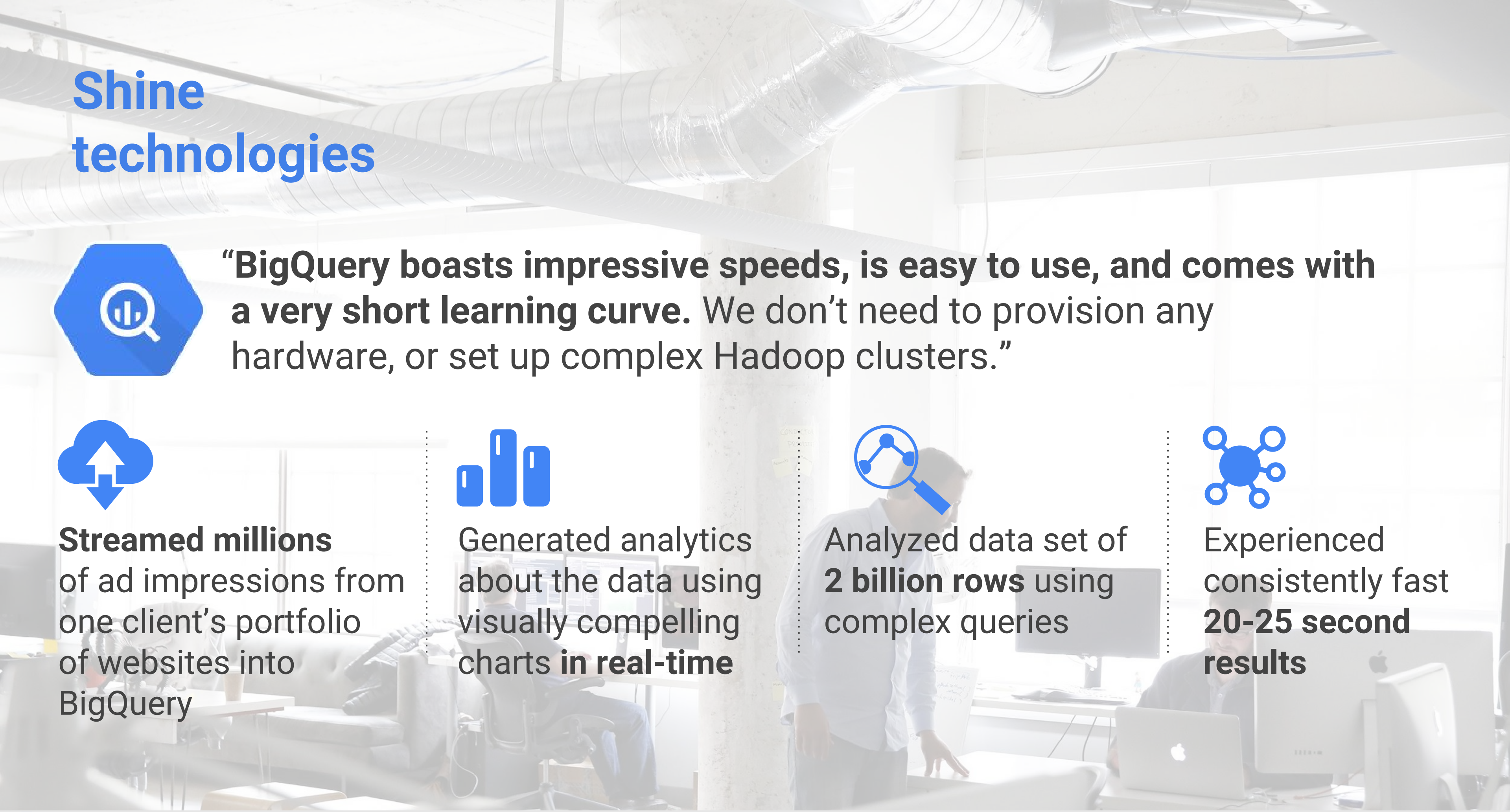
BigQuery (1 of 2)

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

BigQuery (2 of 2)

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

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 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 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 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 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

