



Google Cloud

Cargando Datos a
Bigquery

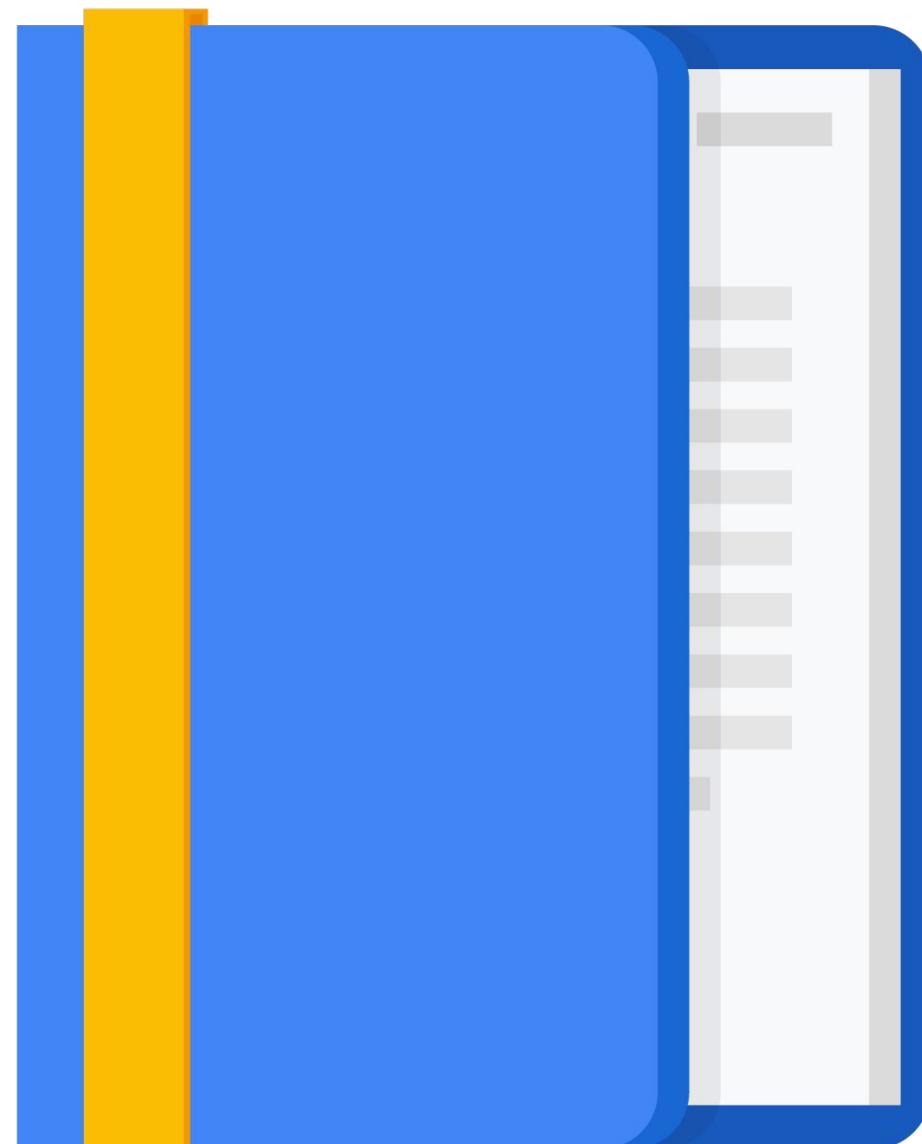
Temario

Cómo llegamos a ELT?

