Ingeniería de datos en Google Cloud

UNIVERSIDAD DE BURGOS

¡Hola!

Soy Mario

Máster en Ingeniería Informática

mario@mjuez.com

@mjuez

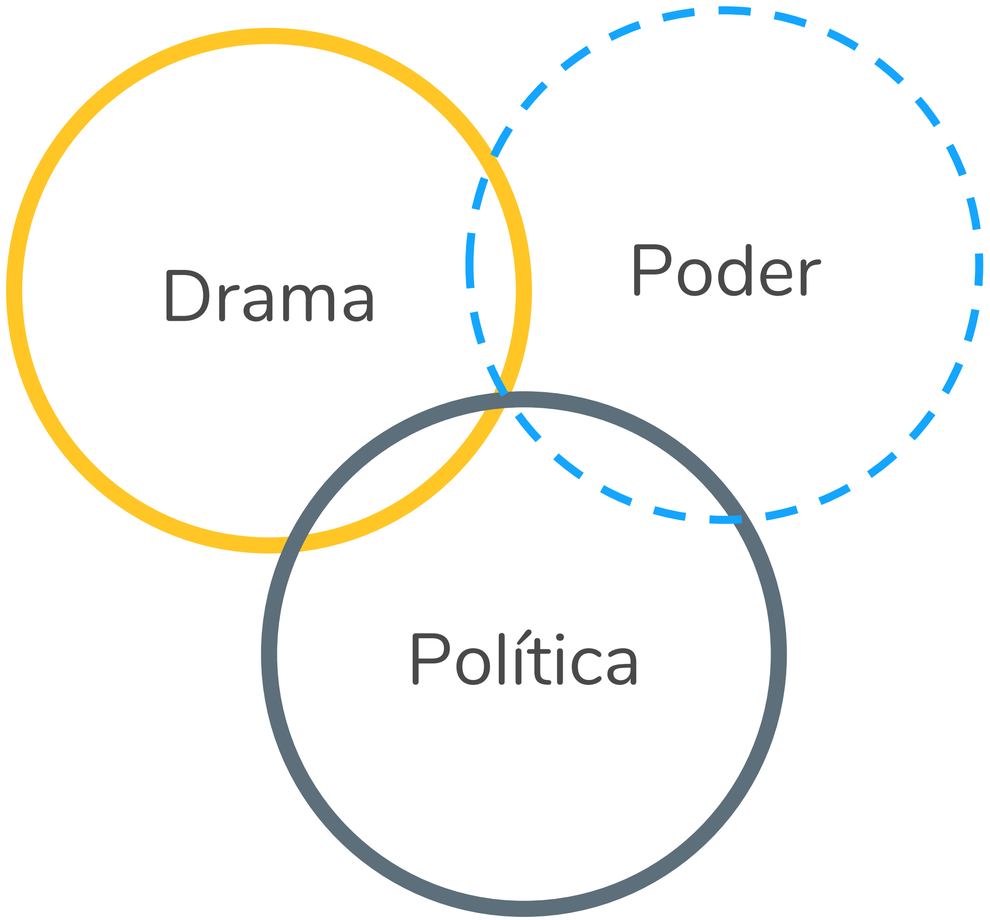
¿De qué me vas a hablar?

- Introducción a Big Data
- Computación en la nube
- Soluciones de Google Cloud para el análisis de datos
- El nuevo modelo de computación sin servidor
- Comparativa entre Google, Amazon y Microsoft Azure

BIG DATA

Para muchos, el nuevo oro

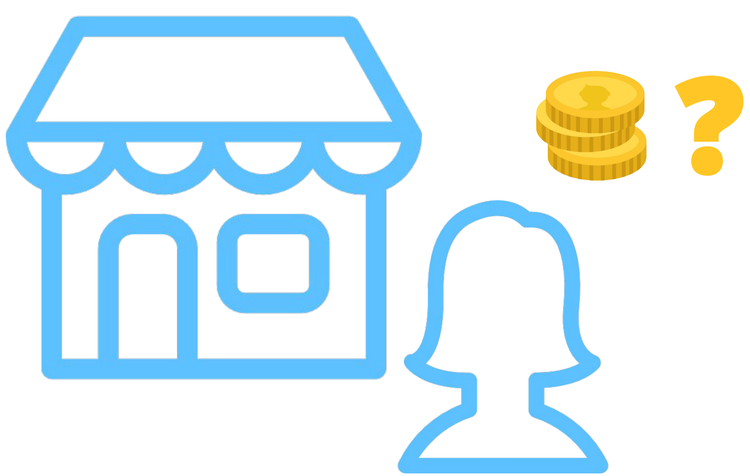
¿De verdad?


