

Arquitectura de datos

M.Sc. Ing. Danny Luis Huanca Sevilla

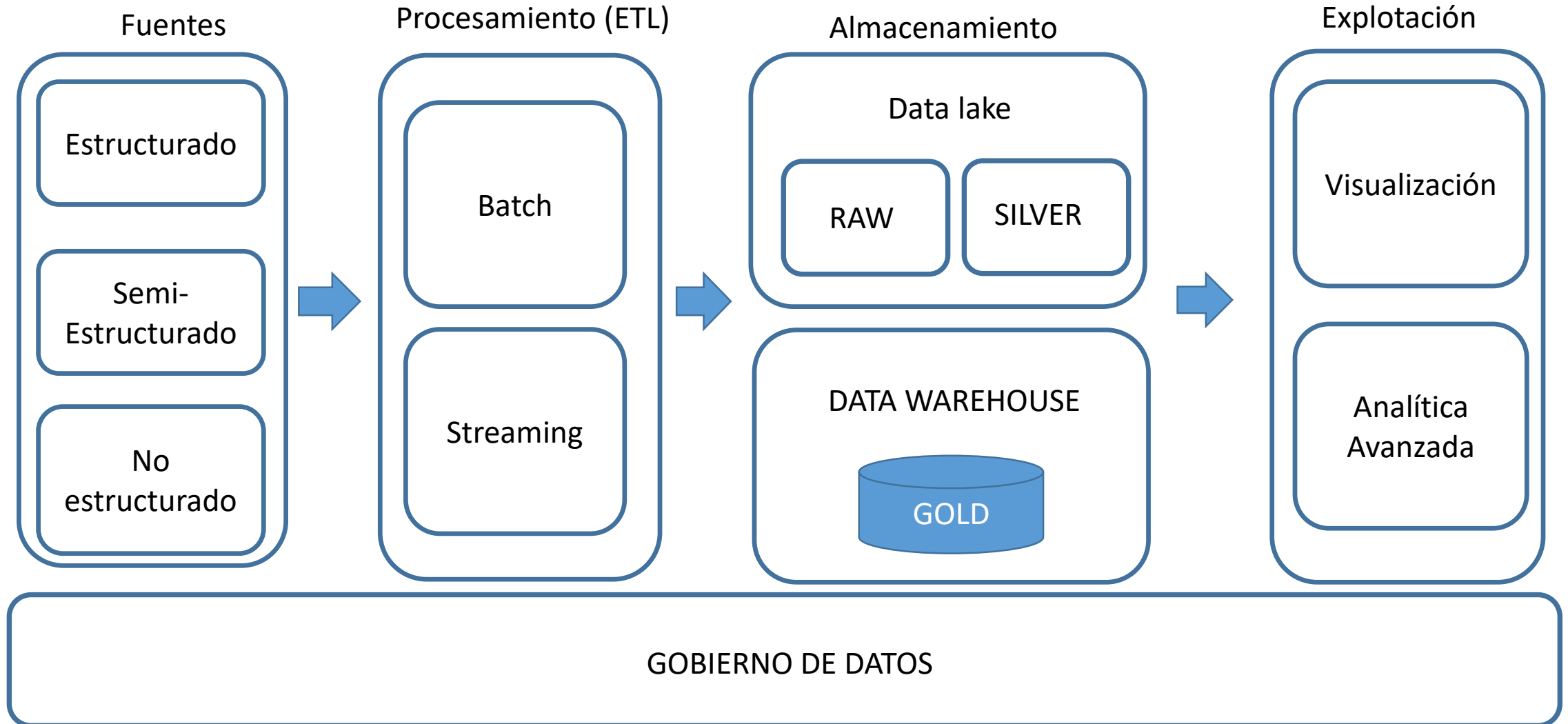
Agenda

- Arquitectura de datos diagrama de alto contexto
- Elementos del framework de arquitectura
- Aplicación al caso TELCO
- Arquitecturas modernas de datos

¿Por qué Arquitectura?



