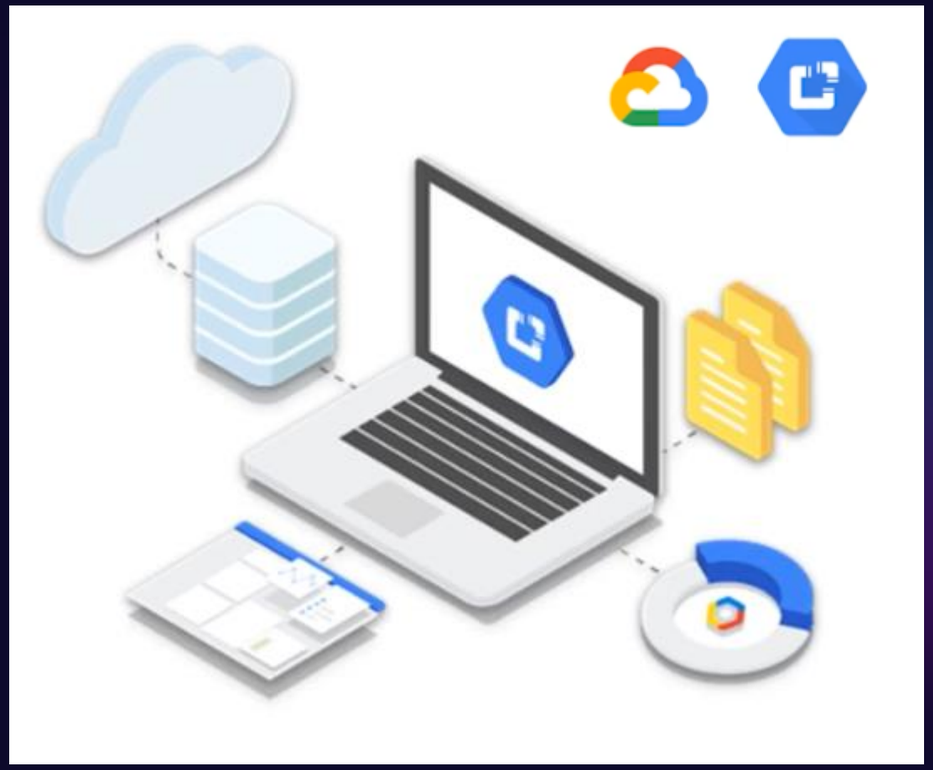
Ecosistema de integración de datos de Google

…conjunto de servicios, herramientas y plataformas que Google Cloud ofrece para facilitar la integración, transformación, gestión y análisis de datos.



Estudiantes:

Andrei Ricardo Saldarriaga
Juan Pablo Mogollon Lozano
John Isidro Roa Reina
Ariel Cifuentes Osorio

18 de mayo de 2024

...algunas soluciones agrupadas por categorías:

Ingesta de Datos

BigQuery Data Transfer

Automatiza la carga de datos desde aplicaciones SaaS y otras fuentes a BigQuery.

Cloud Pub/Sub

Servicio de mensajería en tiempo real que facilita la ingesta y transmisión de datos entre aplicaciones y servicios.

Google Cloud Storage Transfer Service

Permite transferir datos desde otras nubes o sistemas de almacenamiento a Google Cloud Storage.

Almacenamiento de Datos

Google Cloud Storage

Almacenamiento de **objetos escalable y seguro para** datos estructurados y no estructurados.

BigQuery

Almacén de datos totalmente gestionado para análisis de datos a gran escala.

Google Cloud SQL

Base de datos relacional gestionada compatible con MySQL, PostgreSQL y SQL Server.

Google Cloud Spanner:

Base de datos relacional distribuida y gestionada con escalabilidad horizontal.

Google Firestore

Base de datos NoSQL gestionada para desarrollo de aplicaciones web y móviles.

Procesamiento y Transformación de Datos

Cloud Data Fusion

Plataforma de integración de datos como servicio (iPaaS) para diseñar y gestionar pipelines de datos.

Cloud Dataflow

Servicio gestionado para procesamiento de datos en flujo y por lotes basado en Apache Beam.

Cloud Dataproc

Servicio gestionado de Hadoop y Spark para procesamiento de big data.

Google Cloud Functions

Funciones sin servidor que permiten ejecutar código en respuesta a eventos.

Google Cloud Composer

Servicio gestionado de Apache Airflow para la orquestación y gestión de flujos de trabajo de datos.

...algunas soluciones agrupadas por categorías:

Análisis de Datos / Gestión y Descubrimiento de Datos

Google Data Catalog

Herramienta para la gestión, descubrimiento y organización de datos con capacidades de etiquetado y búsqueda.

Google Cloud IAM

Servicio para gestionar el acceso y las políticas de seguridad de los recursos en Google Cloud.

Análisis de Datos BigQuery

Permite análisis SQL de grandes volúmenes de datos de manera rápida y eficiente.
Looker:

Análisis de Datos

Looker

Plataforma de BI (Business Intelligence) que permite crear informes y dashboards interactivos.

Google Data Studio

Herramienta de visualización de datos que permite crear informes y dashboards interactivos.

Machine Learning y IA

AI Platform (Vertex AI):

Plataforma para desarrollo, despliegue y gestión de modelos de aprendizaje automático.

AutoML

Herramienta que permite a los desarrolladores crear modelos personalizados de aprendizaje automático sin necesidad de expertise en machine learning.

