

Diagrama del curso

Módulo 1

¿Qué es la nube?

Módulo 2

Comience con una
plataforma sólida

Módulo 3

Use Google Cloud
para compilar sus
aplicaciones

Módulo 4

¿Dónde se almacenan
los datos?

Módulo 5

Hay una API
para eso

Módulo 6

La nube no es segura,
¿verdad?

Módulo 7

Las redes ayudan

Módulo 8

Deje que Google se
encargue

Módulo 9

Ya tiene los datos,
pero ¿qué está
haciendo con ellos?

Módulo 10

Deje que las
máquinas hagan
el trabajo

Proyecto final



Ya tiene los datos, pero
¿qué está haciendo con ellos?



Introducción a los servicios administrados de macrodatos en la nube



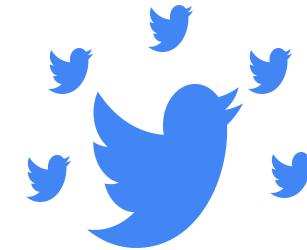
¿Qué tan grande es un petabyte de datos?



Una pila de
disquetes más alta
que 12 edificios
Empire State

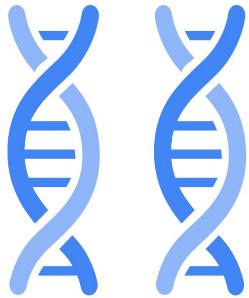


27 años
para
descargar con
4G



Cada tuit
alguna vez
tuiteado...
50 veces

¿Qué tan pequeño es un petabyte de datos?



2 microgramos
de ADN



El equivalente a un
día de videos
subidos a
YouTube

Descripción general de los servicios administrados de macrodatos



[DataProc](#)

Procese
macrodatos con
Hadoop/Spark.



[Dataflow](#)

Analice datos de
transmisión en
tiempo real.



[BigQuery](#)

Modernice las bases
de un almacén de
datos.



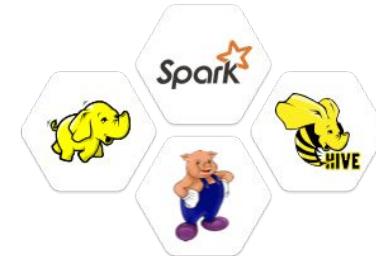
¿Qué es Dataproc?



Dataproc es un servicio administrado que se usa para el procesamiento por lotes, las consultas, la transmisión y el AA