Arquitectura de alto nivel



Fuentes

- Data Estructurada

CUSTOMERID	GENDER	SENIORCITIZEN	PARTNER	DEPENDENTS	TENURE	PHONESERVICE	MULTIPLELINES
1658-BYGOY	Male	1	No	No	18	Yes	Yes
8769-KKTPH	Female	0	Yes	Yes	63	Yes	Yes
5067-XJQFU	Male	1	Yes	Yes	66	Yes	Yes
3957-SQXML	Female	0	Yes	Yes	34	Yes	Yes
5954-BDFSG	Female	0	No	No	72	Yes	Yes
0434-CSFON	Female	0	Yes	No	47	Yes	Yes
1215-FIGMP	Male	0	No	No	60	Yes	Yes
0526-SXDJP	Male	0	Yes	No	72	No	No phone service
0557-ASKVU	Female	0	Yes	Yes	18	Yes	No

- Data Semi estructurada

```
{
  "marcadores": [
    {
      "latitude": 40.416875,
      "longitude": -3.703308,
    },
    {
      "latitude": 40.417438,
      "longitude": -3.693363,
      "description": "Paseo del Prado"
    },
    {
      "latitude": 40.407015,
      "longitude": -3.691163,
      "city": "Madrid",
      "description": "Estación de Atocha"
    }
  ]
}
```

- Data No estructurada



Fuentes

- ¿Pueden dar ejemplo de fuentes Estructurada, no estructurada o semiestructurada?
- Que porcentaje de tipo de fuente utiliza en su organización?

CUSTOMERID	GENDER	SENIORCITIZEN	PARTNER	DEPENDENTS	TENURE	PHONESERVICE	MULTIPLELINES
1658-BYGOY	Male	1	No	No	18	Yes	Yes
8769-KKTPH	Female	0	Yes	Yes	63	Yes	Yes
5067-XJQFU	Male	1	Yes	Yes	66	Yes	Yes
3957-SQXML	Female	0	Yes	Yes	34	Yes	Yes
5954-BDFSG	Female	0	No	No	72	Yes	Yes
0434-CSFON	Female	0	Yes	No	47	Yes	Yes
1215-FIGMP	Male	0	No	No	60	Yes	Yes
0526-SXDJP	Male	0	Yes	No	72	No	No phone service
0557-ASKVU	Female	0	Yes	Yes	18	Yes	No



```
{
  "marcadores": [
    {
      "latitude": 40.416875,
      "longitude": -3.703308,
    },
    {
      "latitude": 40.417438,
      "longitude": -3.693363,
      "description": "Paseo del Prado"
    },
    {
      "latitude": 40.407015,
      "longitude": -3.691163,
      "city": "Madrid",
      "description": "Estación de Atocha"
    }
  ]
}
```

Procesamiento (ETL)

- Es el motor que hace que funcione la Fábrica.
- Este procesamiento puede ser:
 - Batch. Cuando se realiza periódicamente cada cierto tiempo. Ejemplo la carga de suscriptores de una empresa retail diaria.
 - Streaming. Cuando se realiza en tiempo real y la necesidad del negocio lo requiere así
- ¿Qué porcentaje de procesos de acuerdo a la necesidad del negocio son Batch y cuantos son Streaming?

Batch

- Ejemplo proceso batch:
 - Una gerente de Marketing recibe información diaria del consolidado de ventas del día anterior en un informa que contraste lo presupuestado vs lo ejecutado con la finalidad de tomar acciones respecto a los canales de venta.