Drama
Poder
Política


NETFLIX
KEVIN
SPACEY
ROBIN
WRIGHT
A NETFLIX ORIGINAL SERIES
HOUSE of CARDS

TRÁFICO EN TIEMPO REAL

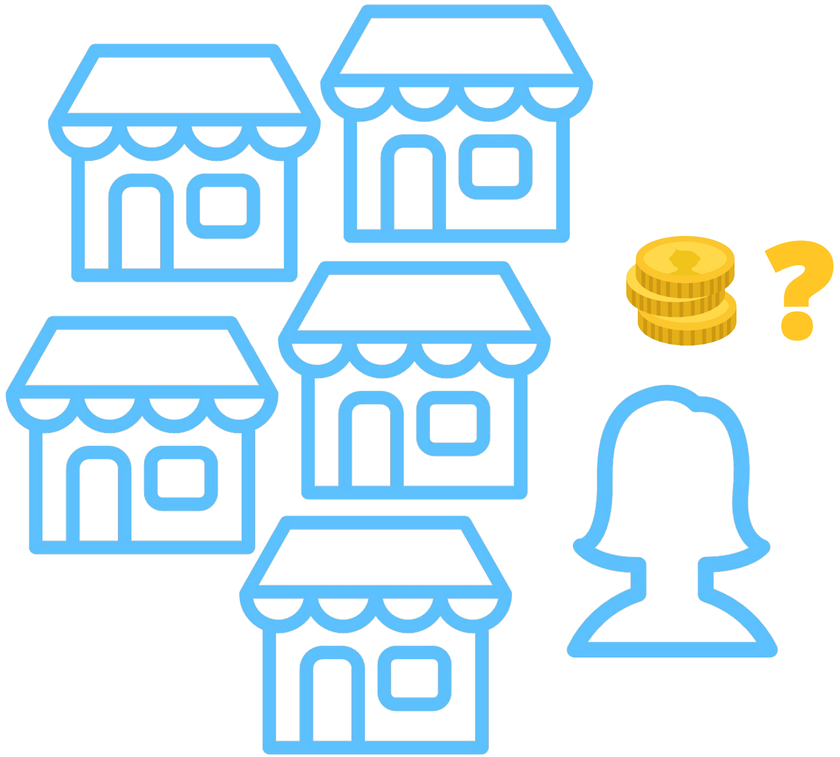
Tenemos un pequeño negocio local

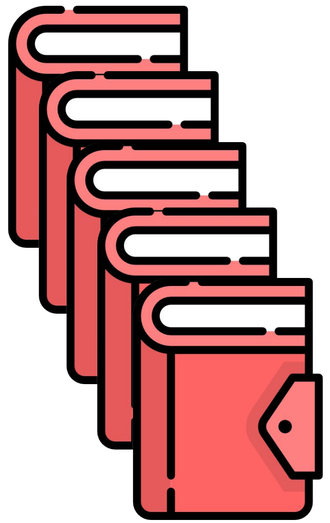


11 DICIEMBRE 2016 (DOMINGO)
LUNES (12 DICIEMBRE 2016)

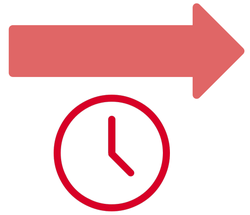
250€

El negocio crece...



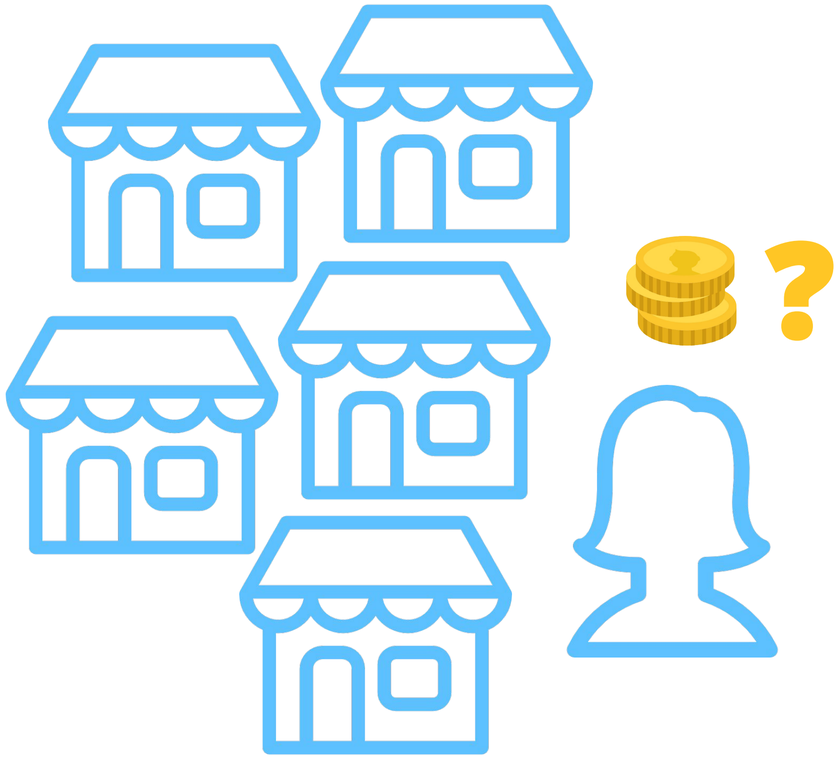


Media de años anteriores
con diferentes pesos por local



1500€

El negocio crece...

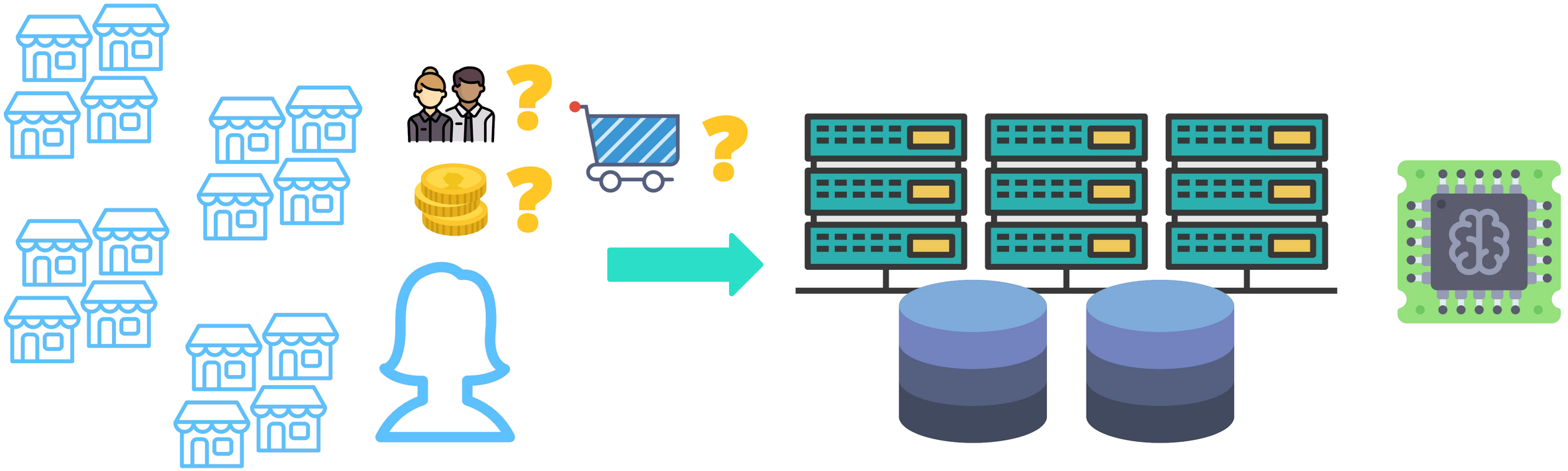


Media de años anteriores
con diferentes pesos por local

1500€

Además de las cajas, ahora también registra
qué productos vende y **a qué horas**.