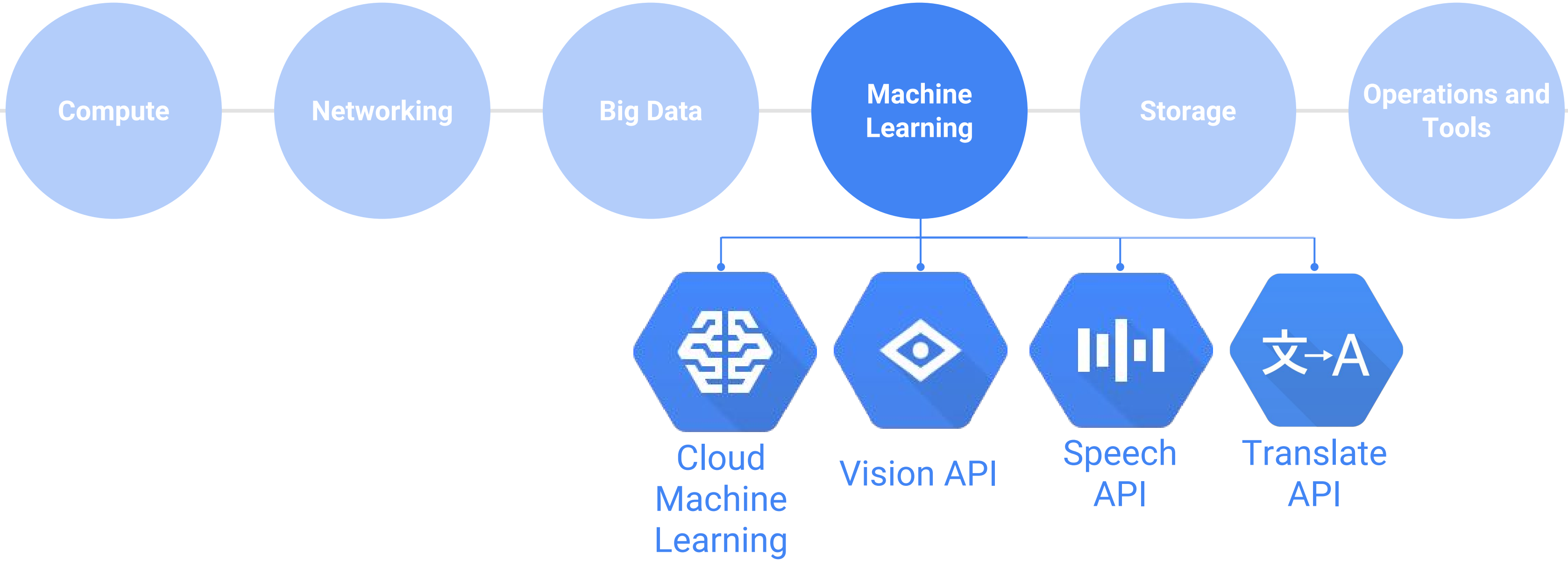
Google Cloud Datalab Beta (2 of 2)

- 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

Google Cloud Machine Learning Platform

TensorFlow

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 of Google Brain

Cloud ML Alpha

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

Machine Learning APIs

Pre-trained machine learning models built by Google

- *Vision*: identify objects, landmarks, text, explicit content detection
- *Translate*: includes language detection
- *Speech*: stream results in real-time, detects 80 languages

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

Speech API Alpha

- 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

Translate API (1 of 2)

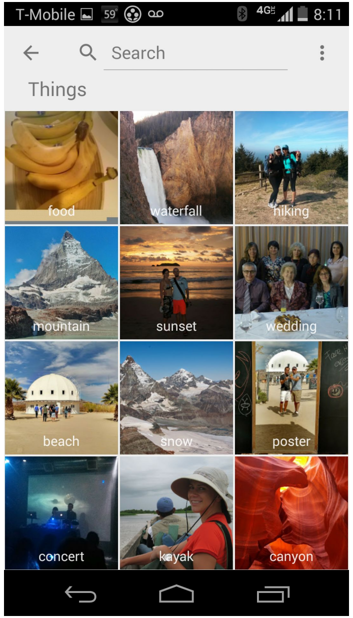
- Translate arbitrary strings between thousands of language pairs
- Programmatically detect a document’s language
- Support for dozens of languages

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.



YouTube

"Ok Google"

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.

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

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

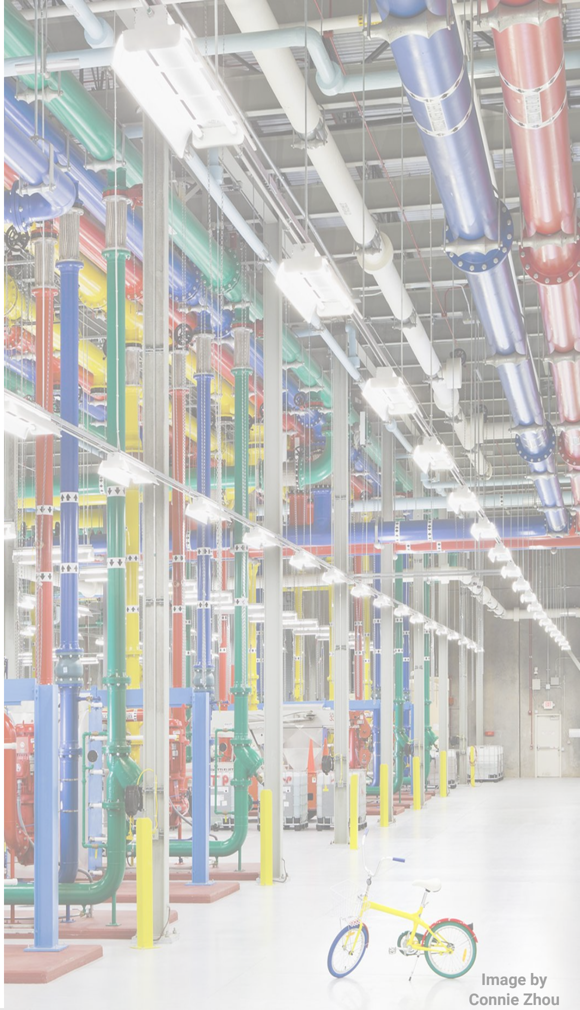


Image by Connie Zhou

Resources

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

cloud.google.com