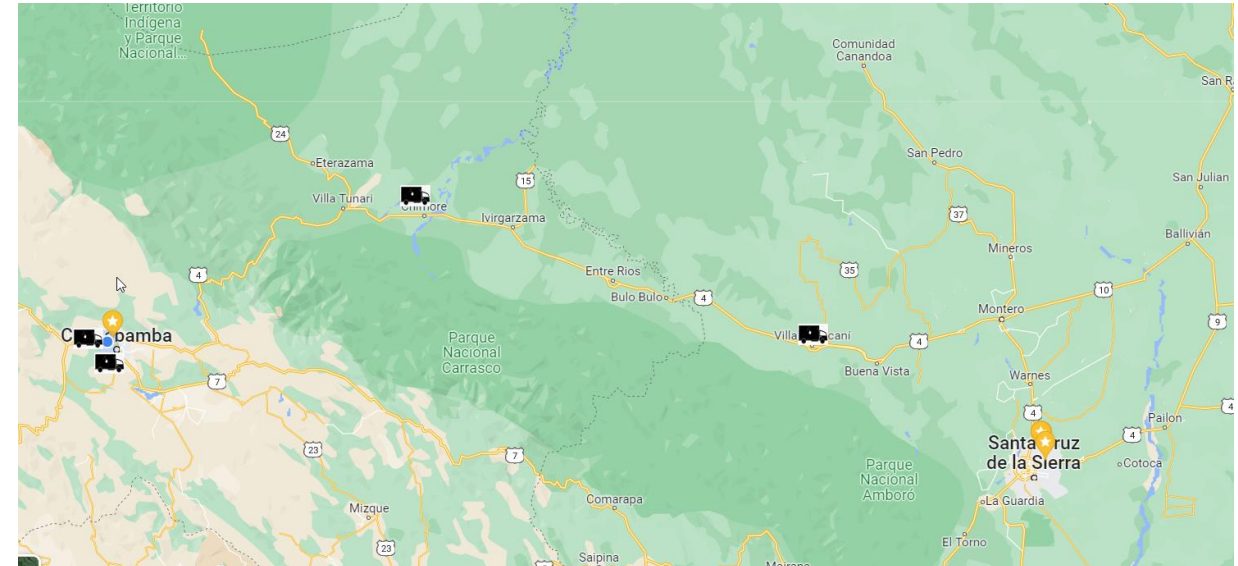
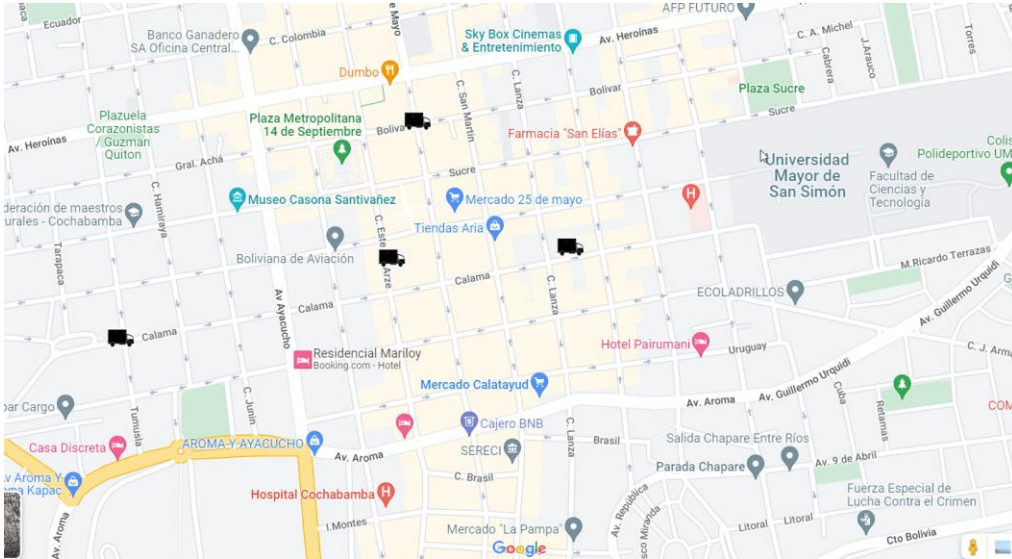


Batch

- ¿Puede dar un ejemplo de proceso batch en su organización?

Streaming

- Ejemplo proceso streaming.
 - Una empresa de transporte necesita la geolocalización en tiempo real de su flota de camiones para tomar acciones respecto a posibles retrasos debido a condiciones climáticas (nieblina) en las zonas montañosas.



Streaming

- ¿Puede dar un ejemplo de proceso Streaming en su organización?

Almacenamiento

- Se refiere al lugar en el que se irán almacenando los datos procesados en el ETL.
- Dependiendo si son estructurados, semiestructurados y no estructurados se pueden **almacenar** en un data lake (semi estructurados o no estructurados) o cuando ya están refinados en un Data warehouse.
- Los datos pasan por etapas (stages) desde zonas raw pasando a zonas silver y finalmente gold

Almacenamiento

- Data lake
 - Un data lake es un repositorio de información que tiene características de Big Data:
 - Volumen – Puede almacenar grandes archivos o mucha cantidad de archivos estructurados o semi estructurados Ej: Fotos de gatos, fotos de documentos de la organización escaneados y que se generaron en el siglo XX.
 - Variedad – Se puede almacenar todo tipo de formatos de información: fotos jpg, videos mp4, archivos Excel xlsx.
 - Por analogía es parecido a Google Drive. Google tiene su versión denominada Google Storage y Amazon tiene los S3