Y sigue creciendo...



Hacia la computación en la nube

Opción 1:
Centro de Datos Físico

Servidores físicos propios
Tu propio edificio
Requiere gestión propia

Opción 2:
Centro de Datos Virtualizado (+ escalabilidad)

Alquiler de recursos virtuales
Pagas por lo que utilizas
Requiere gestión propia

Opción 3:
Servicios sin servidor (Máxima escalabilidad)

Totalmente gestionados
Escalabilidad automática
Permite centrarse en el análisis de datos

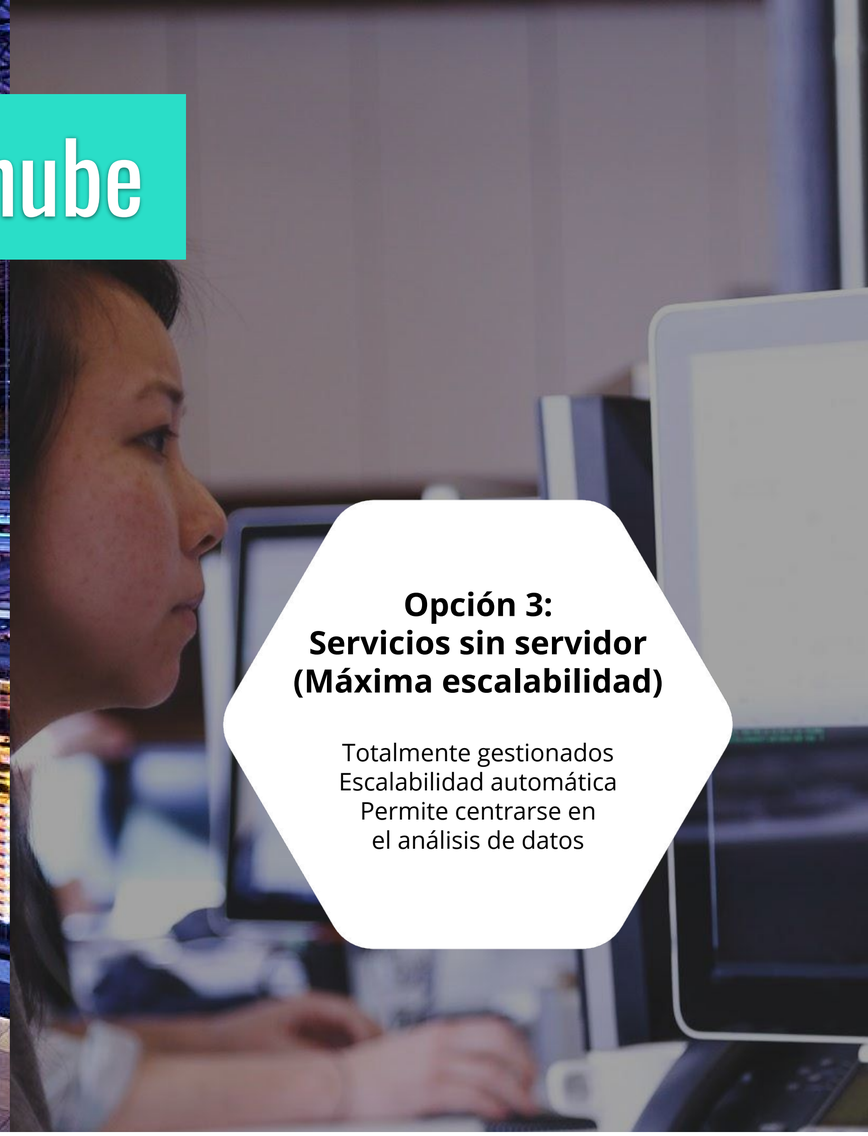
Hacia la computación en la nube

Opción 1:
Centro de Datos Físico

Servidores físicos propios
Tu propio edificio
Requiere gestión propia

Opción 2:
Centro de Datos Virtualizado
(+ escalabilidad)

Alquiler de recursos virtuales
Pagas por lo que utilizas
Requiere gestión propia



Opción 3:
Servicios sin servidor
(Máxima escalabilidad)

Totalmente gestionados
Escalabilidad automática
Permite centrarse en
el análisis de datos