TensorFlow

Biblioteca de código abierto para el desarrollo y entrenamiento de modelos de machine learning.

Orquestación de Flujos de Trabajo

Google Cloud Composer

Servicio gestionado de Apache Airflow para la creación, programación y monitoreo de flujos de trabajo complejos.

Integración de Aplicaciones

Google Cloud Apigee

Plataforma de gestión de API que facilita la creación, gestión y monitoreo de APIs.

Herramientas de Desarrollador

Google Cloud SDK:

Conjunto de herramientas de línea de comandos para interactuar con los servicios de Google Cloud.

Seguridad y Compliance

Google Cloud Security Command Cente

Plataforma centralizada para la gestión de la seguridad y la protección contra amenazas.

Google Storage

Es un servicio de almacenamiento en la nube proporcionado por Google. Permite a los usuarios almacenar y acceder a sus datos de manera segura y escalable desde cualquier lugar.

Ventajas

Almacenamiento escalable y confiable

Acceso desde cualquier dispositivo

Integración con otros servicios de Google

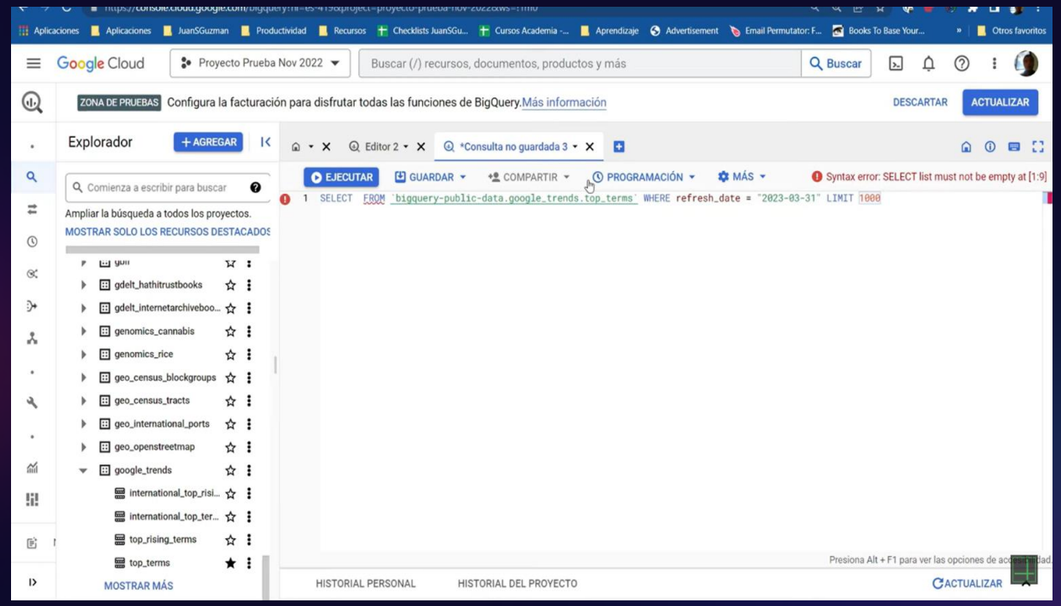
BigQuery

Almacenamiento y Análisis Escalable

BigQuery es un almacén de datos de alto rendimiento que permite procesar y analizar grandes cantidades de datos de manera rápida y eficiente.

Análisis de Datos sin Servidor

Con BigQuery, los usuarios pueden ejecutar consultas SQL sin tener que administrar la infraestructura subyacente, lo que simplifica el proceso de análisis de datos.



Integración con Otros Servicios

BigQuery se integra fácilmente con otras herramientas y servicios de Google Cloud, permitiendo una experiencia fluida de análisis de datos.

Dataflow

Dataflow es un servicio de procesamiento de datos en tiempo real y por lotes que forma parte del ecosistema de integración de datos de Google Cloud. Permite a los desarrolladores crear y ejecutar pipelines de datos escalables y eficientes para procesar grandes volúmenes de información de manera rápida y confiable. Dataflow se encarga de la paralelización, escalado y gestión de los recursos necesarios para el procesamiento de los datos, lo que permite a los equipos centrarse en la lógica del negocio sin preocuparse por la infraestructura subyacente.

1 Procesamiento en tiempo real y por lotes

Permite procesar datos tanto en tiempo real como en lotes, lo que brinda flexibilidad para diferentes tipos de análisis.-

2 Basado en Apache Beam

Utiliza el modelo de programación unificado de Apache Beam, lo que facilita el desarrollo de pipelines de datos consistentes y portables.-

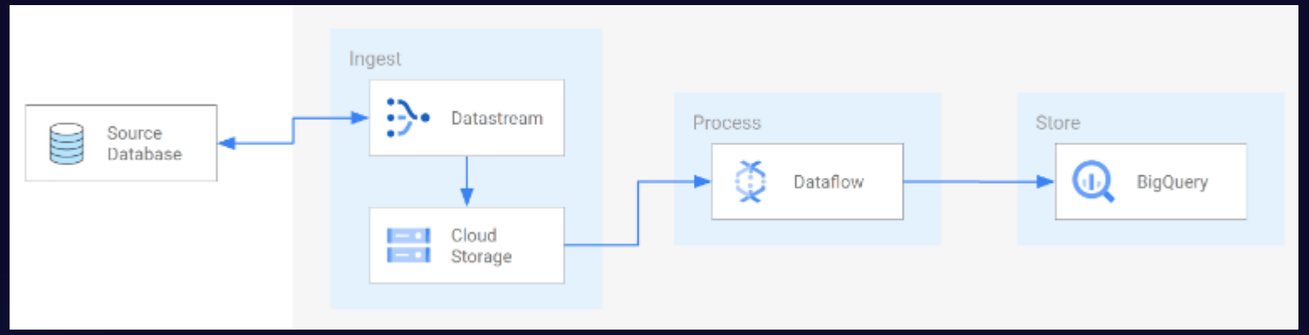
3 Escalabilidad automática

Se ajusta automáticamente al volumen de datos y a la complejidad de la carga de trabajo, lo que garantiza un procesamiento eficiente sin necesidad de configuraciones manuales.