Almacenamiento

- Data lake
- Amazon s3: <https://aws.amazon.com/es/s3/>

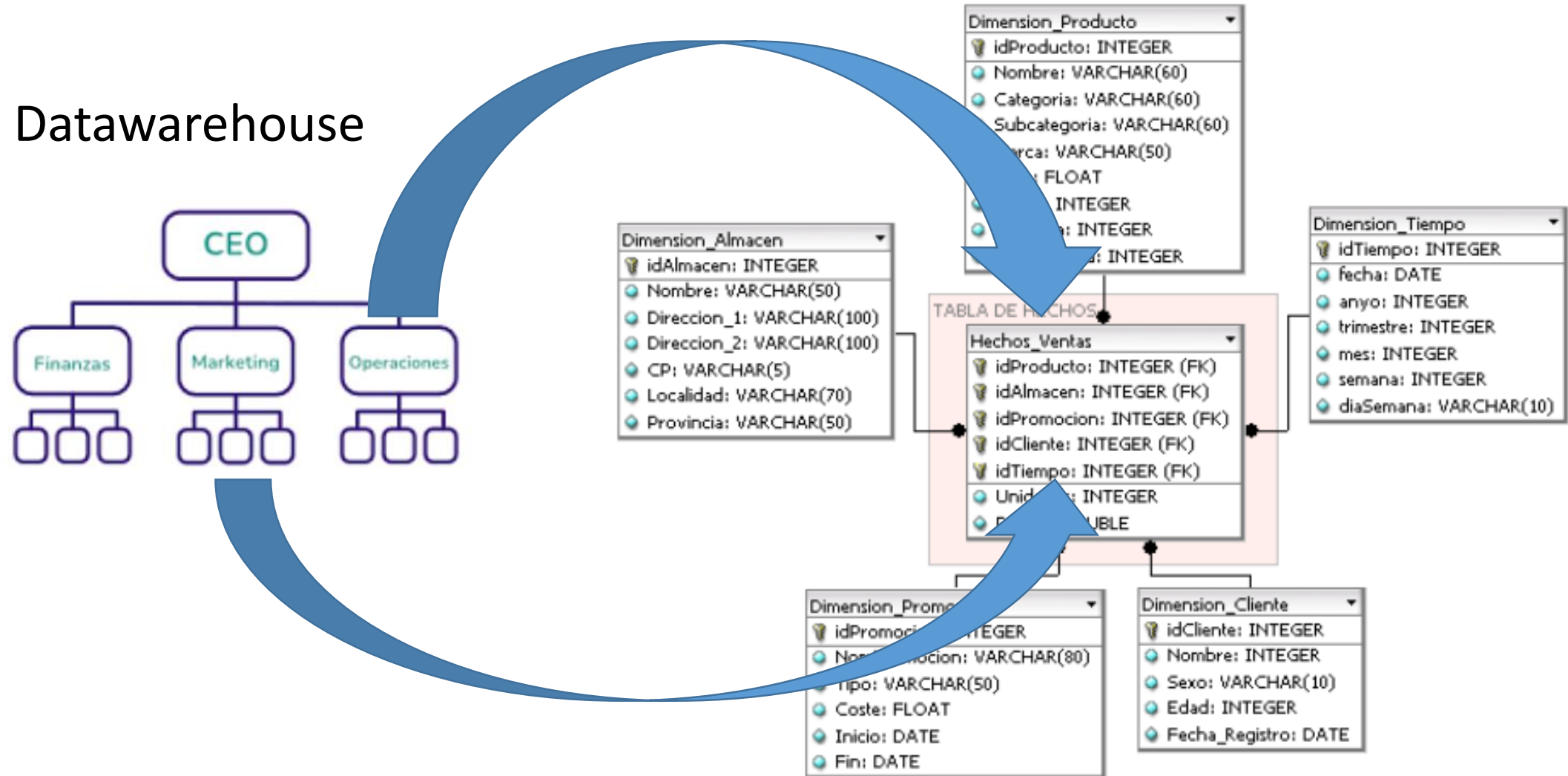


Almacenamiento

- Datawarehouse
 - Es un repositorio de datos estructurados.
 - Utiliza una metodología para la implementación
 - <https://www.kimballgroup.com/>
 - Se basa en crear estructuras multidimensionales, que se construyen por partes denominadas Datamarts que se asocian a una parte del negocio
 - Contienen Tablas de hechos, que normalmente están llenas de métricas de un negocio