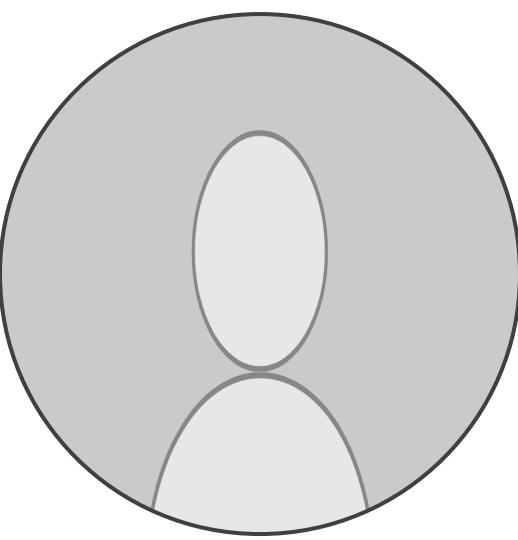
Cargando Datos a Bigquery ELT

Ingesta Bachera

Metodologías de Carga

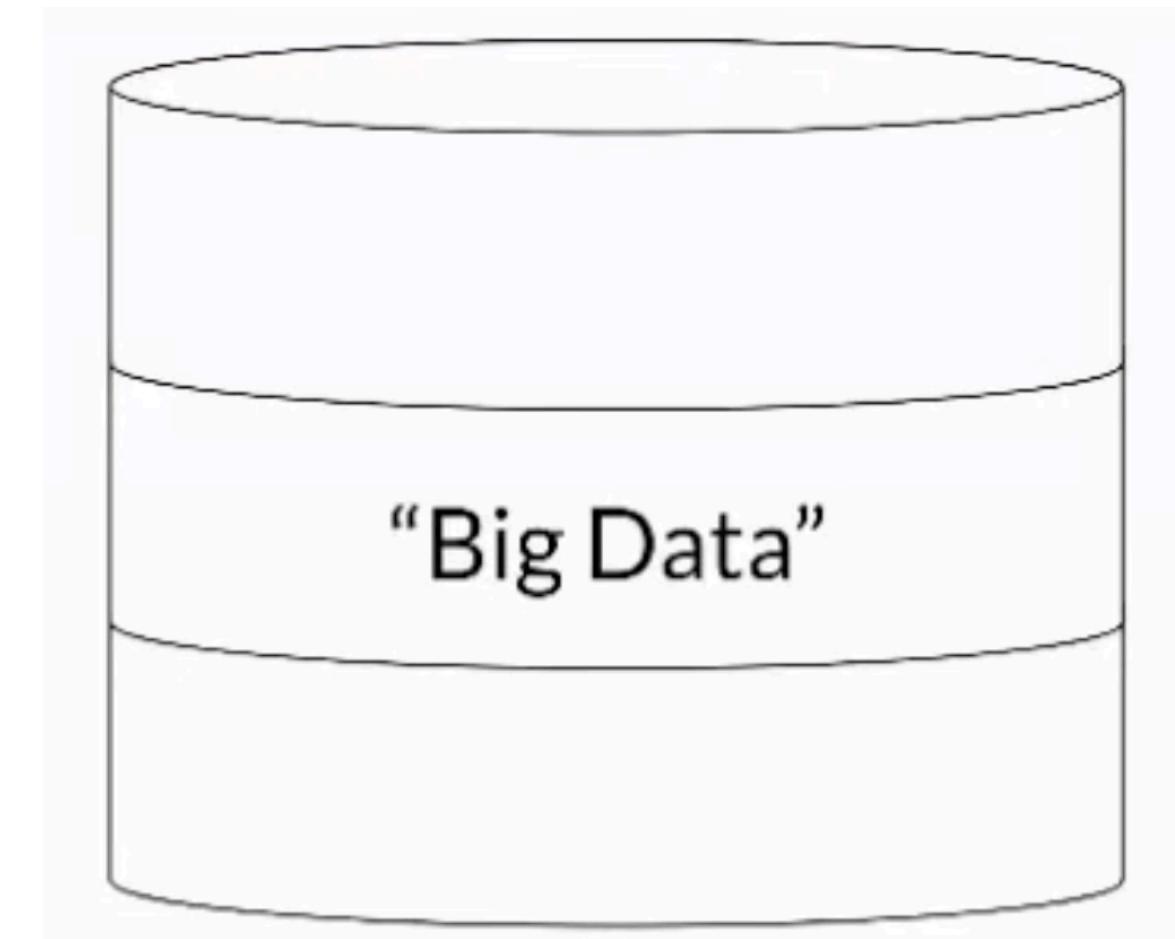


Cómo llegamos a ELT?



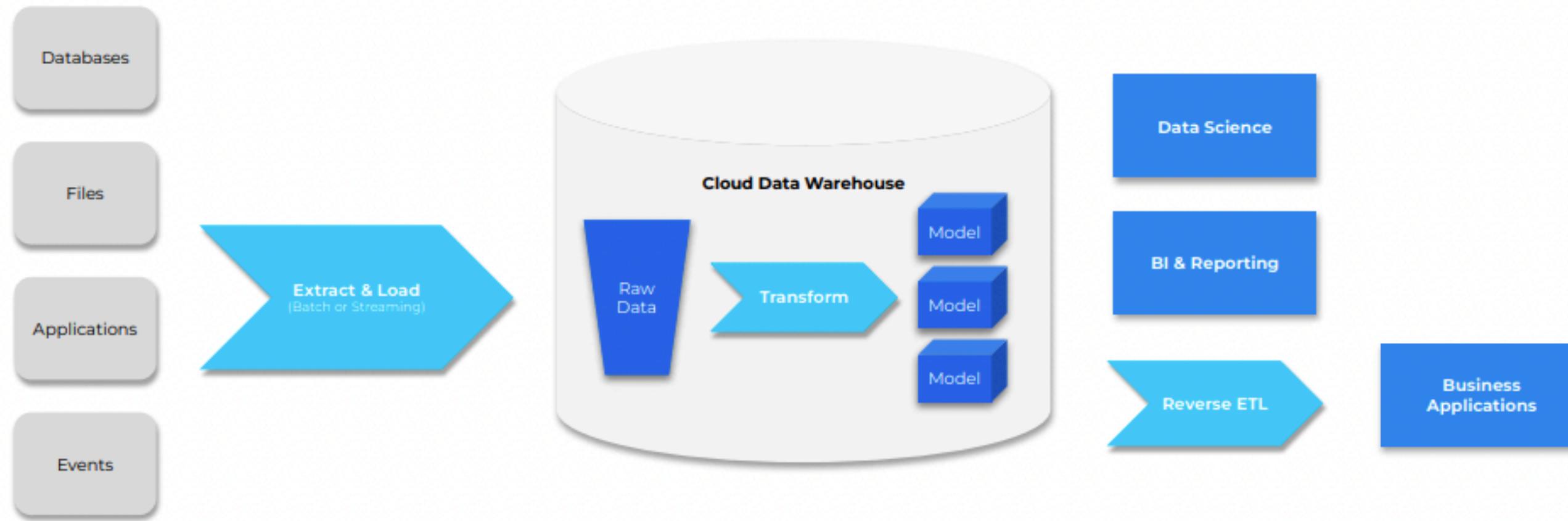
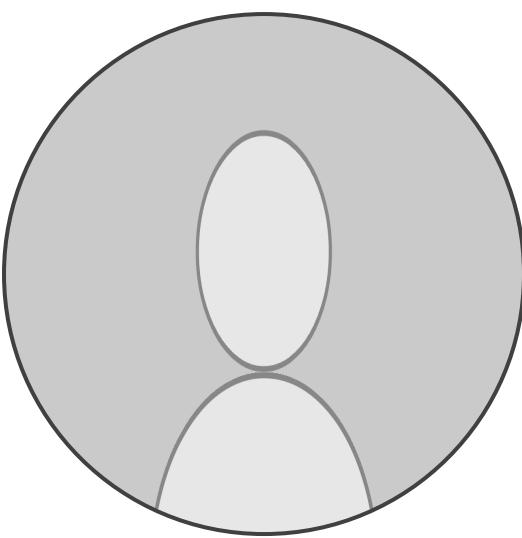
En 2010, el foco de la mayoría de equipos de analítica estaba en infraestructura, computo y almacenamiento para datasets grandes.

