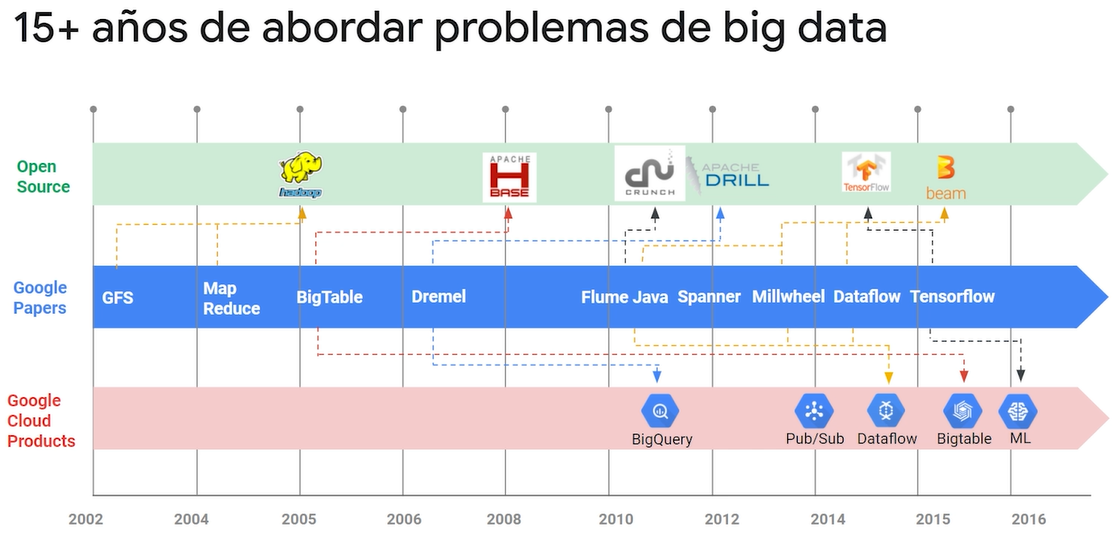
**Introducción a big data**

Google en la industria big data: Open Source, Google Papers y Google Cloud Products

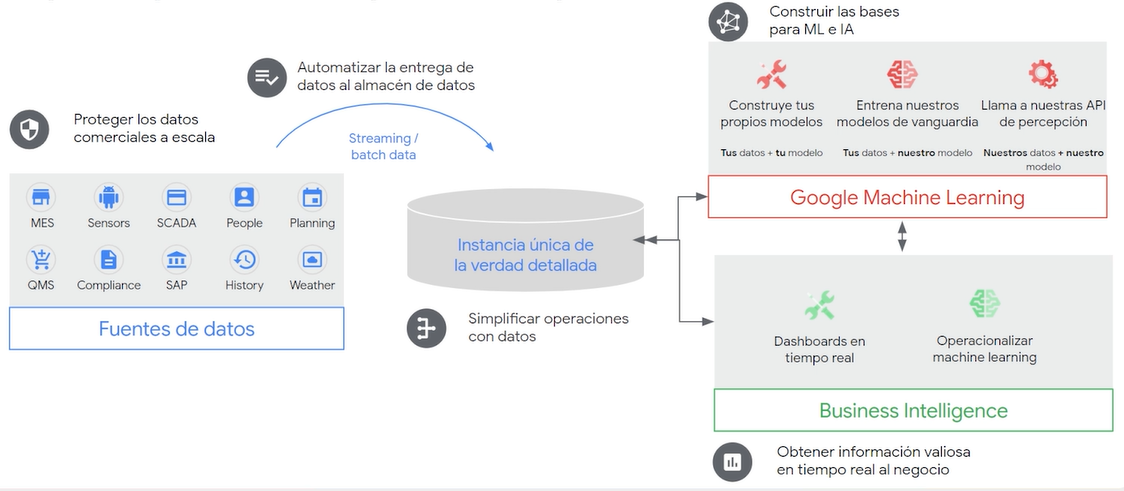


**Complejidad de datalakes y analytics.**

Los datos no estructurados representan el 90% de los datos empresariales

* Aplicaciones Legacy
* Silos de datos en todas partes.
* Cambio de vista sobre el valor de los datos.
* Entorno regulatorio.
* Habilidades limitadas, difícil de reclutar.

**Listo para IA: hiper personalizado, hiper optimizado y en tiempo real**

****

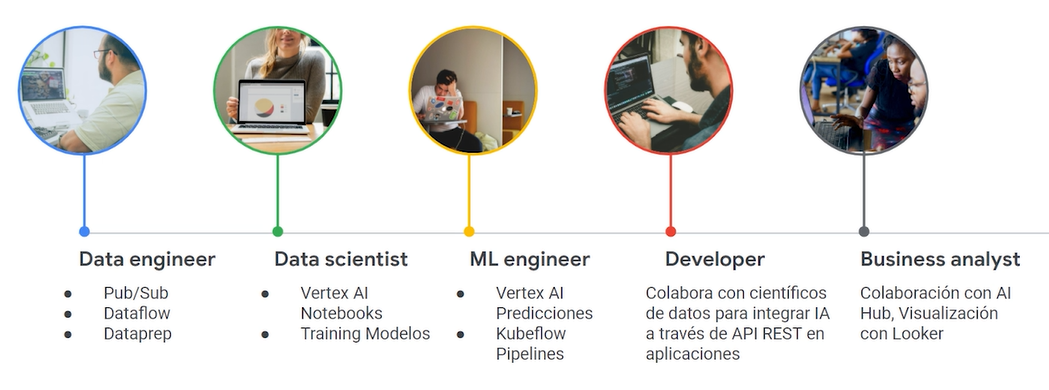
**Base completa para el ciclo de vida de datos y análisis.**