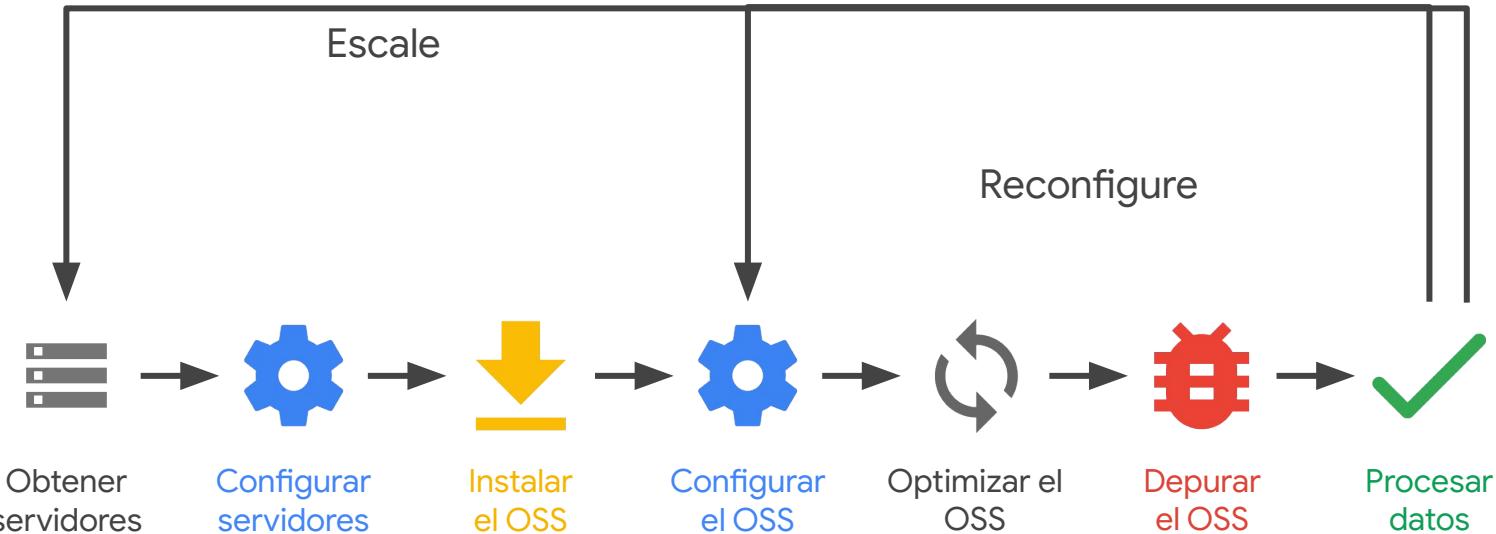
Google Cloud Platform



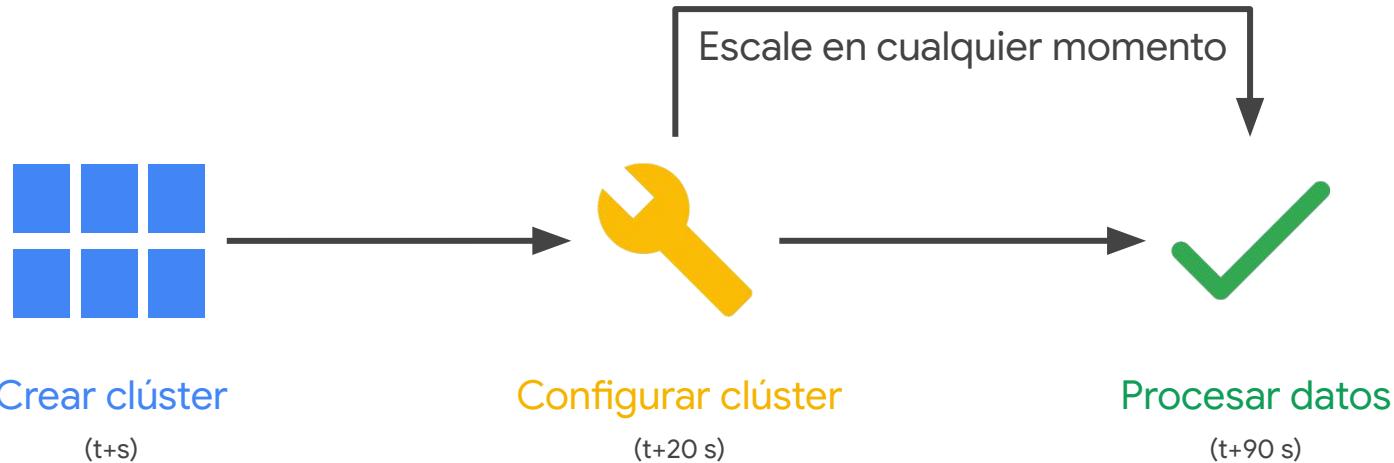
Hadoop y Spark son tecnologías de código abierto que a menudo forman la base del procesamiento de macrodatos



Clústeres típicos de Spark/Hadoop

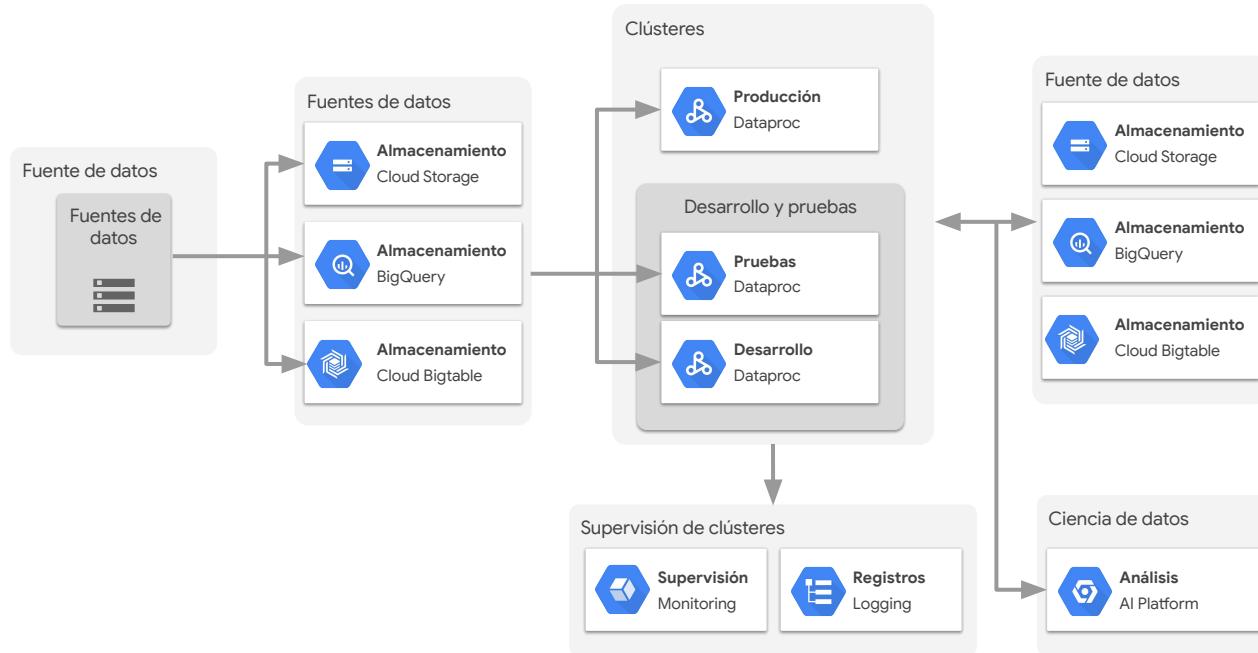


Dataproc separa el almacenamiento del procesamiento



hdfs:// ➔ gs://

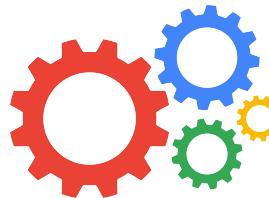
No se necesita administración ni mantenimiento



Dataproc puede ayudar con el procesamiento de registros



La necesidad



La solución



El valor

Se agregan y cargan grandes volúmenes de datos de varias fuentes en bases de datos para que puedan reunirse métricas, y así realizar informes diarios, paneles de administración y análisis.

En la actualidad, se utiliza un clúster local exclusivo para almacenar y procesar los registros con MapReduce.

Cloud Storage proporciona una opción de almacenamiento de bajo costo.

Un clúster efímero de Dataproc se puede crear en menos de 2 minutos.