- ¿Cómo construimos/escalamos nuestra infraestructura ETL?
- ¿Cómo controlamos los gastos de almacenamiento?
- ¿Cómo diseñamos nuestro data warehouse para un buen desempeño?



Como llegamos ELT?

Hoy, el cambio hacia ELT, y data warehouse ha resuelto estos problemas



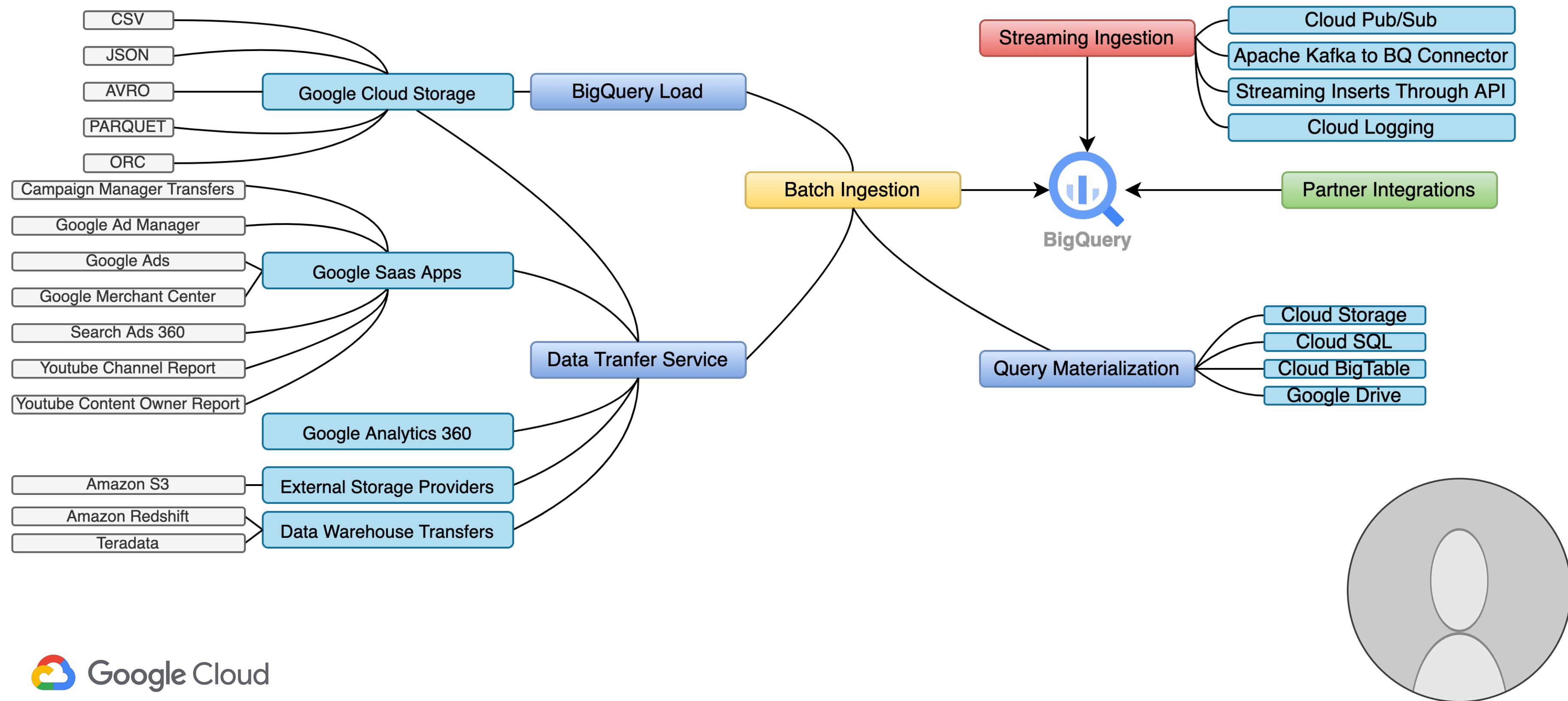
Mover Datos al Warehouse es fácil!

- Herramientas de ingesta han automatizado cerca casi todos los procesos de ingesta.