4 Integración con otros servicios de Google Cloud:

Se integra fácilmente con servicios como BigQuery, Pub/Sub y Cloud Storage, lo que simplifica la implementación de pipelines complejos.

Dataflow



Ventajas

Facilidad de uso

Permite diseñar pipelines de datos de forma visual y con un enfoque basado en código, lo que facilita su adopción por parte de equipos no especializados en programación.-

Tolerancia a fallos

Proporciona mecanismos integrados para manejar fallos y garantizar la integridad de los datos durante el procesamiento.

Desventajas

Costos adicionales

Los costos de Google Cloud Dataflow se basan en el uso de recursos de computación y almacenamiento, así como en el tráfico de datos procesados.

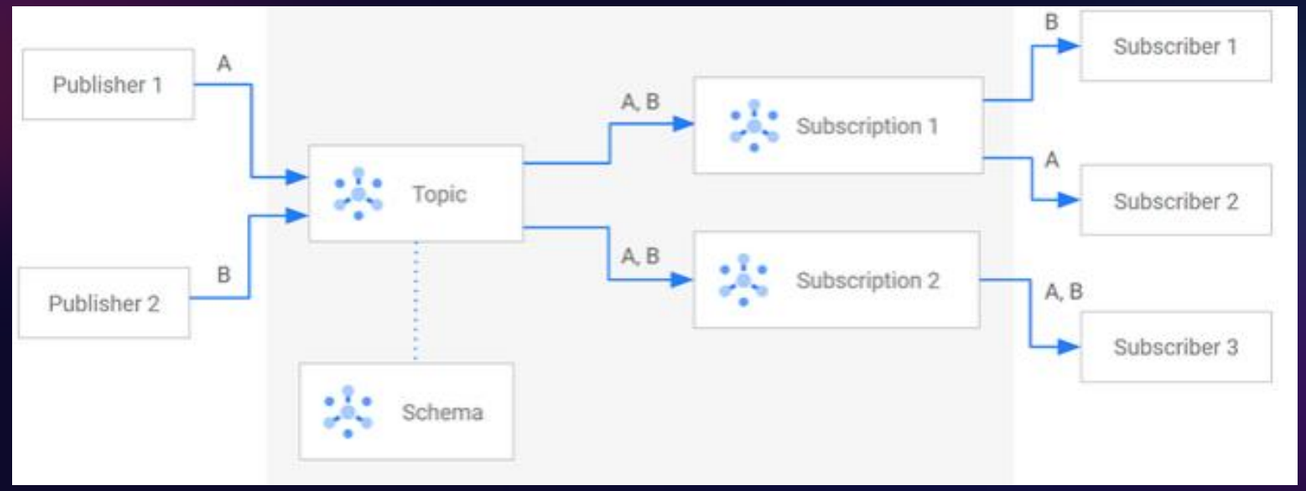
Casos de uso

Análisis de datos en tiempo real para aplicaciones web y móviles.- Procesamiento de logs y eventos para monitoreo y análisis de sistemas.- ETL (Extract, Transform, Load) de datos para integrar datos de diferentes fuentes y formatos.-