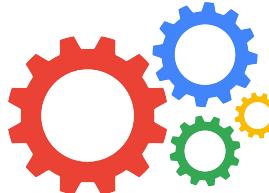
Los datos se procesan mediante el MapReduce existente.

Se ahorra dinero y se reduce la complejidad.

Dataproc puede ayudar con el análisis de datos ad hoc



La
necesidad



La
solución



El
valor

Los analistas están utilizando una shell de Spark, pero están preocupados por el aumento del uso.

No están seguros de cómo escalar su clúster, que se ejecuta en modo independiente.

Se crean clústeres que escalan según la velocidad y mitigan las fallas.

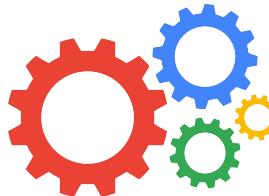
Se puede usar una interfaz web, el SDK de Cloud o una shell de Spark nativa mediante SSH.

Se desbloquea la nube sin complejidad técnica.
Los cálculos complejos se realizan en segundos, no horas.

Dataproc puede ayudar con el aprendizaje automático



La
necesidad



La
solución



El
valor

Se usan las bibliotecas de aprendizaje automático de Spark (MLlib) para ejecutar los algoritmos de clasificación en grandes conjuntos de datos.

Se depende de máquinas en la nube para instalar y personalizar Spark.

Spark y MLlib pueden instalarse en cualquier clúster de Dataproc.

Pueden aplicarse personalizaciones a los clústeres por medio de acciones de inicialización.

Se usa Cloud Monitoring para supervisar flujos de trabajo.

Los recursos pueden concentrarse en los datos, no en la creación y administración de clústeres.

La integración con Google Cloud desbloquea nuevas características de Spark.



¿Qué es Dataflow?



Dataflow ofrece procesamiento simplificado de datos de transmisión y por lotes



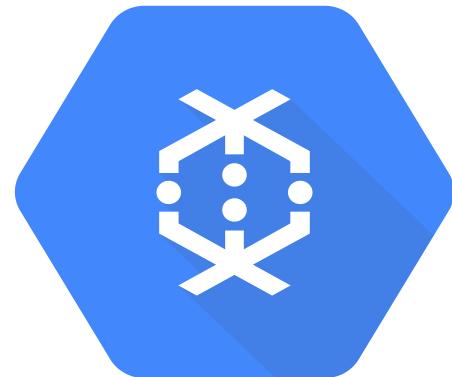
Modelo de programación unificado



Servicio completamente administrado



Integrado



Las plantillas de Dataflow permiten la implementación rápida de tipos estándar de trabajo

The screenshot shows the Google Cloud Platform Dataflow interface for creating a new job from a template. On the left, the main window has a 'Job name' field containing 'my-job-name' and a 'Cloud Dataflow template' dropdown labeled 'Select a template'. Below these are 'Run job' and 'Cancel' buttons. A blue arrow points from the 'Select a template' dropdown to the expanded list of templates on the right. The right panel is a modal or expanded view titled 'Create job from template' which lists various standard Dataflow templates categorized by processing type: 'Get Started' (Word Count), 'Process Data Continuously (stream)' (Cloud Pub/Sub Subscription to BigQuery, etc.), 'Process Data in Bulk (batch)' (Text Files on Cloud Storage to Cloud Pub/Sub, etc.), and specific types like 'Cloud Spanner to Text Files on Cloud Storage'.

Get Started
Word Count

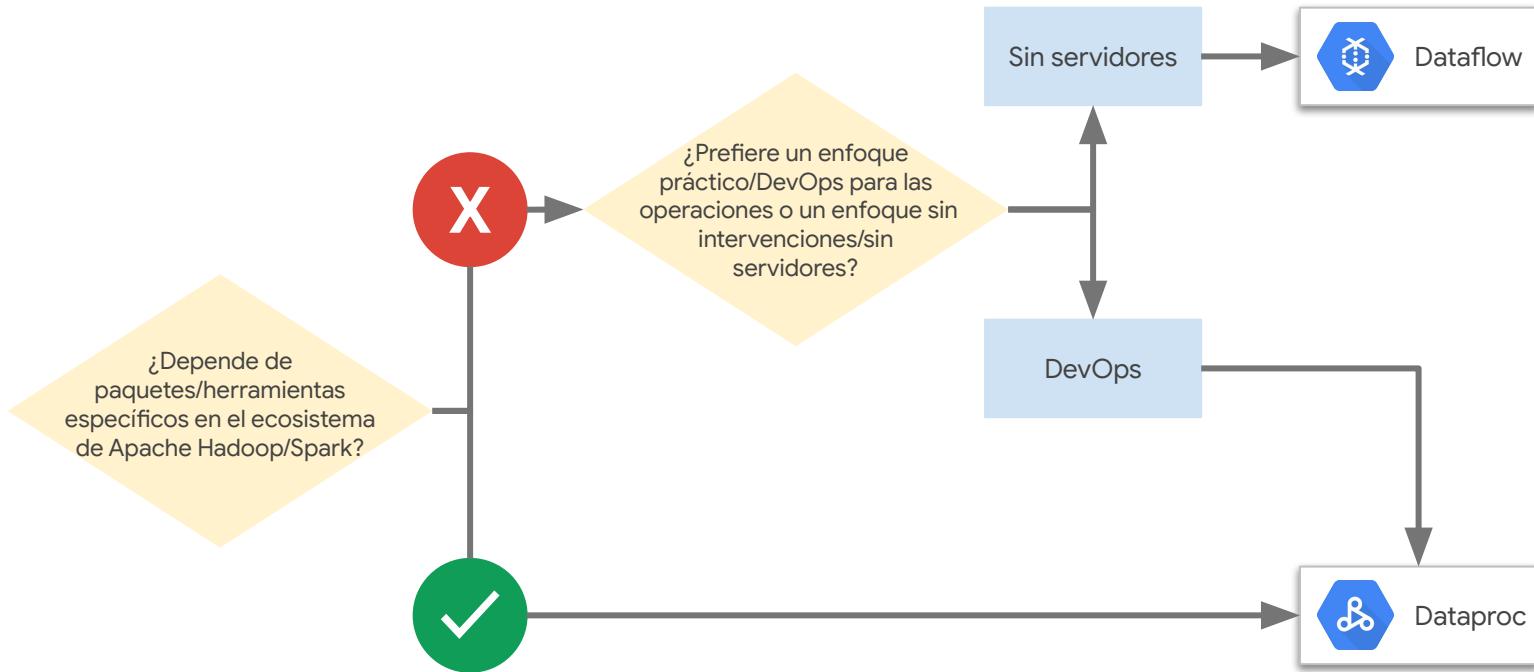
Process Data Continuously (stream)