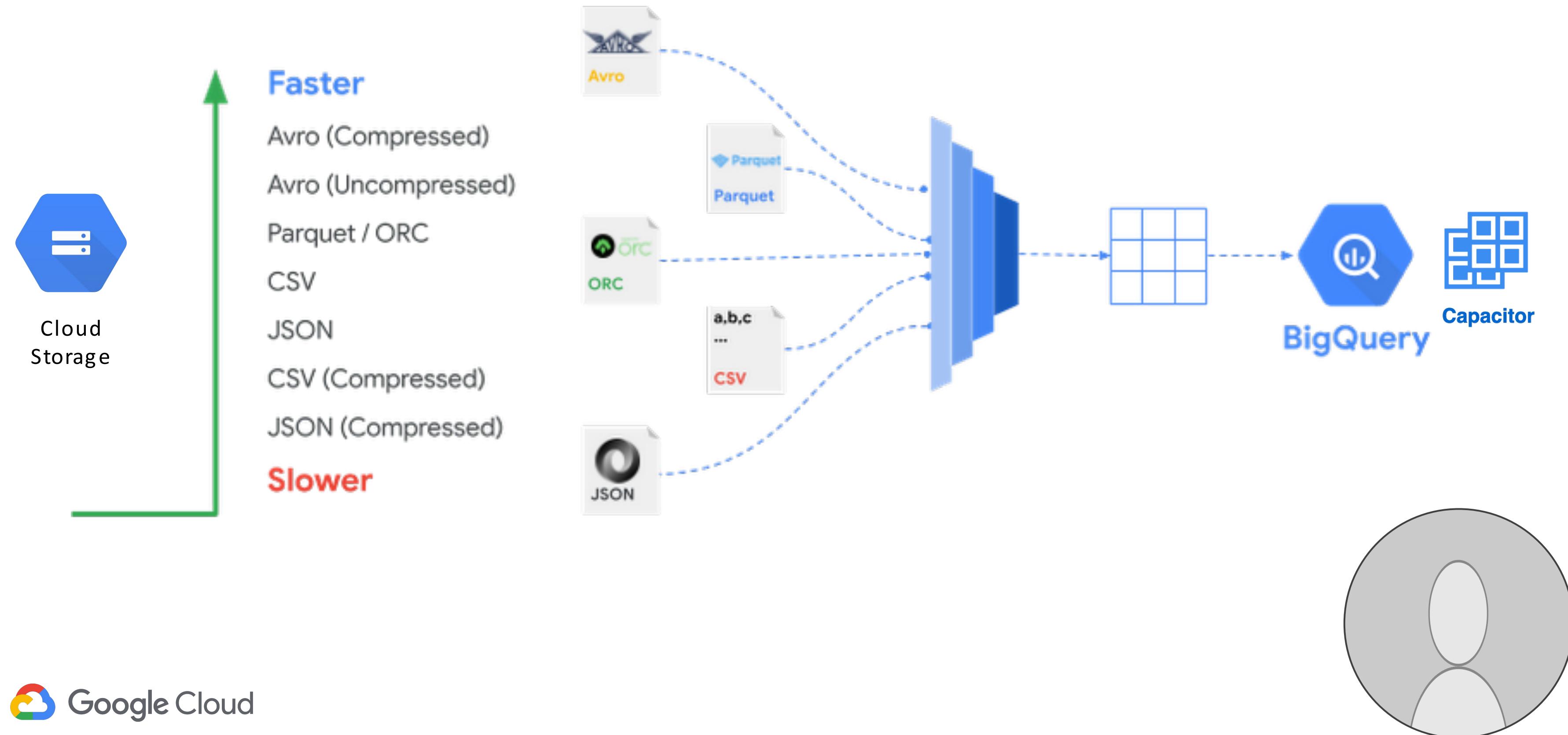
Warehouse remueve el sobrecargo en infraestructura

- Barato almacenamiento elimina la necesidad por un Lake separado.
- Pagos a crédito/ servicios autogestionados elimina la necesidad de infraestructura compleja.
- Transformaciones en SQL elimina la necesidad de tenerlas en frameworks de procesamiento.