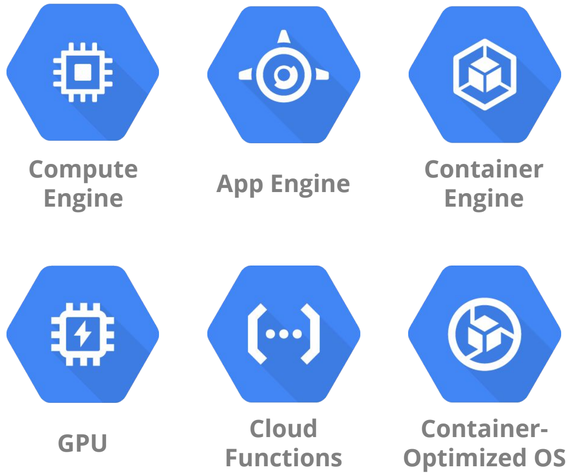
Google Cloud Platform

La nube nos permite enfocarnos en el análisis

ANÁLISIS DE DATOS

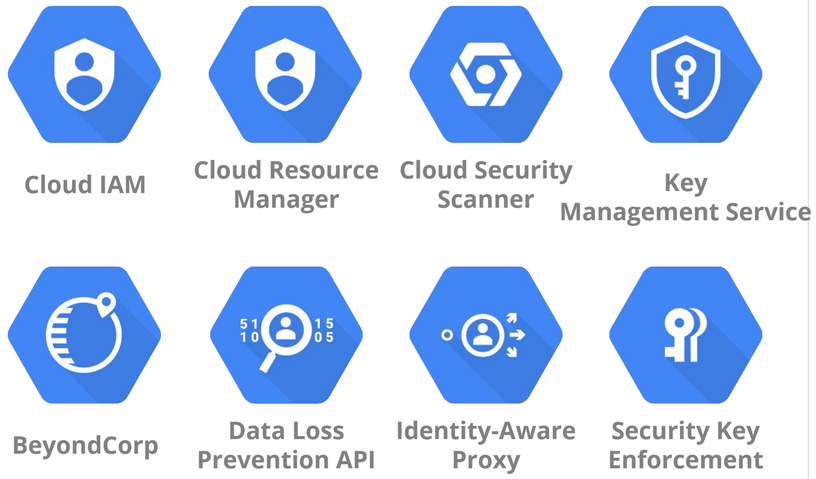
- **ASIGNACIÓN DE RECURSOS**
- **DESPLIEGUE Y CONFIGURACIÓN**
- **MONITORIZACIÓN**
- **MEJORAS DE UTILIZACIÓN**
- **MEJORAS DE RENDIMIENTO**
- **ESCALABILIDAD**