1. Ingesta de datos a cualquier escala.
2. Pipeline de datos de transmisión confiable.
3. Data warehousing y data lake
4. Analítica avanzada.

**¿Como conectar a usuario para crear mas impacto?**

* ML Engineer
* Data Engineer
* Data Scientist
* Developer
* Business analyst
* End User

Servicio que puede usar cada perfil



UNIR ROLES Y SERVICIOS DE UNA ORGANIZACIÓN. ------HERRAMIENTAS DE GOBIOERNO PARA DEMOCRATIZAR CONOCIMIENTO

* 1. Decubrir , clasificar y redactar datos sensibles (GOOGLE DATA LOSS PREVENTION
  2. Gestión de metadata ( DATA CATALOG
  3. Control de acceso (CLOUD IAM)

CAPACIDADES DE DATA LOSS PREVENTION

. Inspección ( encontrar datos sensibles en tus conjuntos de datos)

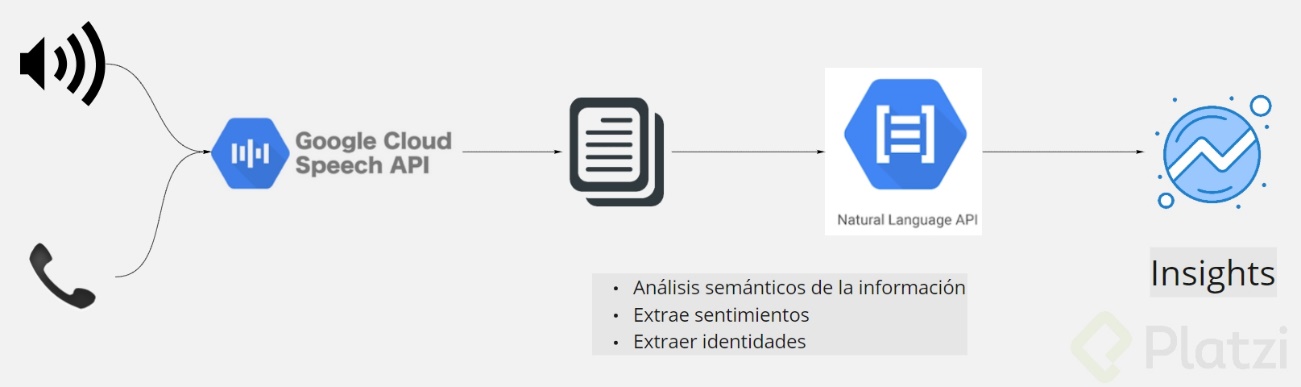
. Desidentificación

. Reidentificaión.

. Análisis del riesgo de reidentificación ( k-anonymity)

**EXPLORANDO G.C.P. PARA BIG DATA**

* Big Query:  
  Es un servicio serverless totalmente escalable que permite trabajar con petabytes de información y lo más importante permite separar el cómputo del procesamiento.
* Pub-Sub:  
  Permite trabajar a través de tópicos, esos tópicos permiten conectarnos con distintas fuentes de información y obtener éstos elementos en tiempo real. Éstos elementos pueden llevarse hasta un DataFlow para poder hacer filtrado, poder hacer ordenamiento de éstos registros y luego cargarlos en Big Query.
* DataFlow:  
  Es un servicio de extracción, transformación y carga de datos que puede funcionar tanto Batch como Near Real Time. Éste servicio permite trabajar en base a jobs y éstos pueden ser a través de un Template y ya vienen pre configurados distintos Templates.
* Data Catalog:  
  Es un servicio totalmente escalable que permite administrar la metadata que tu tienes en tu organización. Ésta metadata recuerda que se genera desde los procesos de ingesta hasta los repositorios de datos, por lo tanto es bastante útil que puedas conectar tus sistemas y poder almacenar ésta data para que distintos usuarios dentro de tu organización puedan conocer cual es la información que tienes cargado en Google Cloud.
* Vertex AI:  
  Almacena todas las funcionalidades para poder hacer y crear Datasets, hacer ingesta, hacer entrenamiento, deployement de éstos modelos de una forma asistida con AutoML Tables o de una forma programática con Jupyter Notebook. Éstas dos opciones están disponibles en Vertex AI.
* Vision:  
  Nos permite resolver una serie de casos de uso enfocados en la detección tanto de video como de imégenes para detectar elementos.
* Document AI:  
  Permite tomar información no estructurada desde distintos formularios, PDF, correo electrónico, y poder extraer ésta información para poder procesarla de una forma ordenada y encontrar insight de éstos textos.
* Speech to Text:  
  Permite resolver casos de uso donde tengo audio, y ese audio necesito hacer una transcripción a texto para después analizarlo con los servicios de lenguaje natural.
* Natural Lenguage:  
  Permite analizar, extraer sentimientos, extraer entidades y todo eso gracias al análisis semántico de información de texto.