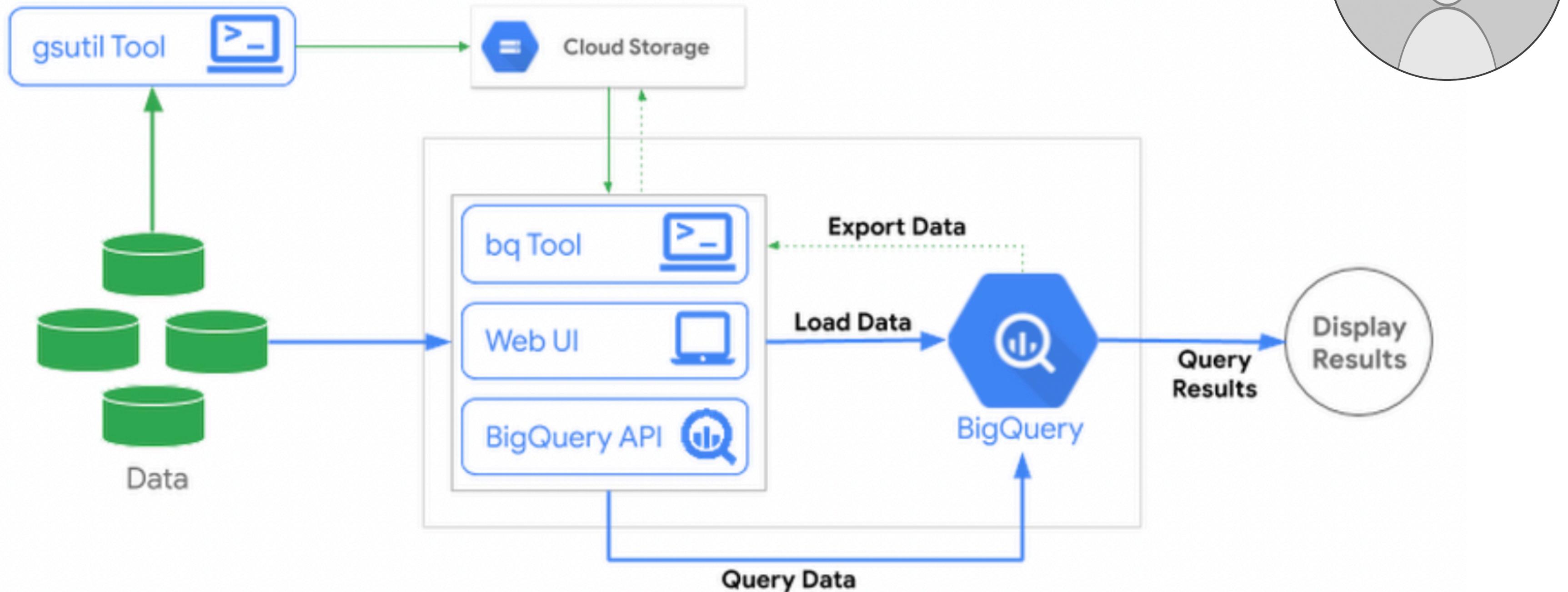
Cargando Datos a Bigquery ELT



Ingesta Bachera – Formatos de Archivo Soportados

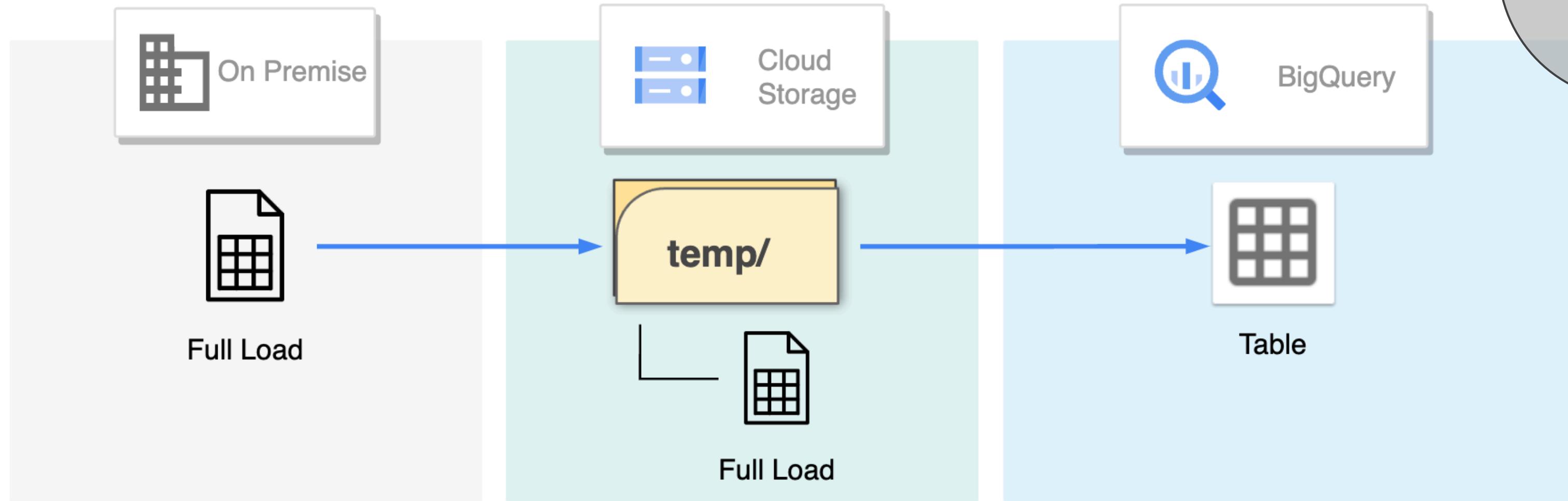
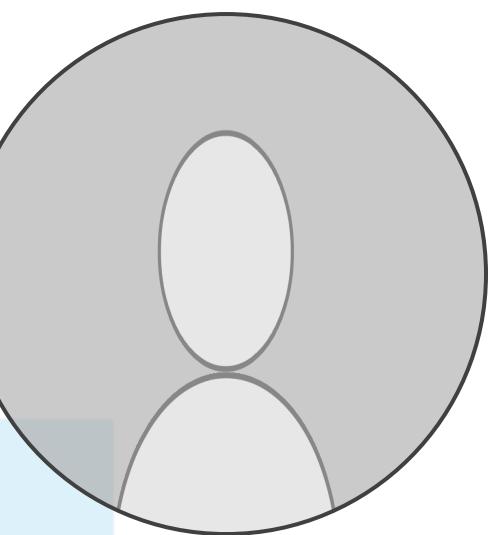