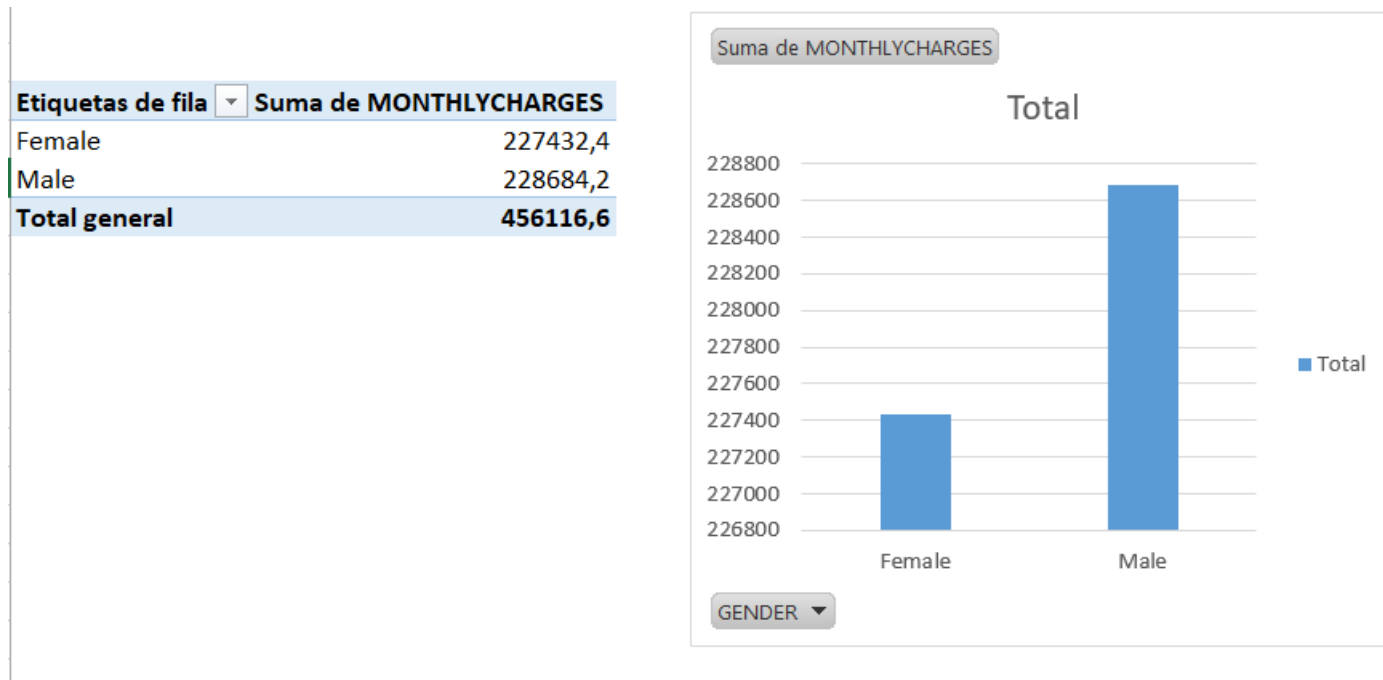
Almacenamiento

- Datawarehouse



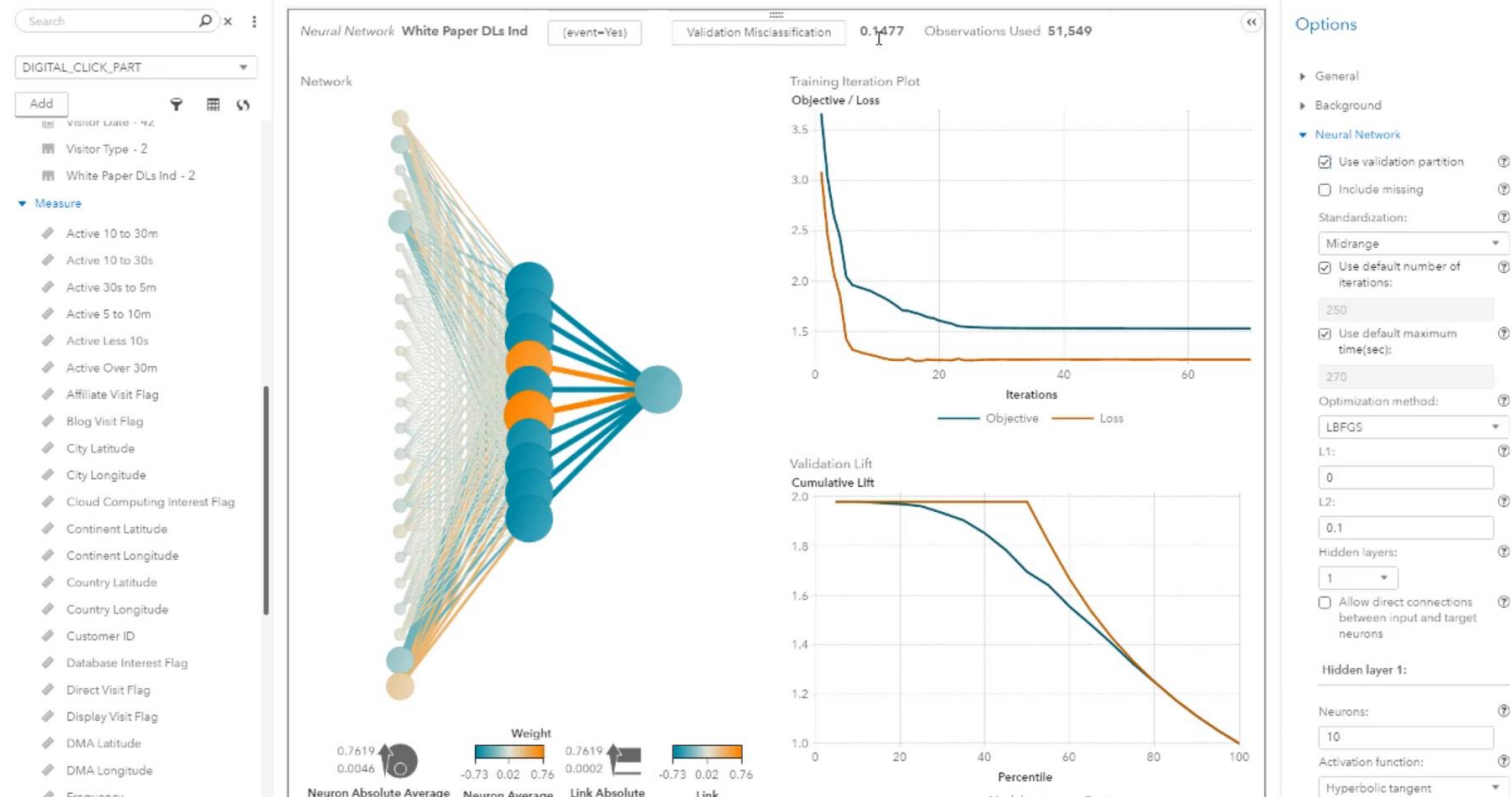
Explotación

- Visualización de datos
 - Herramientas como Tableau, Excel, Power BI



Explotación

- Modelo de churn prectivo



Gobierno de datos

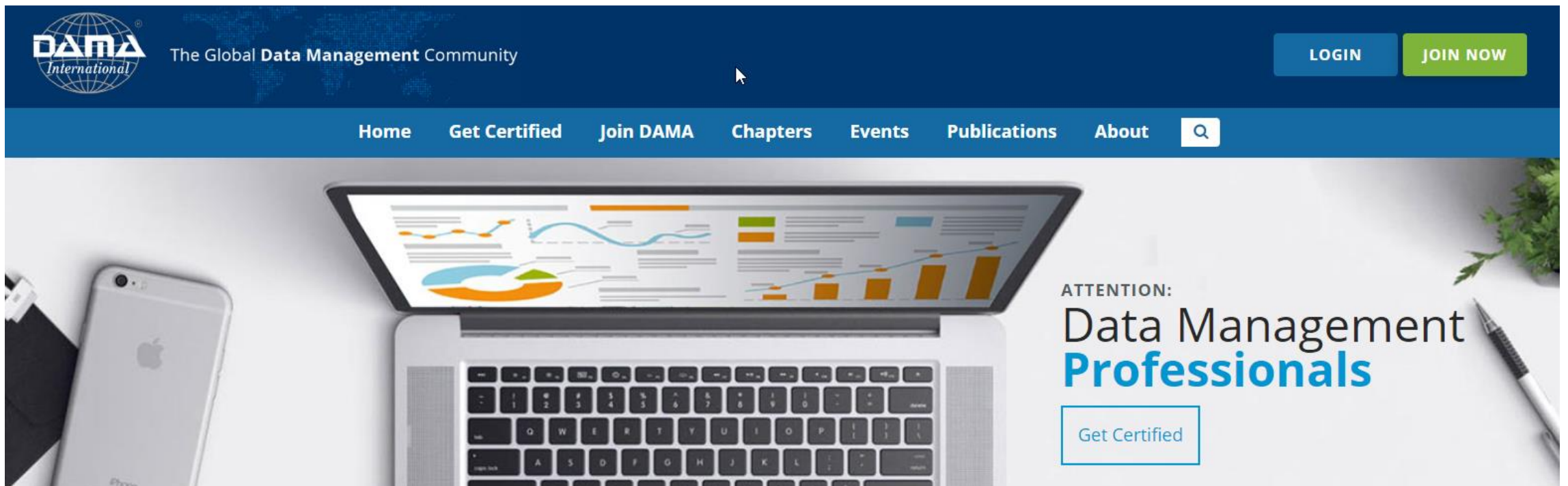
- Principios de Administración de datos
 - Los datos son un activo con características únicas. Intagibles, no se gastan