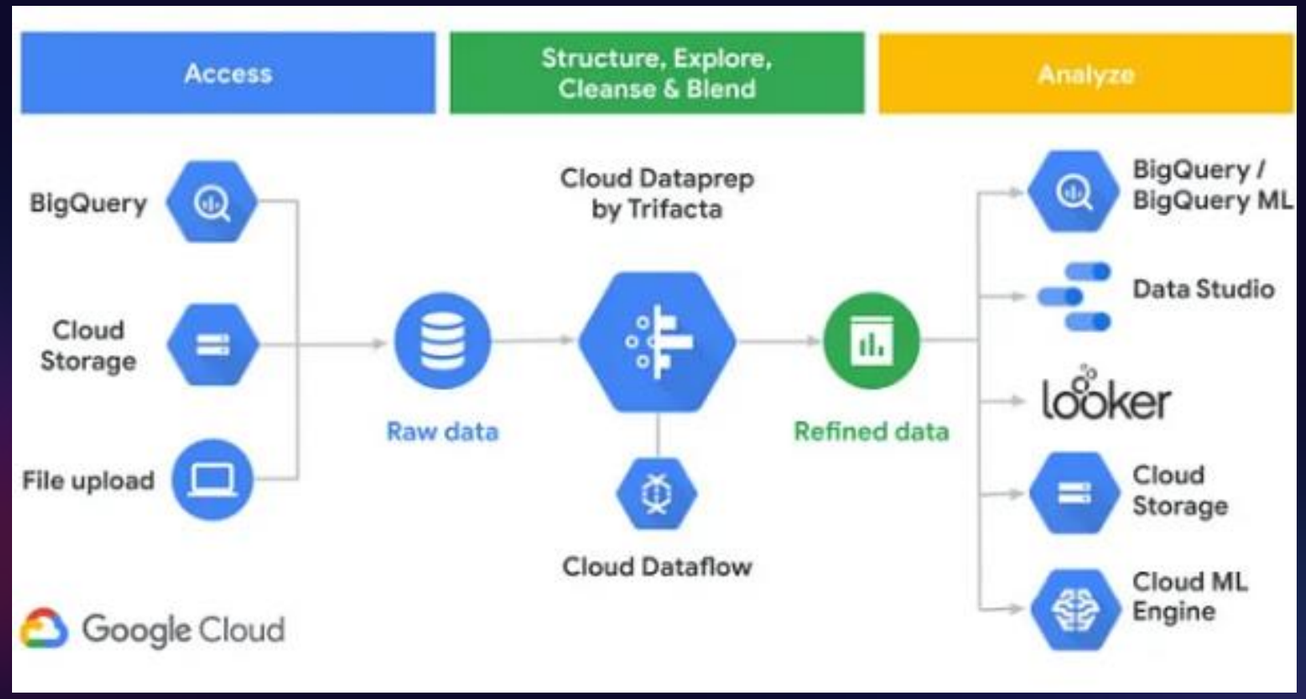
Pub/Sub

Pub/Sub es un servicio de mensajería en tiempo real de Google Cloud que permite una comunicación asíncrona y escalable entre diferentes componentes de una aplicación. Permite a los productores enviar mensajes a uno o más suscriptores sin necesidad de que estos estén conectados simultáneamente.

Pub/Sub facilita la integración de diferentes servicios de Google Cloud, como Dataflow, BigQuery y Dataprep, permitiendo el flujo de datos entre ellos de manera eficiente y en tiempo real.



Dataprep



Dataprep es una herramienta de Google Cloud que facilita la preparación y limpieza de datos de manera intuitiva y visual. Permite a los usuarios analizar, transformar y enriquecer datos de diversos orígenes sin necesidad de conocimientos avanzados de programación.

Con Dataprep, los equipos pueden agilizar el proceso de preparación de datos, lo que los habilita a dedicar más tiempo a generar insights valiosos a partir de esos datos. La herramienta ofrece funcionalidades avanzadas de detección de anomalías, perfilado de datos y generación de flujos de trabajo para automatizar tareas repetitivas.

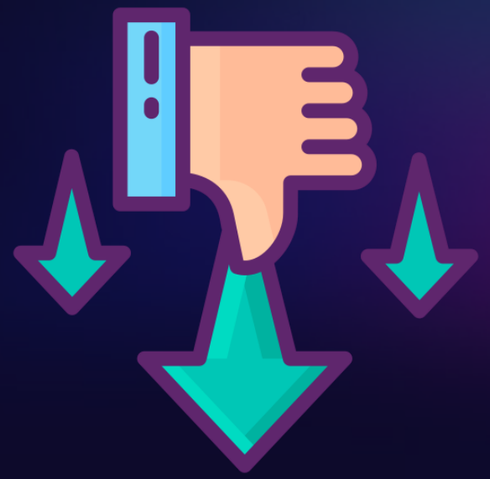
Dataprep

Ventajas

1. Interfaz Gráfica Intuitiva
2. Automatización de Tareas Repetitivas
3. Integración con Otros Servicios de Google Cloud
4. Escalabilidad y Rendimiento
5. Seguridad y Cumplimiento