Ingesta Bachera – Bigquery Load



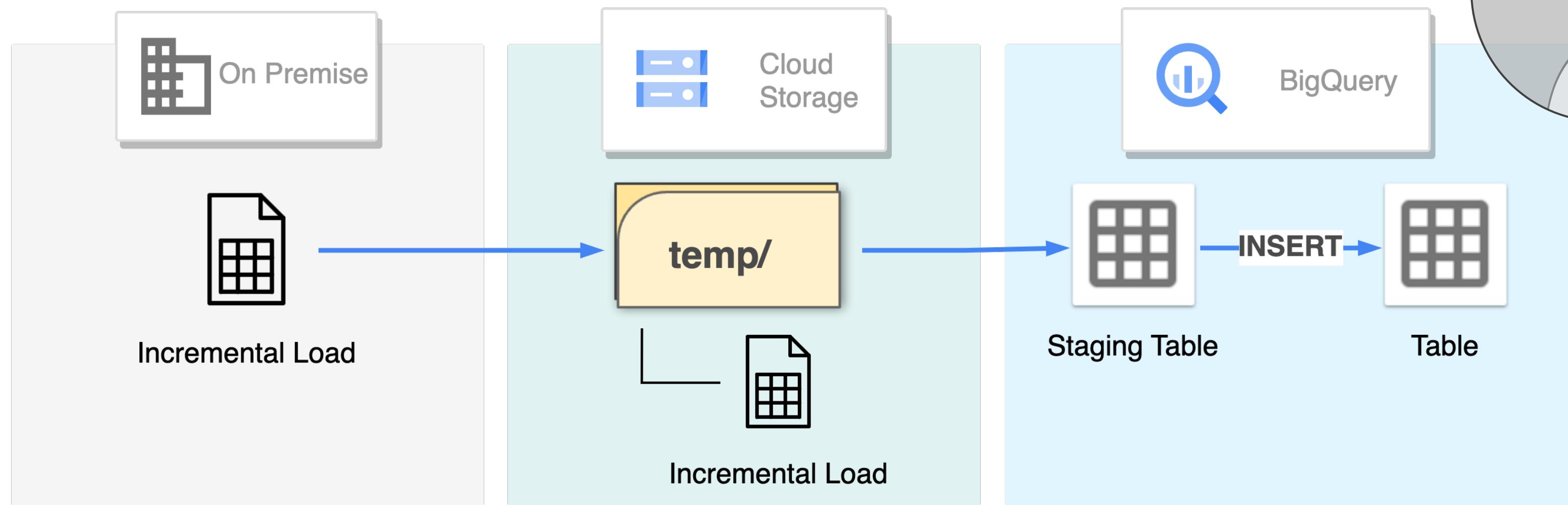
- Ingesta Bachera no tiene costo.

Metodologías de Carga – Full Load



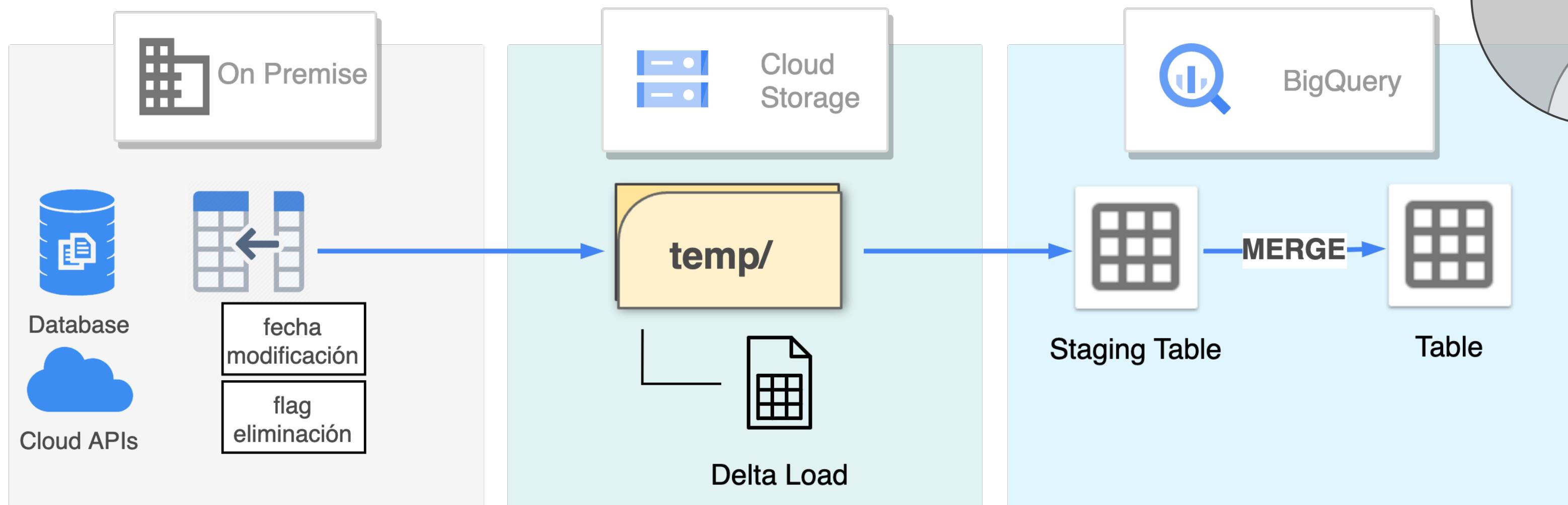
- Poco Volumen
- **No se tiene forma de identificar registros en el tiempo.**
- Dimensiones/ Tablas estáticas.(Ej: ubigeos. catálogos de producto, tiendas, etc)
- Carga Histórica o Una sola vez.
- Frecuencia de actualización es baja.