- Cloud Pub/Sub Subscription to BigQuery
- Cloud Pub/Sub Topic to BigQuery
- Cloud Pub/Sub to Text Files on Cloud Storage
- Cloud Pub/Sub to Avro Files on Cloud Storage
- Cloud Pub/Sub to Cloud Pub/Sub
- Text Files on Cloud Storage to Cloud Pub/Sub
- Text Files on Cloud Storage to BigQuery
- Data Masking/Tokenization using Cloud DLP from GCS to BigQuery

Process Data in Bulk (batch)

- Text Files on Cloud Storage to Cloud Pub/Sub
- Text Files on Cloud Storage to BigQuery
- Cloud Datastore to Text Files on Cloud Storage
- Text Files on Cloud Storage to Cloud Datastore
- Cloud Spanner to Text Files on Cloud Storage
- Cloud Spanner to Avro Files on Cloud Storage
- Avro Files on Cloud Storage to Cloud Spanner
- Cloud BigTable to SequenceFile Files on Cloud Storage
- SequenceFile Files on Cloud Storage to Cloud BigTable
- Cloud Bigtable to Avro Files on Cloud Storage
- Avro Files on Cloud Storage to Cloud Bigtable
- Jdbc to BigQuery

Dataproc frente a Dataflow



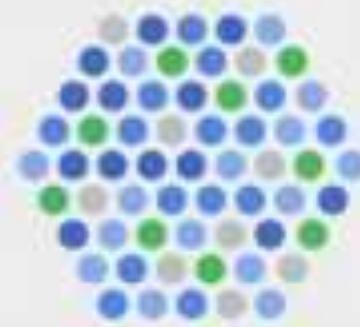
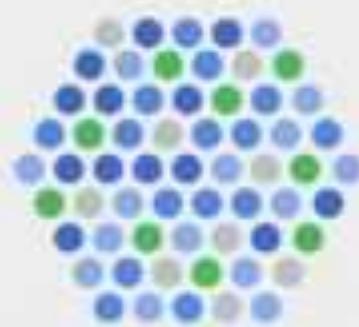
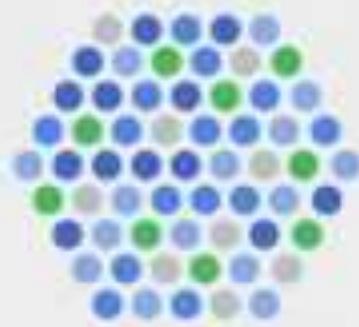
Dataproc vs Dataflow

Workload	Cloud Dataproc	Cloud Dataflow
Stream processing (ETL)	No	Yes
Batch processing (ETL)	Yes	Yes
Iterative processing and notebooks	Yes	No
Machine learning with Spark ML	Yes	No
Preprocessing for machine learning	NO	Yes (with Cloud ML)



BigQuery, el almacén de datos empresarial de Google





BigQuery es la solución de almacenes de datos de Google



Almacén de datos

BigQuery reemplaza la configuración de hardware típica del almacén de datos.



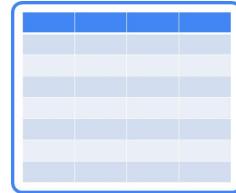
Data mart

BigQuery organiza las tablas de datos en unidades llamadas conjuntos de datos.



Data lake

BigQuery define esquemas y envía consultas directamente en fuentes de datos externas.



Tablas y vistas

Funciona de la misma manera que en un almacén de datos tradicional.



Permisos

Cloud IAM otorga permisos para realizar acciones específicas.