Desventajas

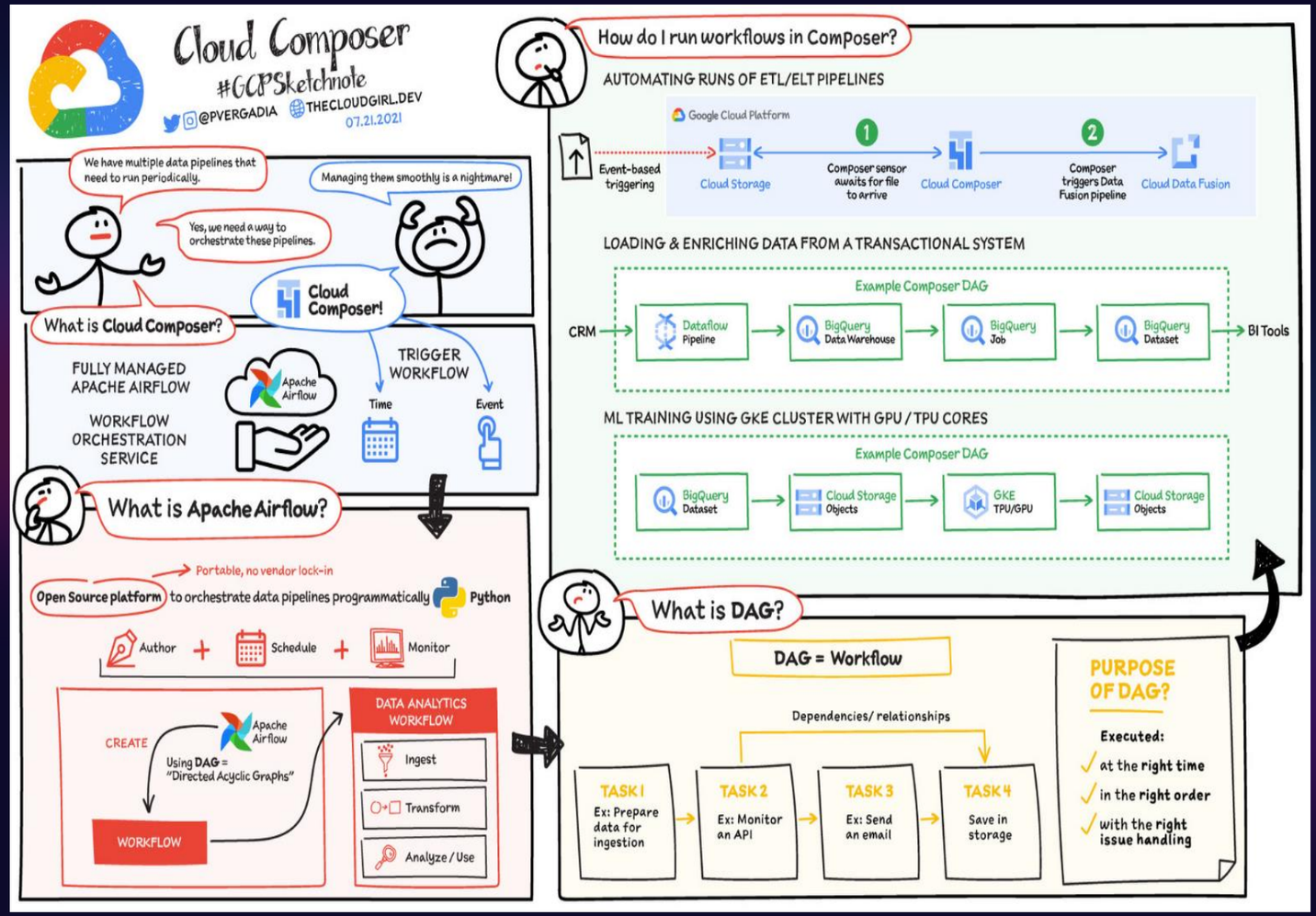
1. Costo
2. Dependencia de Conectividad en la Red
3. Curva de Aprendizaje
4. Compatibilidad con Otros Sistemas



Cloud Composer

Cloud Composer es un servicio de organización de flujos de trabajo completamente administrado que permite crear, programar, supervisar y administrar canalizaciones de flujos de trabajo que se distribuyen en nubes y centros de datos locales.

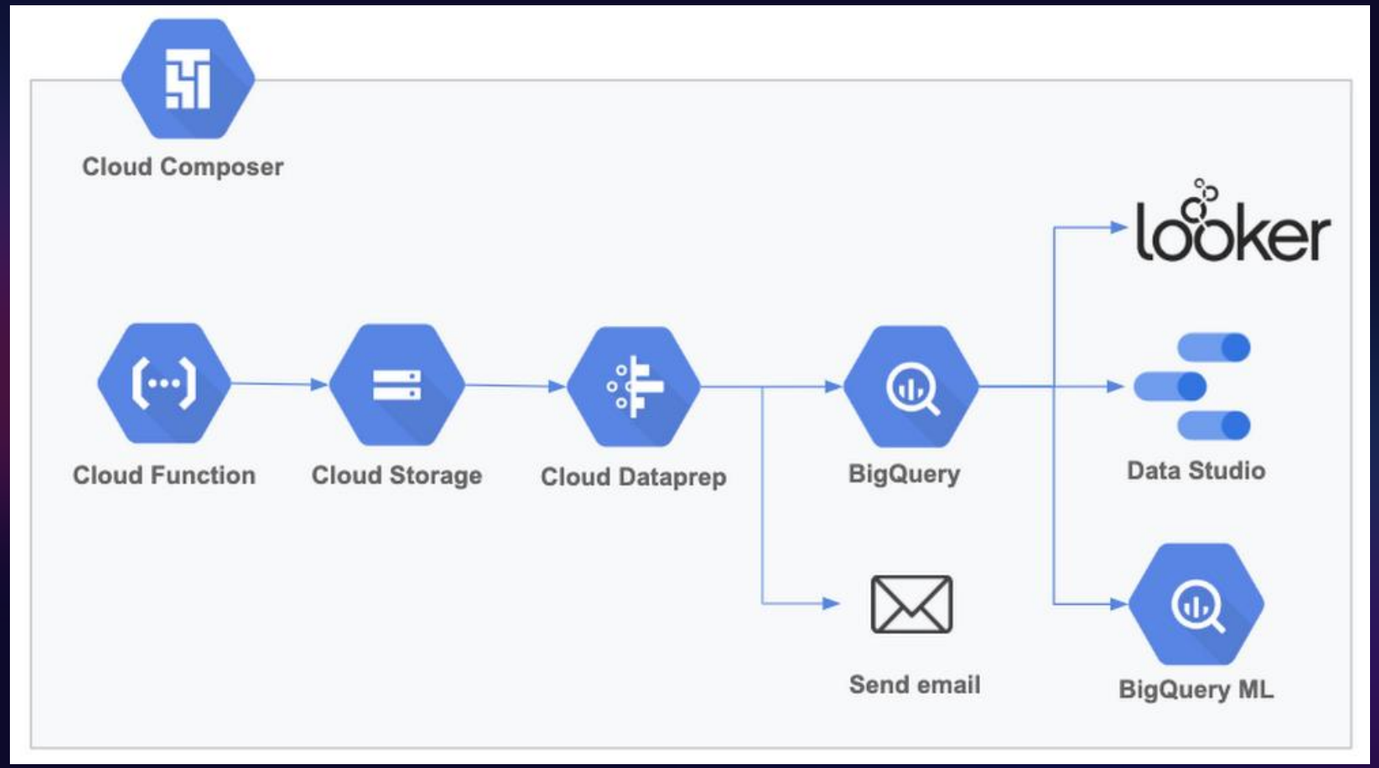
Cloud Composer se basa en el popular proyecto de código abierto Apache Airflow y opera con el lenguaje de programación Python.