Metodologías de Carga – Incremental Load



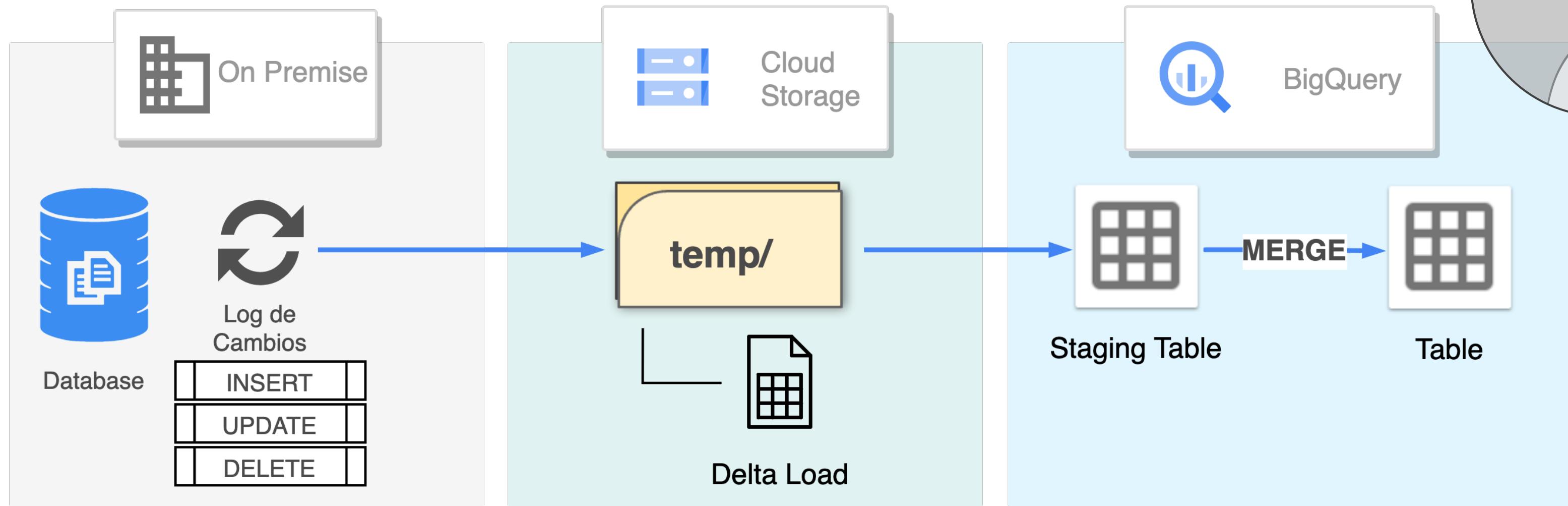
- Gran Volumen
- **Los datos se pueden segmentar por periodo(día o mes)**
- Base de Datos Externas (Ej: RCC, RENIEC, ESSALUD, SUNARP, lista de precios, datos financieros, etc)
- Frecuencia de actualización es baja.

Metodologías de Carga – Incremental Replication



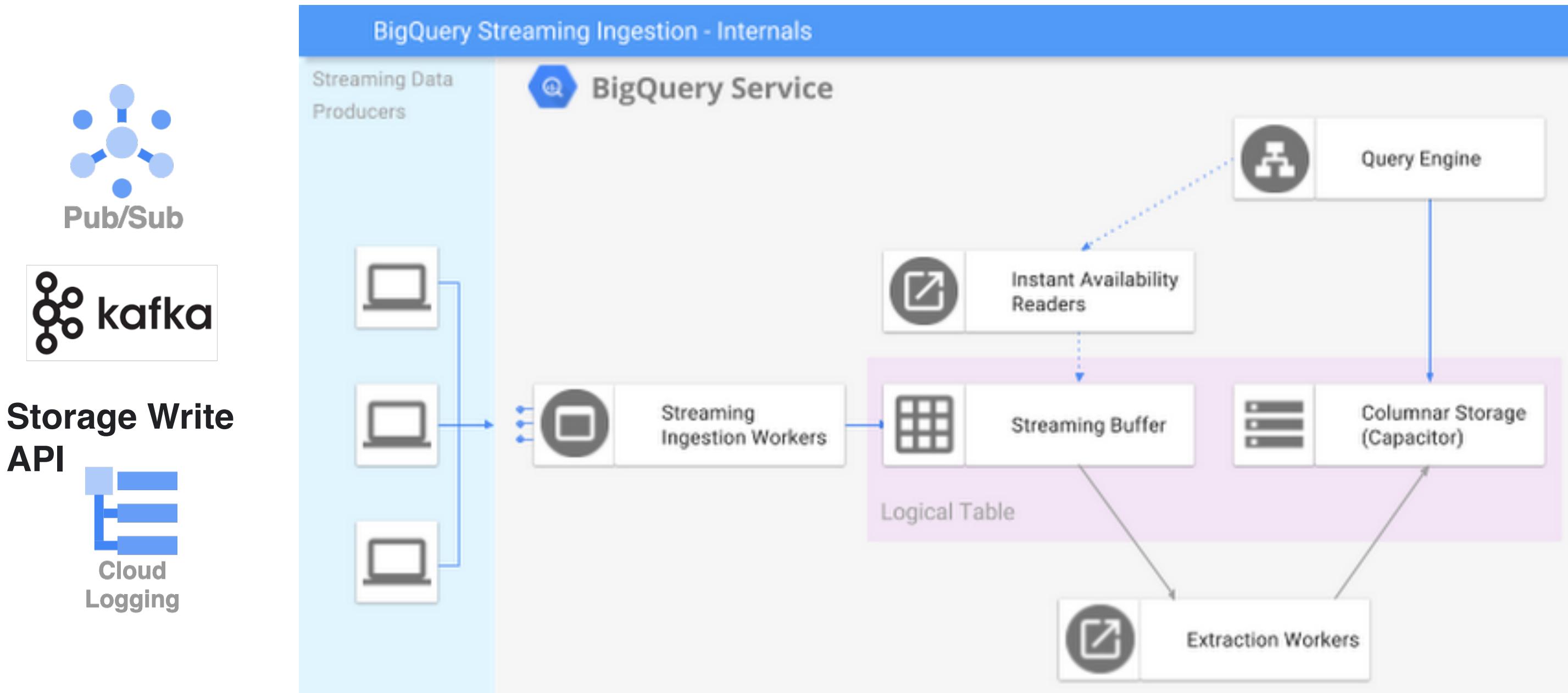
- Gran Volumen
- La base de datos tiene **campos adicionales de auditoría**: **última fecha de actualización y flag de eliminación** (en caso quieras mover también las eliminaciones).
- Tablas transaccionales(Ej: Salesforce, Dynamics 365, Big Table, etc)
- Frecuencia de actualización es alta.