 - Los datos pueden ser expresados en valores económicos. Ejemplo que pasaría si un dato no se entrega a tiempo o si se pierde?
 - Administrar datos implica administrar la calidad de datos
 - Administrar datos implica manejar la metadata
 - Administra datos requiere equipos multidisciplinarios.
 - Se requiere una perspectiva empresarial coordinada con arquitectos empresariales y arquitectos de datos.
 - La administración de datos debe estar acorde al ciclo de vida de un dato en la organización

Gobierno de datos

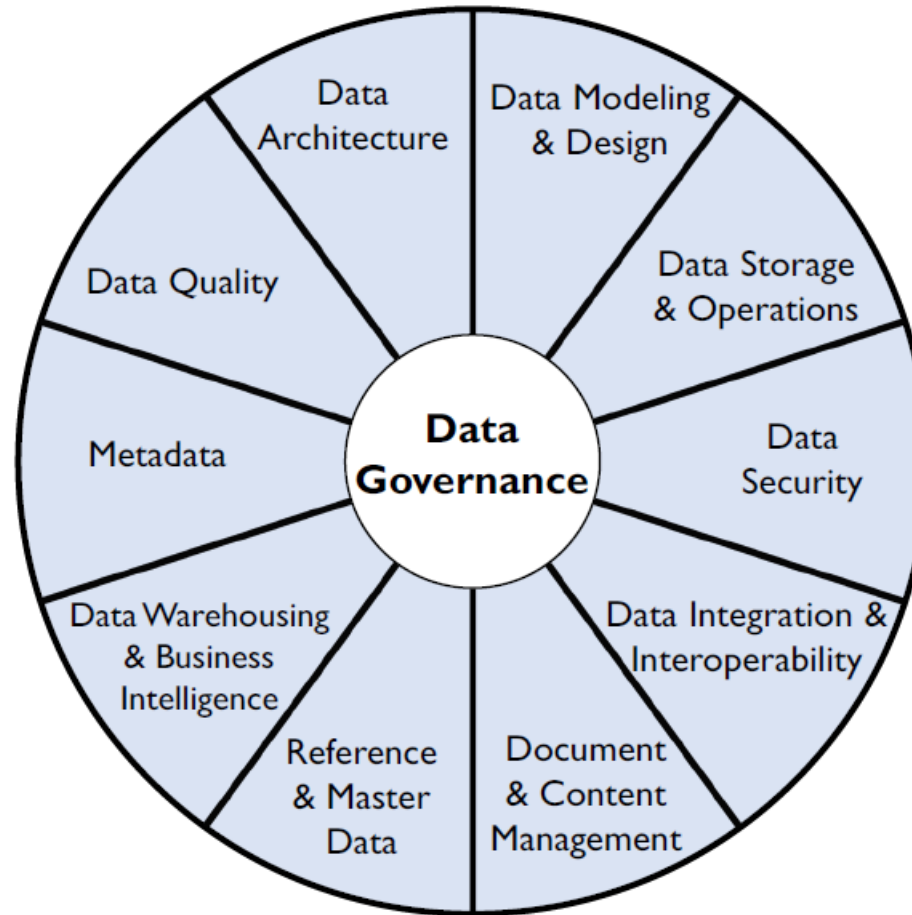
- Principios de Administración de datos
 - Administrar datos implica manejar el riesgo asociado a los mismos. Seguridad de la información.
 - La gestión de datos debe manejar las decisiones que tome IT respecto de los mismos. La tecnología debe servir y no dirigir a las necesidades estratégicas del negocio.
 - Una efectiva administración de datos requiere un compromiso de liderazgo. Se requiere coordinación entre áreas, colaboración y compromiso.