BigQuery ML permite a los usuarios crear y ejecutar modelos de AA en BigQuery por medio de consultas de SQL estándar

- 1 [Ejecute iniciativas de AA sin mover los datos de BigQuery.](#)
- 2 [Realice iteraciones en modelos en SQL dentro de BigQuery para aumentar la velocidad de desarrollo.](#)
- 3 [Automatice las tareas comunes de AA y el ajuste de hiperparámetros.](#)



BigQuery es un servicio completamente administrado



Vencimiento de los datos



Optimización de consultas



Copias de seguridad



Hardware



Administración de almacenamiento



Actualizaciones

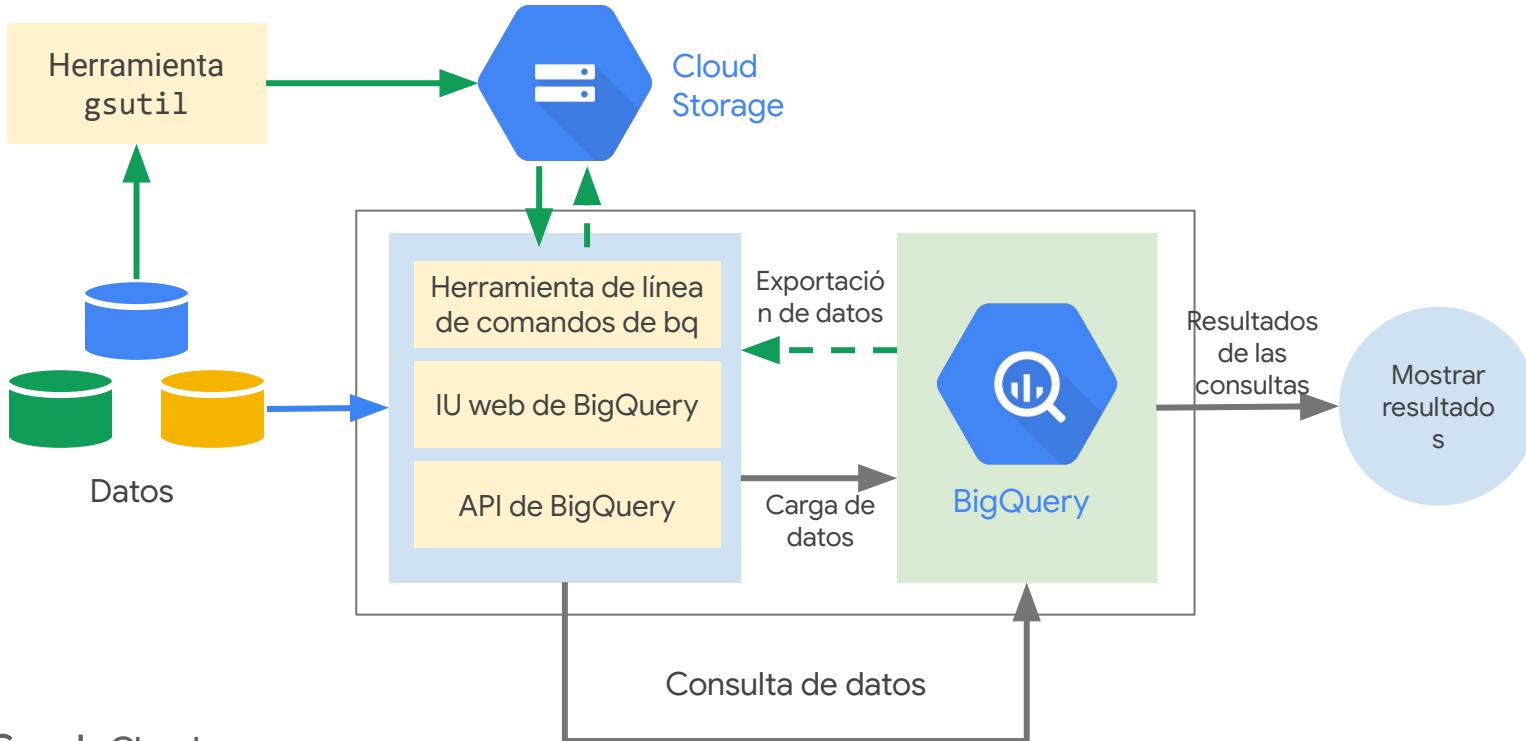


Recuperación ante fallas

Tendrá más horas de trabajo reales, ya que no deberá preocuparse por tareas comunes.



Como cargar datos en BigQuery



Test

¿Cuál de las siguientes afirmaciones acerca de BigQuery es verdadera?

- A. Se debe aprovisionar un clúster antes de usar BigQuery.
- B. BigQuery es un servicio completamente administrado.
- C. BigQuery es un frontend administrado que usa Cloud Storage.
- D. BigQuery almacena sus datos a través de discos persistentes.

Test

¿Cuál de las siguientes afirmaciones acerca de BigQuery es verdadera?

- A. Se debe aprovisionar un clúster antes de usar BigQuery.
- B. BigQuery es un servicio completamente administrado.
- C. BigQuery es un frontend administrado que usa Cloud Storage.
- D. BigQuery almacena sus datos a través de discos persistentes.

Test

¿Qué servicio administrado debería usar si quiere realizar el lift-and-shift de un clúster existente de Hadoop sin tener que reescribir su código de Spark?

- A. Dataproc
- B. Dataflow
- C. BigQuery
- D. Cloud Bigtable

Test

¿Qué servicio administrado debería usar si quiere realizar el lift-and-shift de un clúster existente de Hadoop sin tener que reescribir su código de Spark?