Metodologías de Carga – Batch CDC Replication



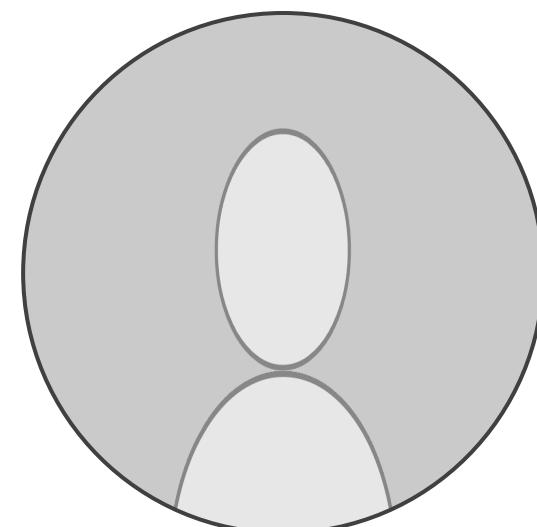
- Gran Volumen
- La base de datos tiene activo logs de CDC y las tablas tienen PK.
- Tablas transaccionales (Bases de Datos internas de aplicativos de la compañía).
- Frecuencia de actualización es alta.

Metodologías de Carga – Streaming

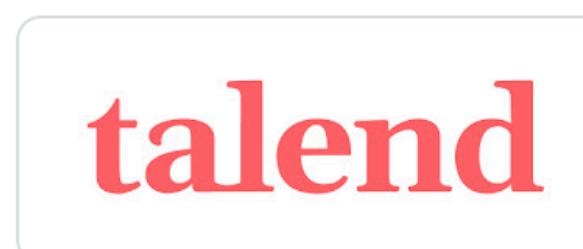
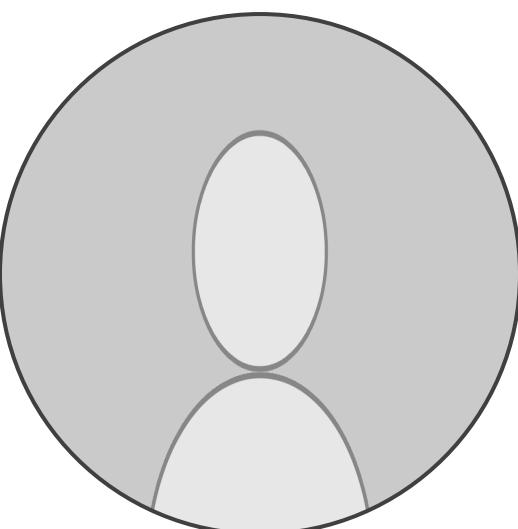


- Gran Volumen
- **Solo INSERTS.**
- Datos de inserciones (**IOT DATA, REST API**).
- Frecuencia de registro es alta.

Google Cloud



Herramientas de Replicación e Integración de Datos



Herramientas de Replicación e Integración de Datos

Company	Url	ETL	ELT		Data Preparation Solution	Streaming Replication
			Incremental Replication	Batch CDC Replication		
Matillion	https://www.matillion.com/	x				
Informatica	https://www.informatica.com/	x		x		
Snaplogic	https://www.snaplogic.com/	x	x			
Skyvia	https://skyvia.com/		x			
Funnel	https://funnel.io/		x			
Integrate.io	https://www.integrate.io/	x	x		x	
Striim	https://www.striim.com/				x	
Fivetran	https://www.fivetran.com/				x	
Rivery	https://rivery.io/				x	
DataFusion	https://cloud.google.com/data-fusion	x			x	
Talend	https://www.talend.com/	x	x		x	
Alteryx	https://www.alteryx.com/					x
Trifacta	https://www.trifacta.com/					x
Confluent	https://www.confluent.io/					x
RudderStack	https://www.rudderstack.com/					x

Algunos datos se pueden usar en su lugar sin importarlos a las tablas de BigQuery mediante fuentes de datos externas