Composer

Planificación y Orquestación de Flujos de Trabajo

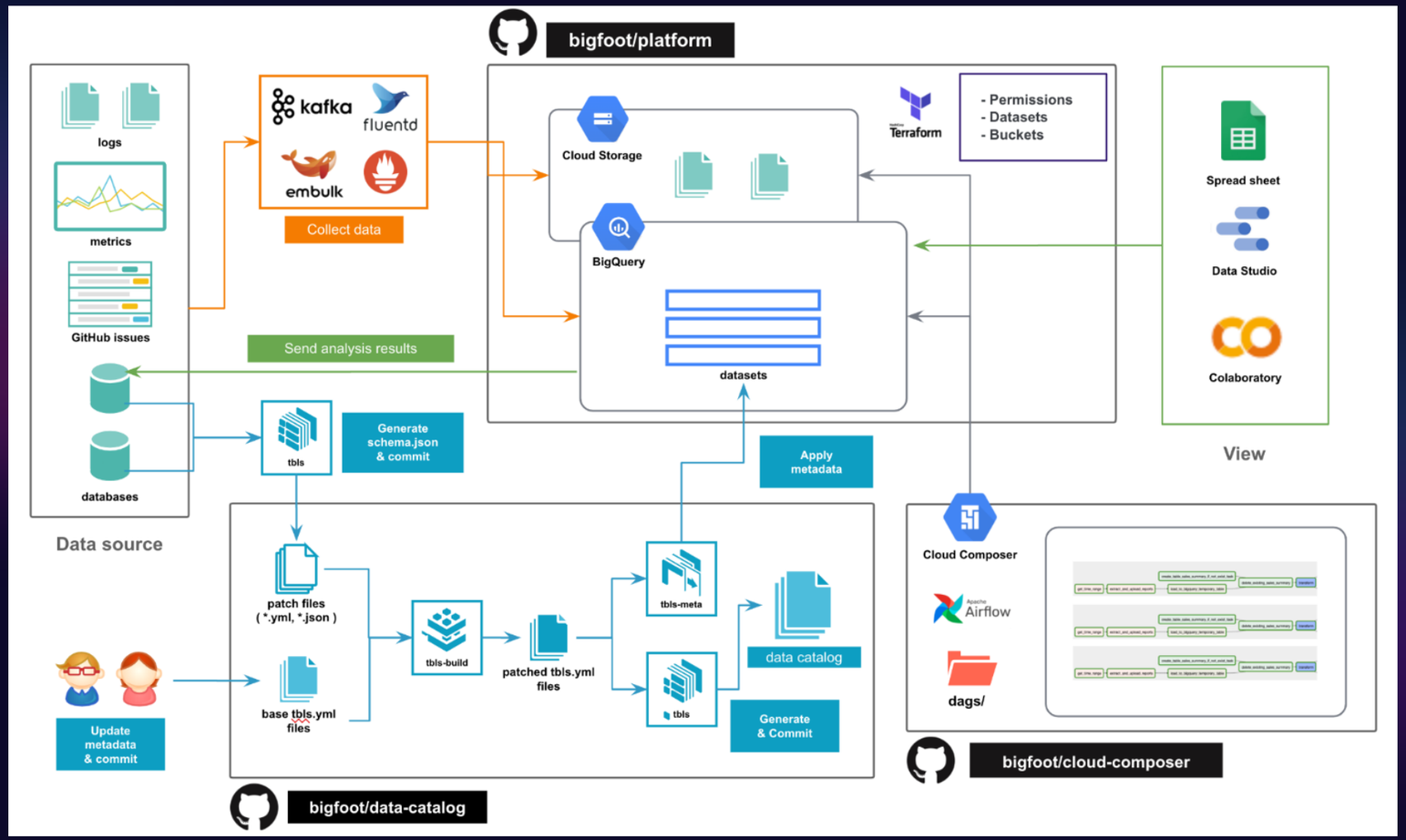
Los flujos de trabajo se definen como Grafos Acíclicos Dirigidos (DAGs), lo que permite a los usuarios especificar el orden y la dependencia entre las tareas de manera flexible.