- A. Dataproc
- B. Dataflow
- C. BigQuery
- D. Cloud Bigtable



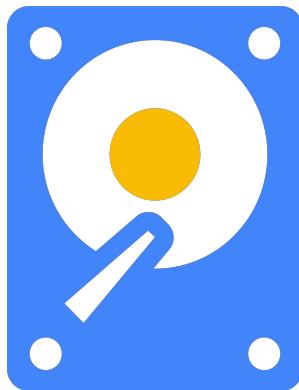
Deje que las máquinas
hagan el trabajo



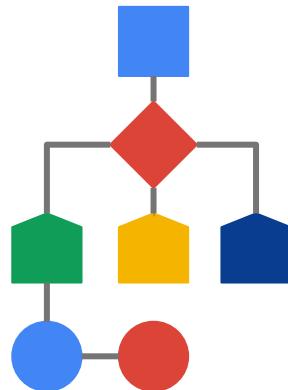
Introducción al aprendizaje automático



El AA usa algoritmos estándar para generar estadísticas predictivas a partir de datos y tomar decisiones repetidas



Datos



Algoritmo



Estadística
predictiva

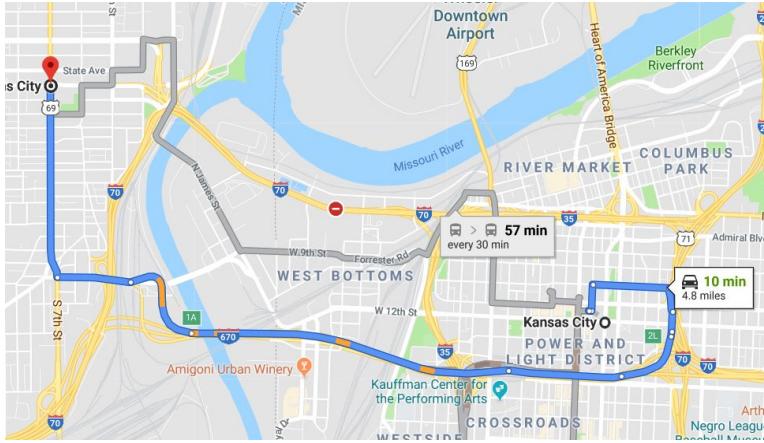


Decisión

El AA usa algoritmos estándar



Calcule los impuestos que debe pagar



¿Cuánto tardará en llegar a su casa?