Computación



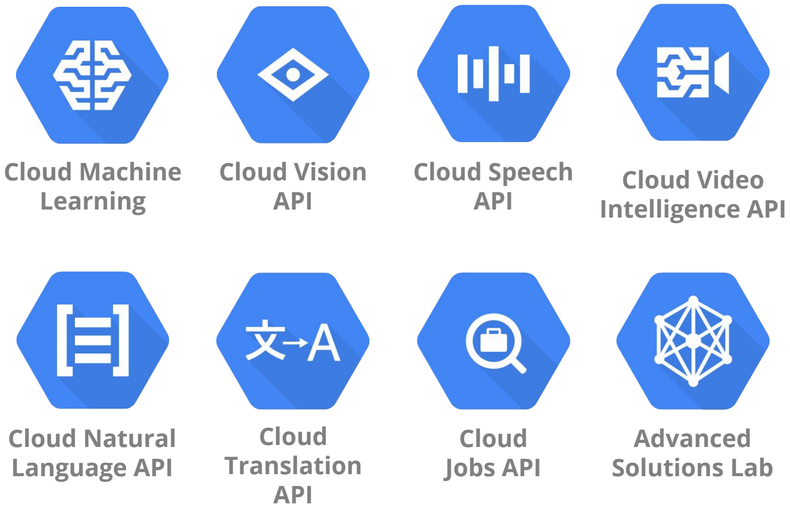
Big Data

Identidad y Seguridad

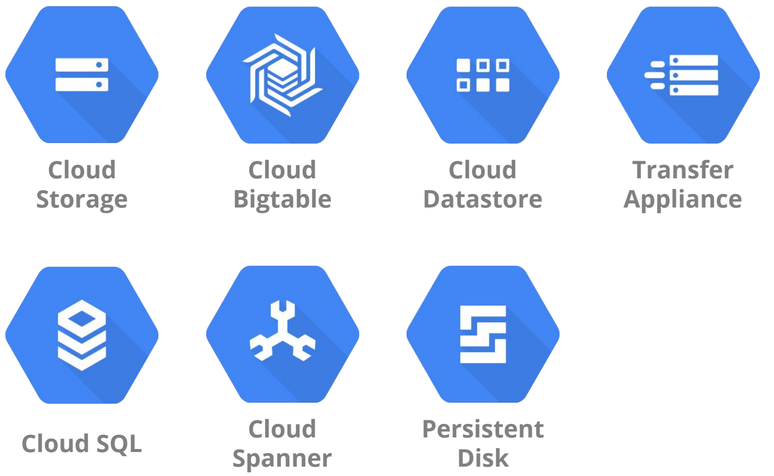


Internet de las cosas

Machine Learning



Almacenamiento



Google Cloud Platform

Computación

Compute Engine

Big Data

BigQuery

Cloud Dataflow

Cloud Dataproc

Cloud Pub/Sub

Machine Learning

Cloud Machine Learning

Cloud Vision API

Cloud Speech API

Cloud Translation API

Almacenamiento

Cloud Storage

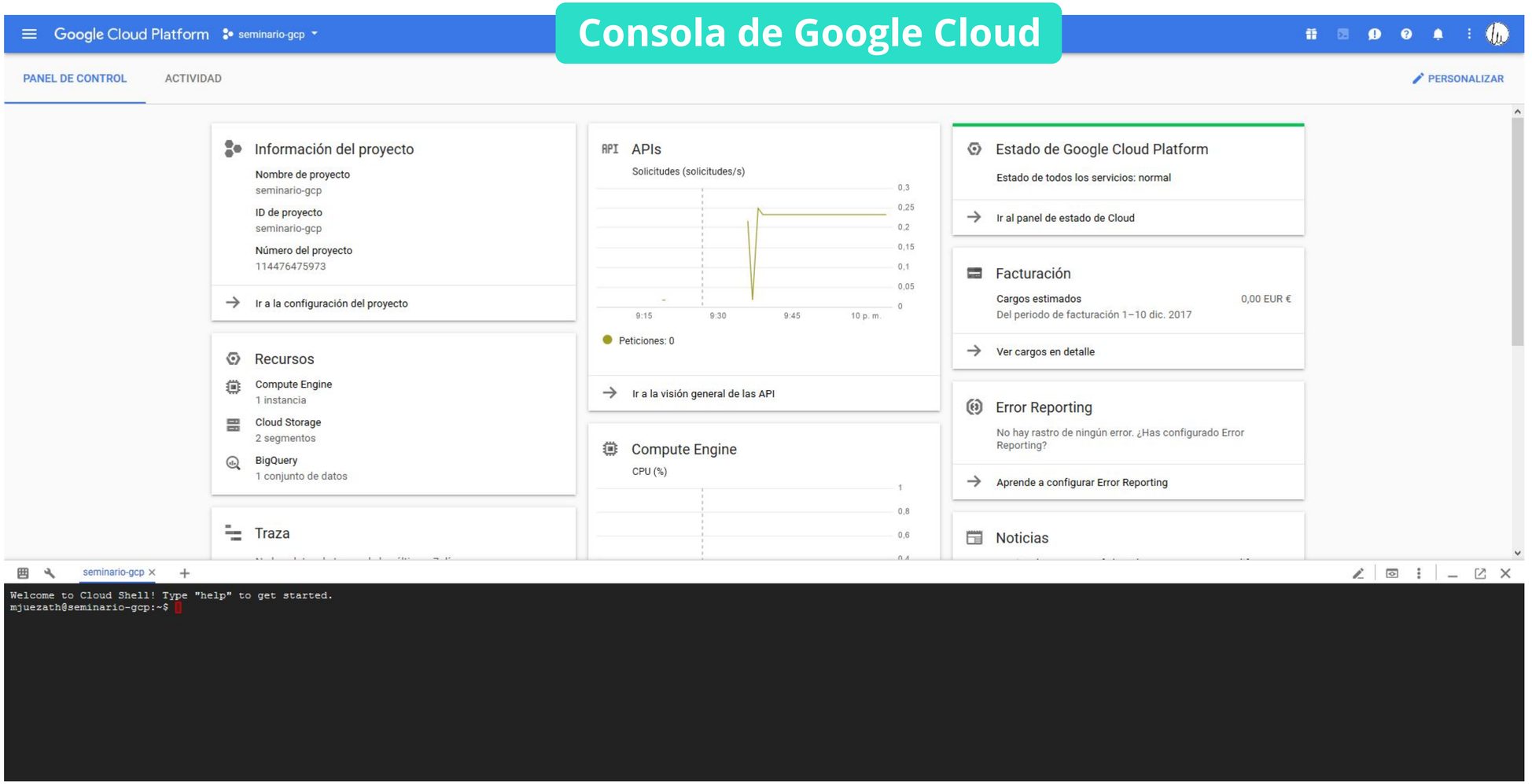
Cloud Bigtable

Cloud SQL

Google Cloud Platform

¿Dónde se gestiona todo?

Consola de Google Cloud



Google Compute Engine

Máquinas virtuales **escalables** de **alto rendimiento**

- Configuraciones predefinidas
- Configuraciones personalizadas
- Sistemas operativos Linux o Windows
- Facturación por minuto (con diversas opciones de descuentos)
- El usuario se encarga de la administración de la MV
- Posibilidad de hacer agrupaciones de MVs (clúster virtual)
- Máquinas virtuales no garantizadas (preemptive)

Google Cloud Storage

Almacenamiento centralizado

- Durabilidad y redundancia (posibilidad multiregional)
- Alta disponibilidad
- Alta escalabilidad
- Posibilidad de archivado de datos inactivos (más económico)
- Análisis de datos desde Compute Engine (alternativa a HDFS)

Bases de datos

Cloud SQL

- MySQL administrada
- Precisa de instancia GCE
- Escalabilidad de terabytes

BigTable

- NoSQL administrada (HBase)
- Precisa de instancia GCE
- Escalabilidad de petabytes
- Baja latencia

Cloud Dataproc

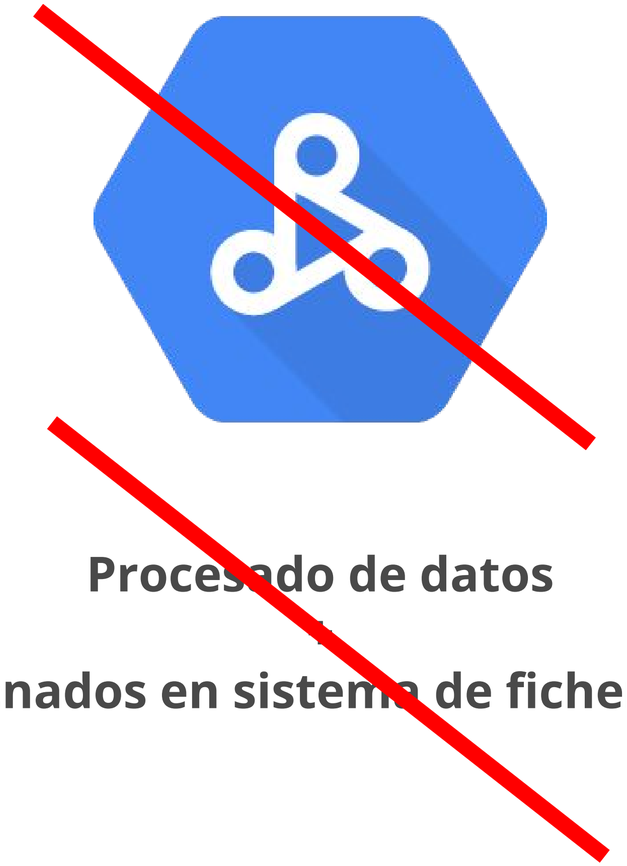
Procesamiento de datos paralelo y Machine Learning

- Hadoop, Spark, Pig, y Hive administrados.
- Clústeres de tamaño flexible
- Posibilidad de uso de máquinas virtuales no garantizadas
- Integrado con Google Cloud Storage (alternativa a HDFS)

Separación Almacenamiento-Cómputo

Procesado de datos
+
Almacenamiento de datos en sistema de ficheros local HDFS

Separación Almacenamiento-Cómputo



Procesado de datos

Datos almacenados en sistema de ficheros local HDFS

Separación Almacenamiento-Cómputo

Almacenamiento de datos (Rendimiento similar a HDFS)

Procesado de datos

Separación Almacenamiento-Cómputo

Almacenamiento de datos
(Rendimiento similar a HDFS)

Procesado de datos

Ejemplo:

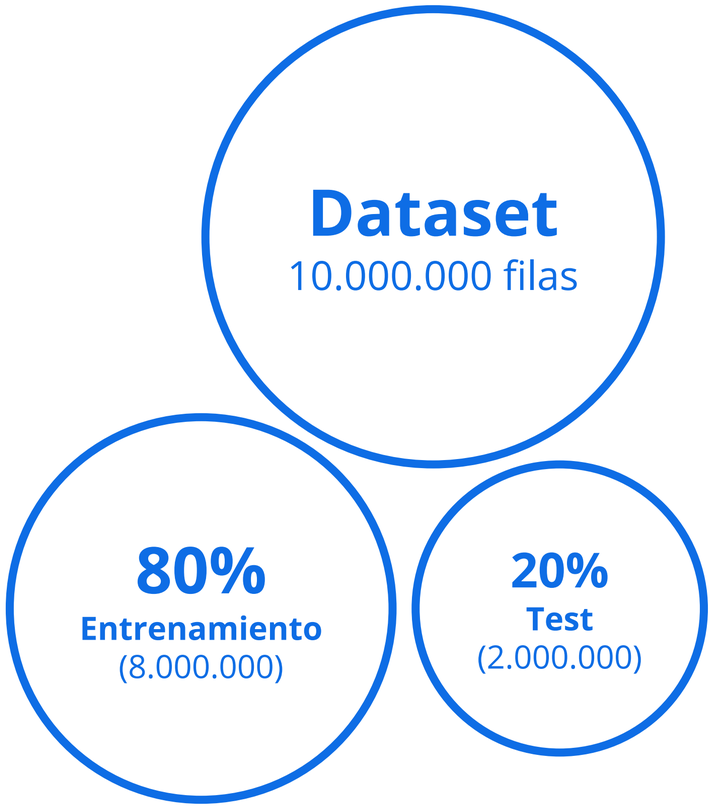
Machine Learning en Cloud Dataproc

PySpark

Regresión Lineal

Predicción del peso de un recién nacido dados:

- Edad de la madre
- Edad del padre
- Semanas de gestación
- Ganancia de peso de la madre
- Puntuación Apgar



Ejemplo:

Machine Learning en Cloud Dataproc (pasos entrenamiento)

Ejemplo:

Machine Learning en Cloud Dataproc (pasos predicción)

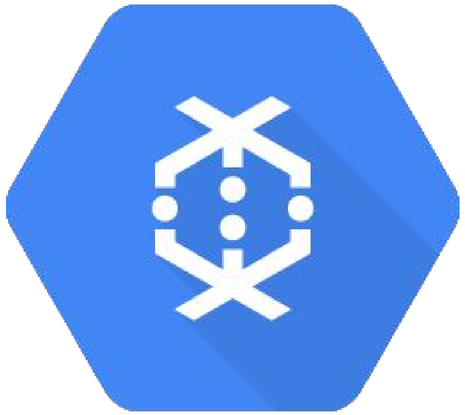
BigQuery

Almacén de grandes cantidades de datos para su análisis

- No requiere servidor
- Análisis de grandes cantidades de datos mediante SQL
- Análisis en tiempo real
- Escalabilidad de petabytes
- Integrado con Dataflow, Spark y Hadoop
- Facturación por almacenamiento y consulta

Cloud Dataflow

Servicio de procesamiento de datos



- No requiere servidor
- Escalabilidad automática
- Asignación automática de recursos
- Implementación de Apache Beam
- Integrado con GCS, Pub/Sub, Bigtable y BigQuery
- Integrado con Apache Kafka y HDFS

Cloud Pub/Sub

Servicio de mensajería asíncrona

- No requiere servidor
- Escalabilidad automática
- Entrega de mensajes garantizada
- APIs y librerías en hasta 7 lenguajes

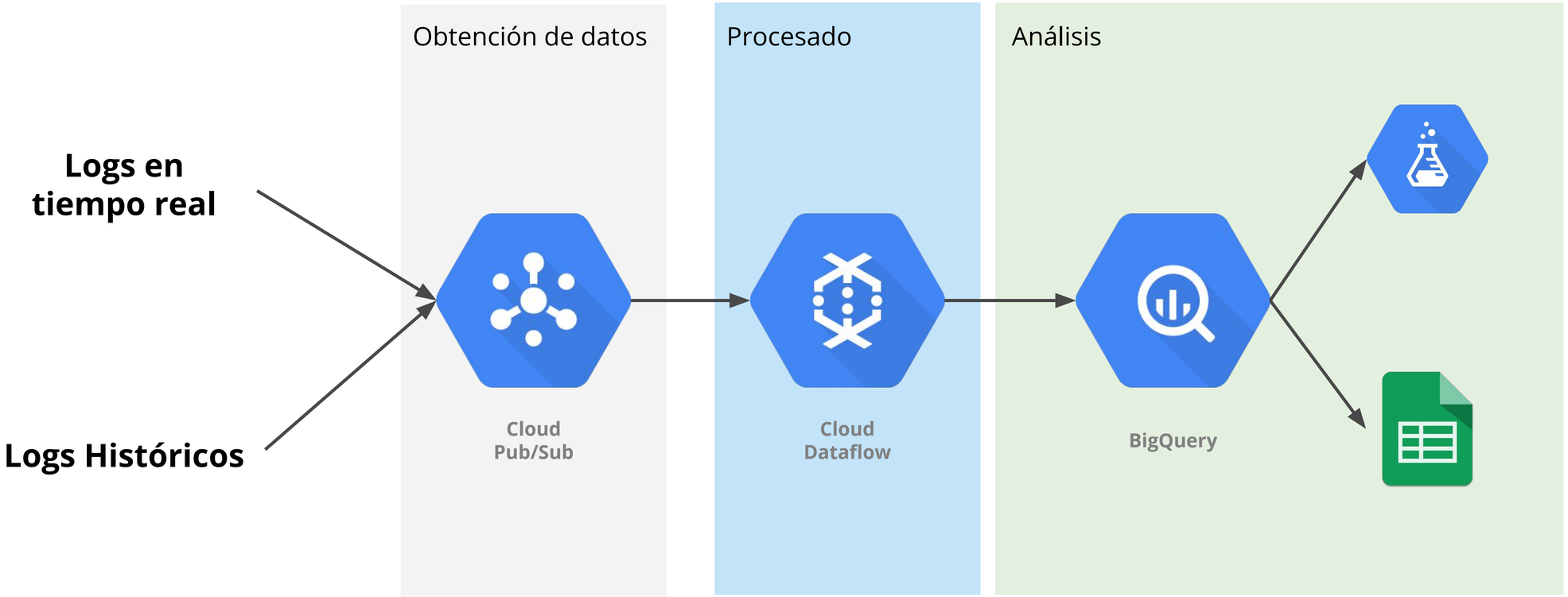
Caso de uso:

Tratamiento de datos de log (datos desestructurados)

```
64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/edit/Main/Double_bounce_sender?topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
64.242.88.10 - - [07/Mar/2004:16:06:51 -0800] "GET /twiki/bin/rdiff/TWiki/NewUserTemplate?rev1=1.3&rev2=1.2 HTTP/1.1" 200 4523
64.242.88.10 - - [07/Mar/2004:16:10:02 -0800] "GET /mailman/listinfo/hsdivision HTTP/1.1" 200 6291
64.242.88.10 - - [07/Mar/2004:16:11:58 -0800] "GET /twiki/bin/view/TWiki/WikiSyntax HTTP/1.1" 200 7352
64.242.88.10 - - [07/Mar/2004:16:20:55 -0800] "GET /twiki/bin/view/Main/DCCAndPostFix HTTP/1.1" 200 5253
64.242.88.10 - - [07/Mar/2004:16:23:12 -0800] "GET /twiki/bin/oops/TWiki/AppendixFileSystem?template=oopsmore¶m1=1.12¶m2=1.12 HTTP/1.1" 200 11382
64.242.88.10 - - [07/Mar/2004:16:24:16 -0800] "GET /twiki/bin/view/Main/PeterThoeny HTTP/1.1" 200 4924
64.242.88.10 - - [07/Mar/2004:16:29:16 -0800] "GET /twiki/bin/edit/Main/Header_checks?topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12851
64.242.88.10 - - [07/Mar/2004:16:30:29 -0800] "GET /twiki/bin/attach/Main/OfficeLocations HTTP/1.1" 401 12851
64.242.88.10 - - [07/Mar/2004:16:31:48 -0800] "GET /twiki/bin/view/TWiki/WebTopicEditTemplate HTTP/1.1" 200 3732
64.242.88.10 - - [07/Mar/2004:16:32:50 -0800] "GET /twiki/bin/view/Main/WebChanges HTTP/1.1" 200 40520
64.242.88.10 - - [07/Mar/2004:16:33:53 -0800] "GET /twiki/bin/edit/Main/Smtpd_etrn_restrictions?topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12851
64.242.88.10 - - [07/Mar/2004:16:35:19 -0800] "GET /mailman/listinfo/business HTTP/1.1" 200 6379
64.242.88.10 - - [07/Mar/2004:16:36:22 -0800] "GET /twiki/bin/rdiff/Main/WebIndex?rev1=1.2&rev2=1.1 HTTP/1.1" 200 46373
64.242.88.10 - - [07/Mar/2004:16:37:27 -0800] "GET /twiki/bin/view/TWiki/DontNotify HTTP/1.1" 200 4140
64.242.88.10 - - [07/Mar/2004:16:39:24 -0800] "GET /twiki/bin/view/Main/TokyoOffice HTTP/1.1" 200 3853
64.242.88.10 - - [07/Mar/2004:16:43:54 -0800] "GET /twiki/bin/view/Main/MikeMannix HTTP/1.1" 200 3686
64.242.88.10 - - [07/Mar/2004:16:45:56 -0800] "GET /twiki/bin/attach/Main/PostfixCommands HTTP/1.1" 401 12846
64.242.88.10 - - [07/Mar/2004:16:47:12 -0800] "GET /robots.txt HTTP/1.1" 200 68
64.242.88.10 - - [07/Mar/2004:16:47:46 -0800] "GET /twiki/bin/rdiff/Know/ReadmeFirst?rev1=1.5&rev2=1.4 HTTP/1.1" 200 5724
64.242.88.10 - - [07/Mar/2004:16:49:04 -0800] "GET /twiki/bin/view/Main/TWikiGroups?rev=1.2 HTTP/1.1" 200 5162
```

Caso de uso:

Tratamiento de datos de log (datos desestructurados)



Cloud ML

Servicio de Machine Learning

- No requiere servidor
- Motor TensorFlow
- Escalabilidad automática
- Despliegue, almacenamiento, y monitorización de modelos

Ejemplo:

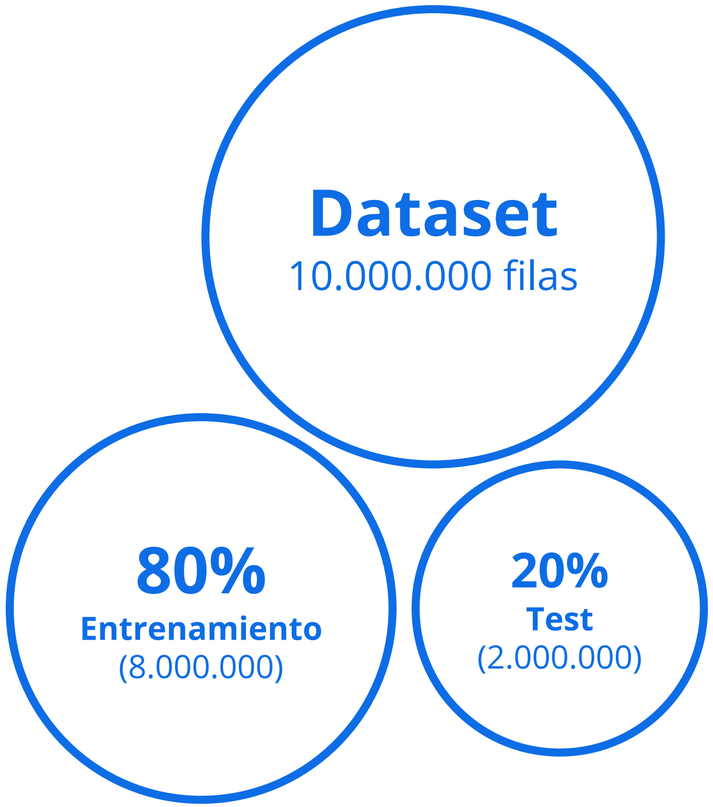
Machine Learning en Cloud ML

TensorFlow

Regresión Lineal

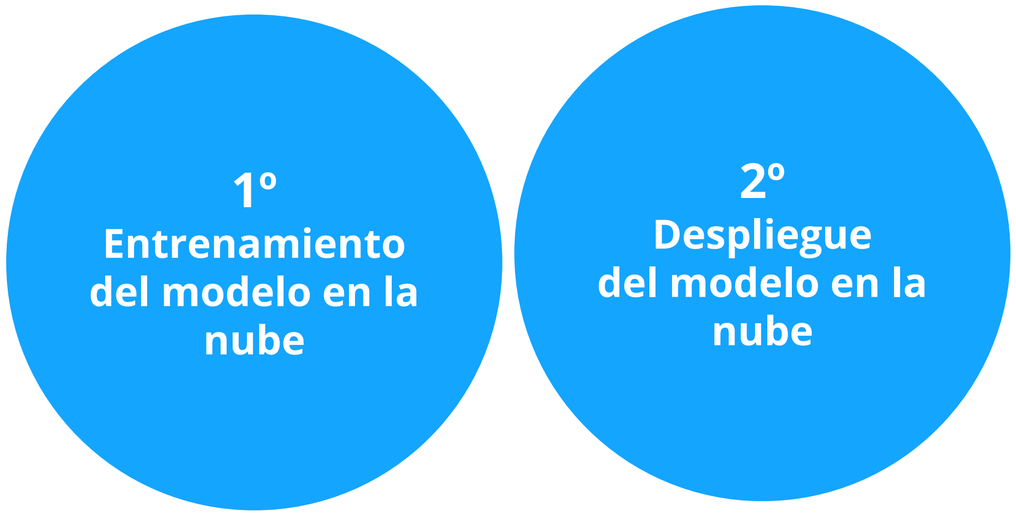
Predicción del peso de un recién nacido dados:

- Edad de la madre
- Edad del padre
- Semanas de gestación
- Ganancia de peso de la madre
- Puntuación Apgar



Ejemplo:

Machine Learning en Cloud ML (pasos entrenamiento)



Ejemplo:

Machine Learning en Cloud ML (pasos predicción)

Modelos ML ya entrenados

Cloud Vision API

- Reconocimiento de objetos
- Extracción de texto
- Detección de atributos de la imagen

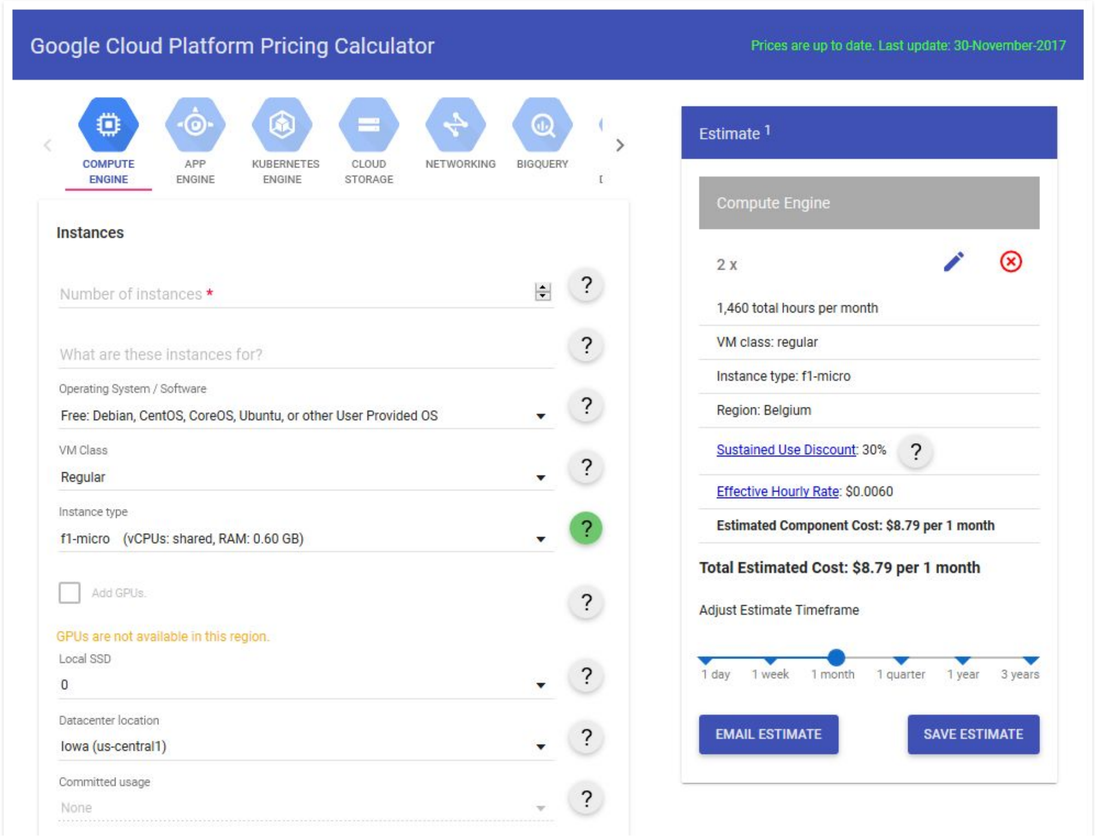
Cloud Speech API

- Conversión voz-texto
- Más de 80 idiomas
- Resultados en tiempo real

Cloud Translation API

- Traducción de textos
- Más de 100 idiomas
- Detección de idioma

Todo esto está muy bien pero...
¿Cuánto me va a costar?



Sí, existen alternativas a Google

Google Cloud	Amazon Web Services	Microsoft Azure
Compute Engine	Elastic Compute Cloud	Virtual Machines
Cloud Storage	Simple Storage Service	Azure Blob Storage
Cloud SQL	Relational Database Service	SQL Database
Cloud Bigtable	DynamoDB	Table Storage
Cloud Dataproc	Elastic MapReduce	HDInsight, Batch
BigQuery	RedShift	Data Lake Analytics, Data Lake Store
Cloud Dataflow	Kinesis	Stream Analytics
Cloud Pub/Sub	Kinesis	Event Hubs / Service Bus
Cloud ML	Amazon Machine Learning	Machine Learning Studio

https://cloud.google.com/free/docs/map-aws-google-cloud-platform

https://cloud.google.com/free/docs/map-azure-google-cloud-platform

¡Gracias!

¿Preguntas?

Puedes encontrar esta presentación y sus ejemplos en:
https://github.com/mjuez/seminario-gcp