Composer

Integración con Otros Servicios de Google Cloud

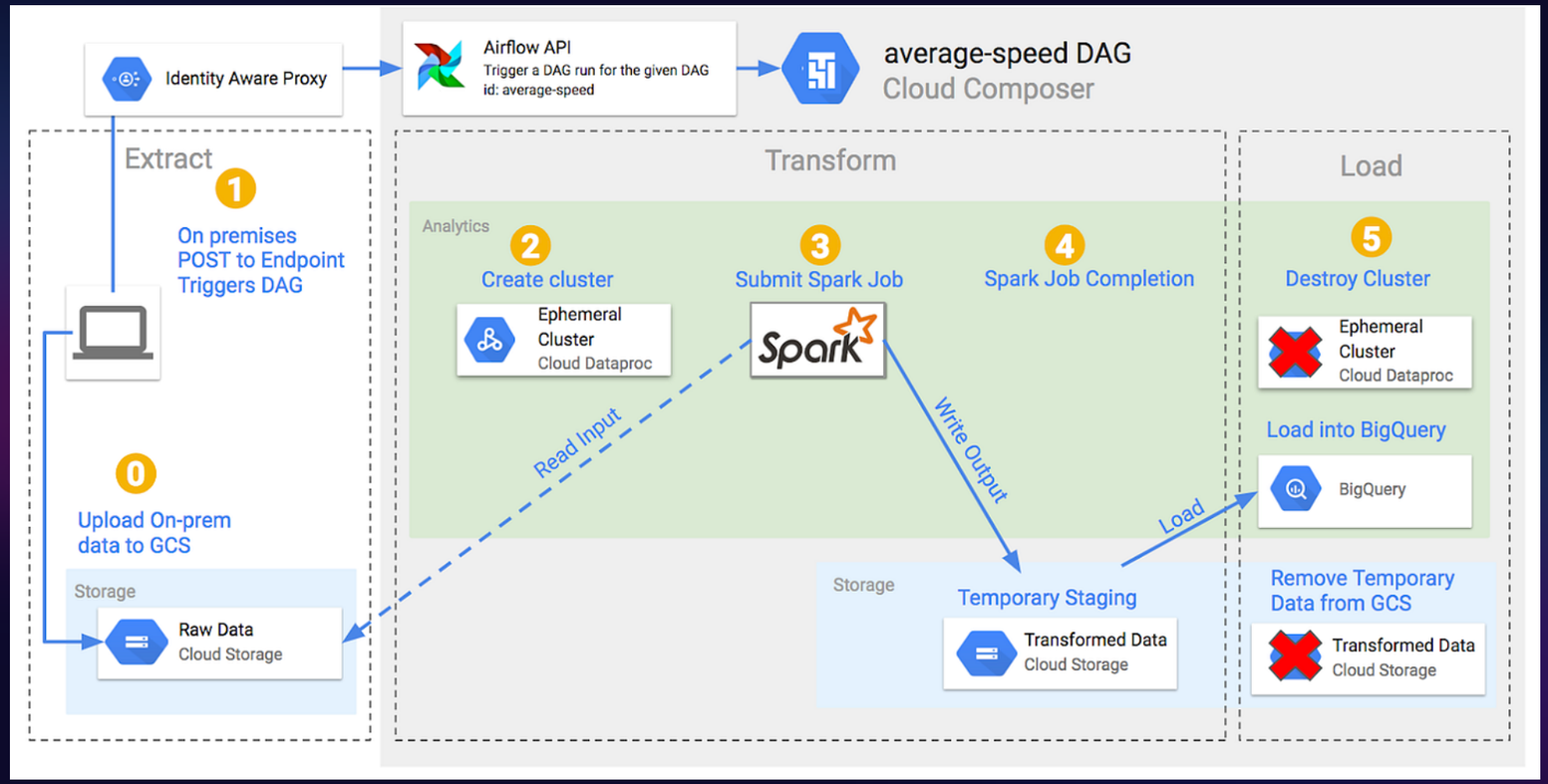
Composer se integra sin problemas con otros servicios de Google Cloud como BigQuery, Dataflow y Pub/Sub, lo que facilita la construcción de soluciones de integración de datos robustas.



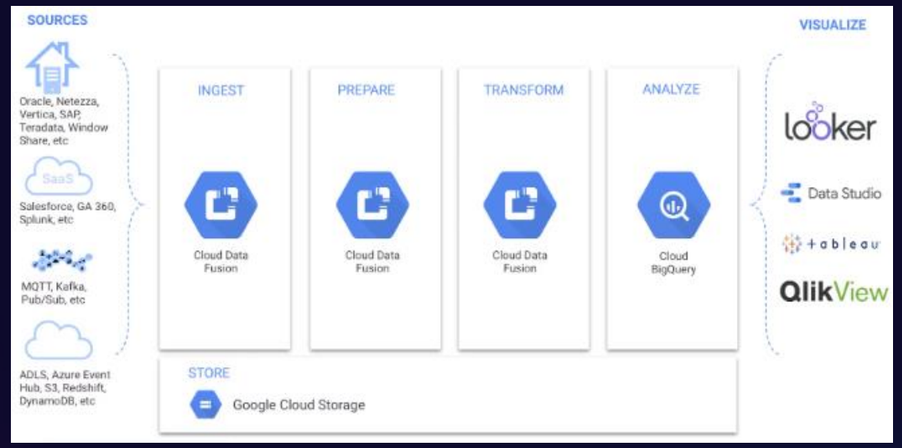
Composer

Basado en Apache Airflow

Composer aprovecha la flexibilidad y potencia de Apache Airflow, un motor de orquestación de código abierto, para crear flujos de trabajo personalizados y escalables.



Data Fusion



Data Fusion es una poderosa herramienta de Google que permite integrar y procesar datos de múltiples fuentes en tiempo real. Ofrece una interfaz visual intuitiva para diseñar y ejecutar flujos de trabajo de integración de datos sin necesidad de escribir código.

Creación de pipelines sin código

Permite diseñar pipelines de datos de forma visual y sin necesidad de escribir código, lo que facilita su uso por parte de usuarios no técnicos.