El entrenamiento de modelos requiere ejemplos



Ejemplos de declaraciones de impuestos



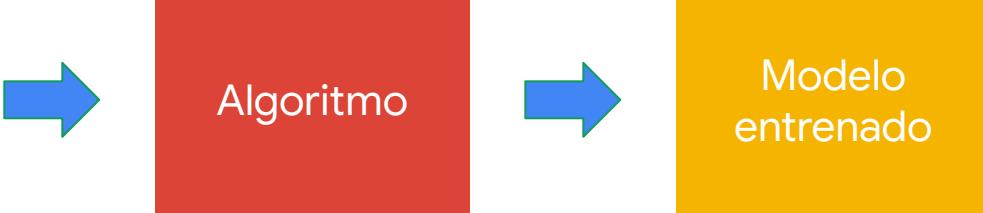
Ejemplos de viajes

Entrene un modelo de AA con ejemplos

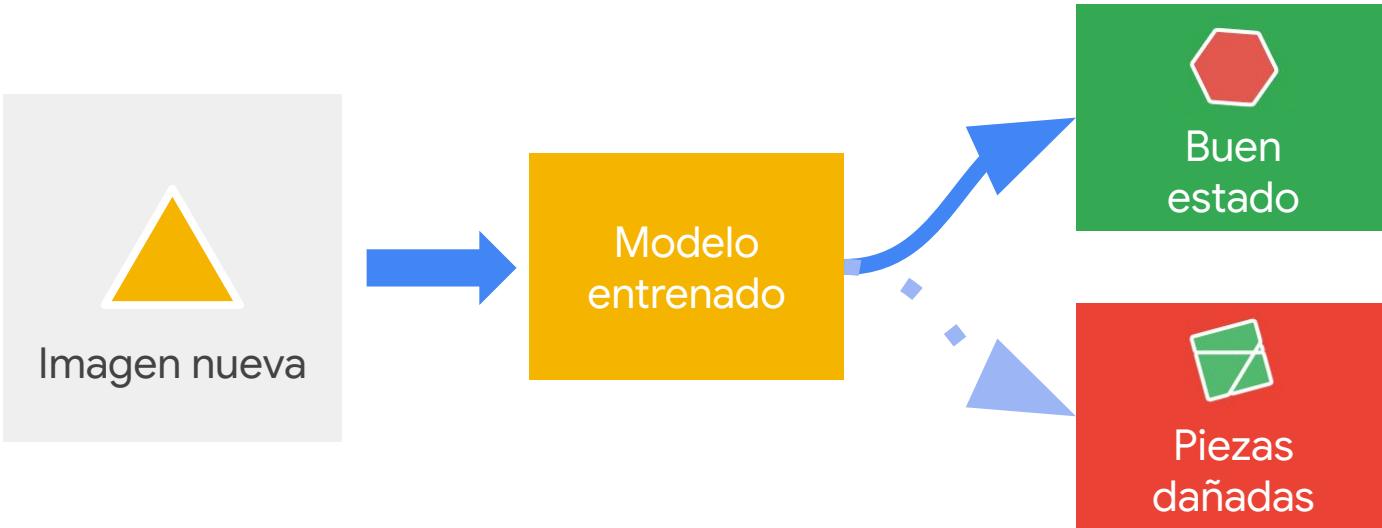
Etiquetas



Entradas



Realice predicciones con un modelo entrenado



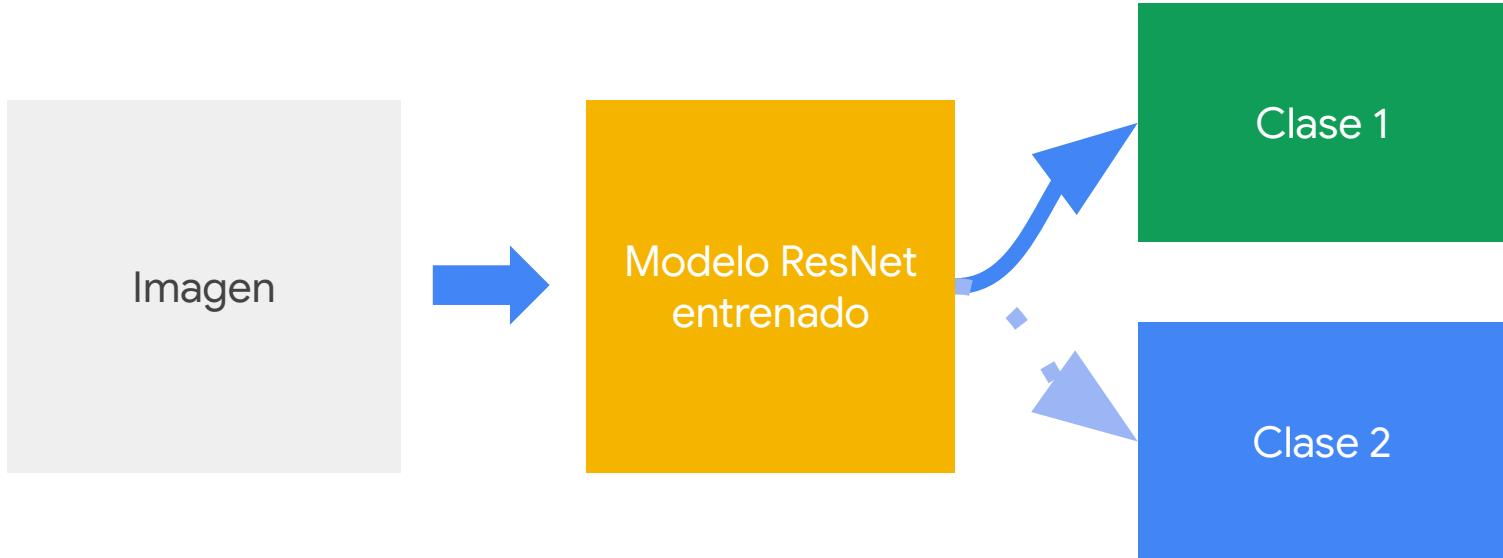
Casos de uso de algoritmos estándar

1 Detecte un patrón en una imagen.

2 Prediga el futuro de una serie temporal.

3 Comprenda o transcriba voz humana o texto.

Un algoritmo estándar para clasificación de imágenes



El mismo algoritmo aplicado a otros datos genera un modelo diferente



El algoritmo es el mismo, pero el modelo entrenado es diferente



Modelo de
imagen
(entrenado)



Nenúfar

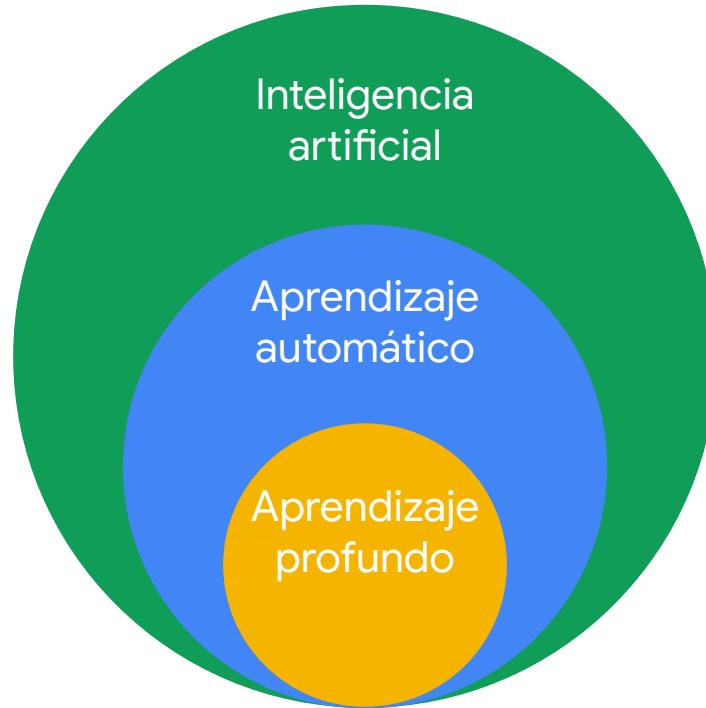


Modelo de
imagen
(entrenado)