**Que es un Cloud Dtaflow**  
Es un servicio de procesamiento de datos totalmente administrado que simplifica el desarrollo y la administración de flujos y pipelines.  
-Acelera el desarrollo de streaming y bach  
-Gestion y operaciones simplificadas  
-Construir sobre una base para machine learning

**Qué es BigQuery**  
Data warehouse empresarial totalmente administrado y serverless.  
-Escala de petabytes y responde muy rápido  
-Disponible en multiples regiones  
-Separación de almacenamiento y cómputo.

**Componentes importantes:**  
-Es SQL  
-Rest API  
-Web UI, CLI  
-LibrerÍa para clientes virtual

**Como funciona?**  
A traves de un esquema federado de lectura llevamos nuestros archivos ( json, video, PDF, etc.) hacia BigQuery  
-Procesamiento y transformacion  
-Procesamiento: OCR, voz a texto, clúster, etc.

**Google Cloud Pub/Sub**  
Mensajería impulsada por eventos para la ingesta de datos y el movimiento de los mismos.

* Ingestion y entrega de eventos escalables/ persistentes.
* Patrón de publicación.
* Servicio global.
* Integrado con otras herramientas.

QUERY DATAFLOW SQL EDITOR

SELECT

TUMBLE\_START('INTEVARL 1 MINUTE') as period\_start,

SUM(passenger\_count) AS pickup\_count

FROM pubsub.topic.`pubsub-public-data`.`taxirides-realtime`

WHERE

ride\_status = "pickup"

GROUP BY

TUMBLE(event\_timestamp, 'INTERVAL 1 MINUTE')

**Product manager**  
-Bi interactivo  
-EDW en Spreadsheet  
-Consultas en lenguaje natural

**Data analyst**  
-Data warehouse interminable  
-Data pipelines autogestionados  
-Modelo de datos catálogo  
-Machine Learning en SQL

**Data engineer**  
-Infraestructura autónoma  
-Variedad de herramientas/lenguajes  
-Calidad/linaje de datos  
-Capacidades en tiempo real

**Data scientist**  
-Notebooks portables  
-Evaluación y selección de modelos  
-Desarrollo point anda click  
-Colaboración

**ML developer**  
-Imagenes y videos  
-Análisis de sentimiento extracción de entidades  
-Chatbots, comandos de voz  
-Enrutamiento de flotas previsión

**ML engineer**  
-Alojamiento de modelos escalables  
-ML CI/CD y orquestación  
-Procedencia y linaje  
-Mejoras y reentrenamiento

**Vertex Ai**  
-Construye en lo mejor  
-Acelera tiempo a valor  
-Confianza y responsabilidad  
\_Obtenemos: \_  
-Velocidad en los modelos de producción  
-Los mejores algoritmos como servicio de Google  
-Flujo de trabajo de ML  
-Gestión y gobierno de modelos

**Product manager**  
-Bi interactivo  
-EDW en Spreadsheet  
-Consultas en lenguaje natural

**Data analyst**  
-Data warehouse interminable  
-Data pipelines autogestionados  
-Modelo de datos catálogo  
-Machine Learning en SQL

**Data engineer**  
-Infraestructura autónoma  
-Variedad de herramientas/lenguajes  
-Calidad/linaje de datos  
-Capacidades en tiempo real

**Data scientist**  
-Notebooks portables  
-Evaluación y selección de modelos  
-Desarrollo point anda click  
-Colaboración

**ML developer**  
-Imagenes y videos  
-Análisis de sentimiento extracción de entidades  
-Chatbots, comandos de voz  
-Enrutamiento de flotas previsión

**ML engineer**  
-Alojamiento de modelos escalables  
-ML CI/CD y orquestación  
-Procedencia y linaje  
-Mejoras y reentrenamiento

**Vertex Ai**  
-Construye en lo mejor  
-Acelera tiempo a valor  
-Confianza y responsabilidad  
\_Obtenemos: \_  
-Velocidad en los modelos de producción  
-Los mejores algoritmos como servicio de Google  
-Flujo de trabajo de ML  
-Gestión y gobierno de modelos

**Aplicaciones de IA fáciles de desplegar**  
+Modelos y servicios de la mas alta calidad  
+Fácil personalización e implementación  
+Valor comercial comprobado  
**Aplicaciones de IA fáciles de desplegar**  
-Vison  
-Lenguaje  
-Conversación  
-Datos estructurados

**Vertex IA te permite manejar todos esos casos de uso anteriores**