Basado en Apache Spark y Apache Hadoop

Utiliza tecnologías de procesamiento de datos ampliamente utilizadas en la industria para garantizar un alto rendimiento y escalabilidad

Integración con otros servicios de Google Cloud:

Se integra con servicios como BigQuery, Cloud Storage y Pub/Sub, lo que facilita la integración de pipelines con la infraestructura existente en Google Cloud

Data Fusion

Ventajas

1 Facilidad de uso

Permite a usuarios no técnicos crear pipelines de datos de forma intuitiva y sin necesidad de conocimientos avanzados de programación.

2 Escalabilidad y rendimiento

Utiliza tecnologías como Apache Spark para ofrecer alta escalabilidad y rendimiento en el procesamiento de datos.

3 Integración con Google Cloud

Se integra fácilmente con otros servicios de Google Cloud, lo que simplifica la implementación de pipelines complejos.

Desventajas

1 Limitaciones en la complejidad de los pipelines:

Al no requerir código, puede tener limitaciones en la complejidad de los pipelines que se pueden crear sin necesidad de escribir código personalizado.

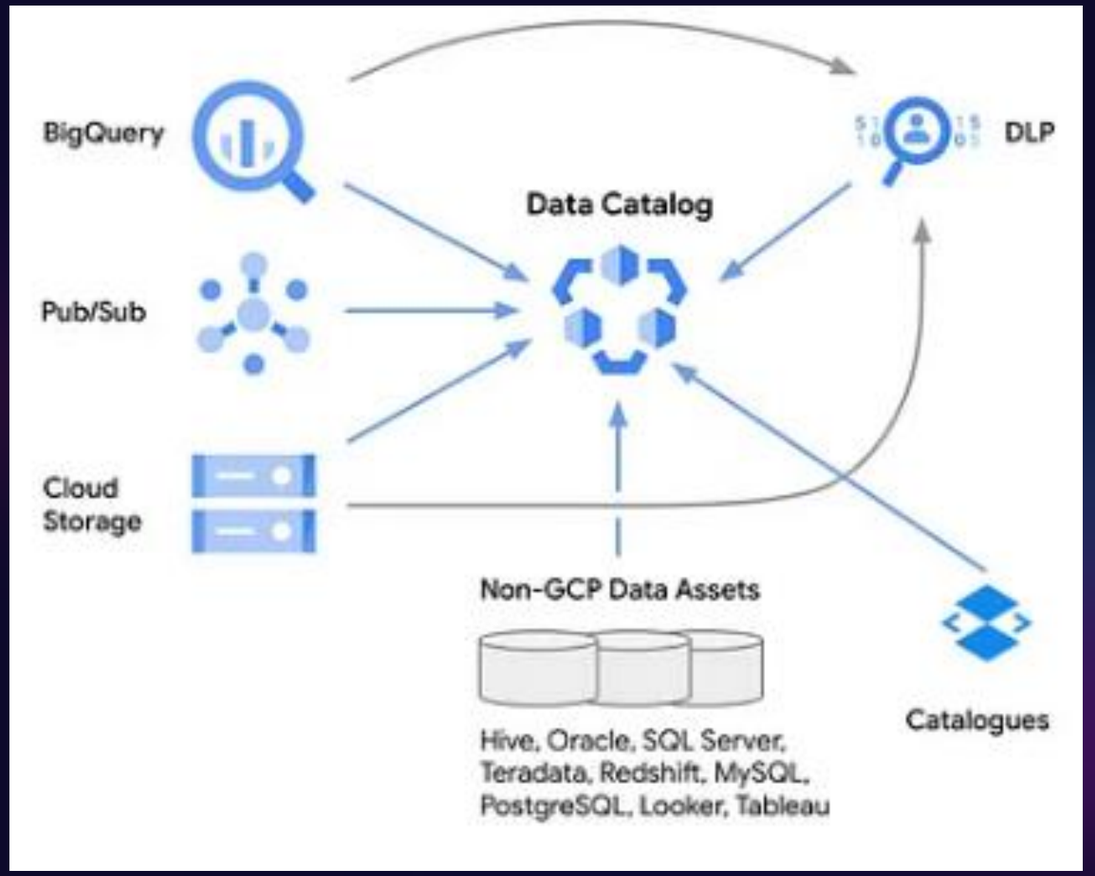
Ejemplos de Uso

- Ingestión de datos desde diferentes fuentes, como bases de datos, archivos y servicios web.
- Transformación de datos para análisis y generación de informes.
- Integración de datos en almacenes de datos para análisis posterior.

Data Catalog

Google Data Catalog es una herramienta de catálogo de datos totalmente administrada que permite a las organizaciones descubrir, comprender y gobernar todos sus datos tanto internos como externos. Con Data Catalog, los usuarios pueden buscar, indexar y organizar fácilmente los activos de datos de la empresa para impulsar análisis y datos eficaces.

Data Catalog facilita la gestión y el descubrimiento de datos al proporcionar un catálogo central, intuitivo y basado en metadatos que ayuda a los usuarios a encontrar rápidamente los datos que necesitan. Esto permite a las empresas obtener más valor de sus datos y tomar mejores decisiones.



Conclusión

En resumen, hemos explorado el poderoso ecosistema de integración de datos de Google, que ofrece una variedad de herramientas y servicios para gestionar de manera eficiente los datos a lo largo de todo el ciclo de vida. Desde el almacenamiento y procesamiento en BigQuery y Dataflow, hasta la orquestación con Composer y la integración con otras fuentes de datos a través de Pub/Sub y Data Fusion, este ecosistema brinda soluciones integrales para satisfacer las necesidades de integración de datos de las empresas modernas.