Pieza
dañada

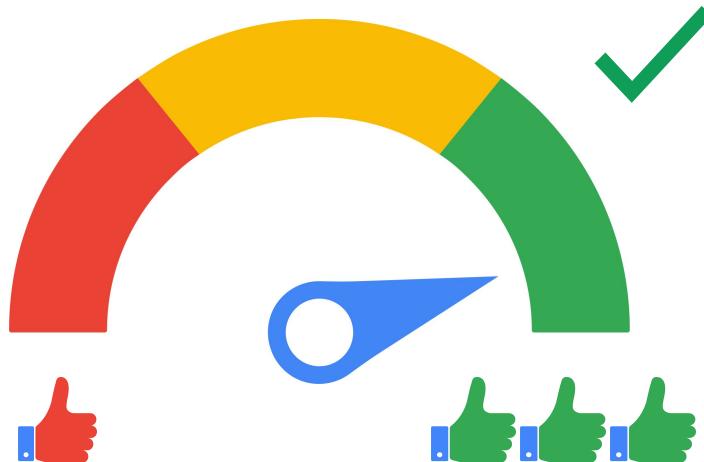
El AA es un tipo de IA



El impacto del AA es la escala

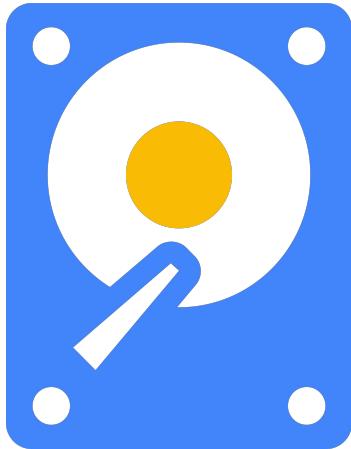


No se trata de ahorrar dinero

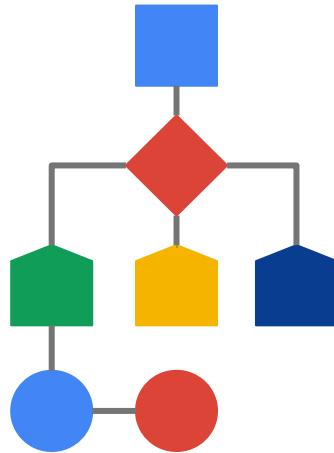


Sino de hacerlo a mayor escala

Ya no hay barreras de acceso



Datos

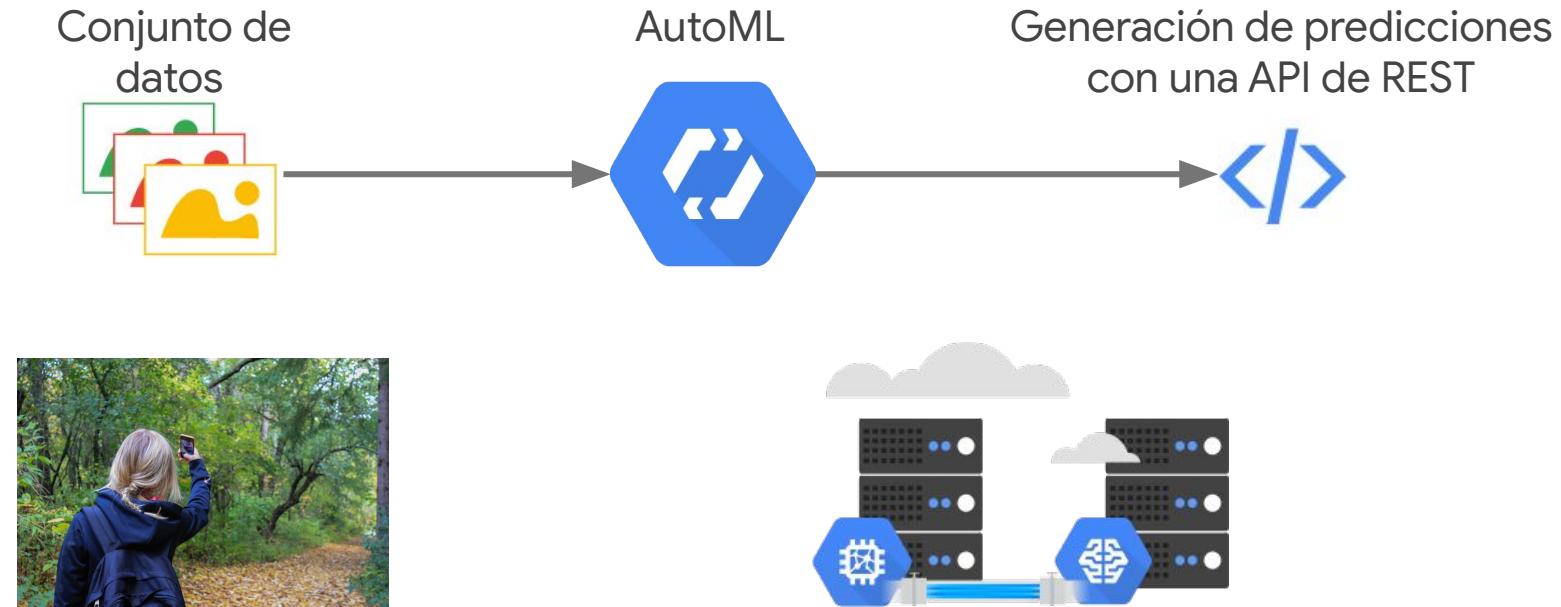


Algoritmo



Hardware
Software

¿Cómo se transforma esto en un modelo de AA para identificar hojas enfermas?



El espectro del aprendizaje automático de Google Cloud

Frameworks de AA

Más control para usuarios avanzados



TensorFlow



AI Platform

AutoML

Use sus propios datos



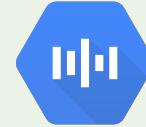
Vision
Video Intelligence
Natural Language
Translation
Tablas de datos

Modelos de AA previamente entrenados

Listos para usar



API de Vision



API de Speech-to-Text



API de Cloud Talent Solution



API de Cloud Translation



API de Cloud Natural Language



API de Video Intelligence

Para divertirse con el AA: Corre, Dibuja



quickdraw.withgoogle.com

Las aplicaciones de IA modernas usan el aprendizaje profundo