DAMA

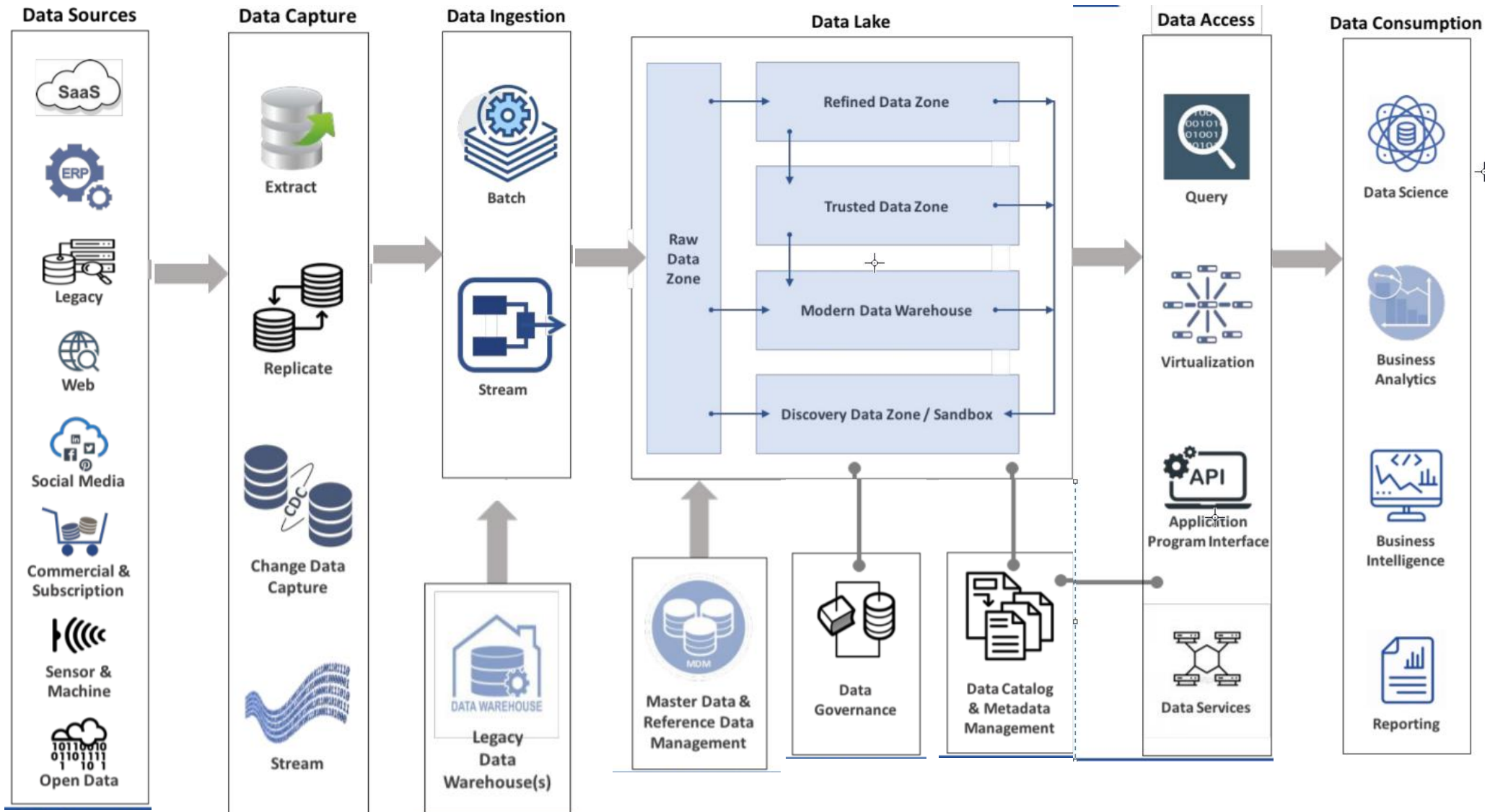
- <https://www.dama.org/cpages/home>



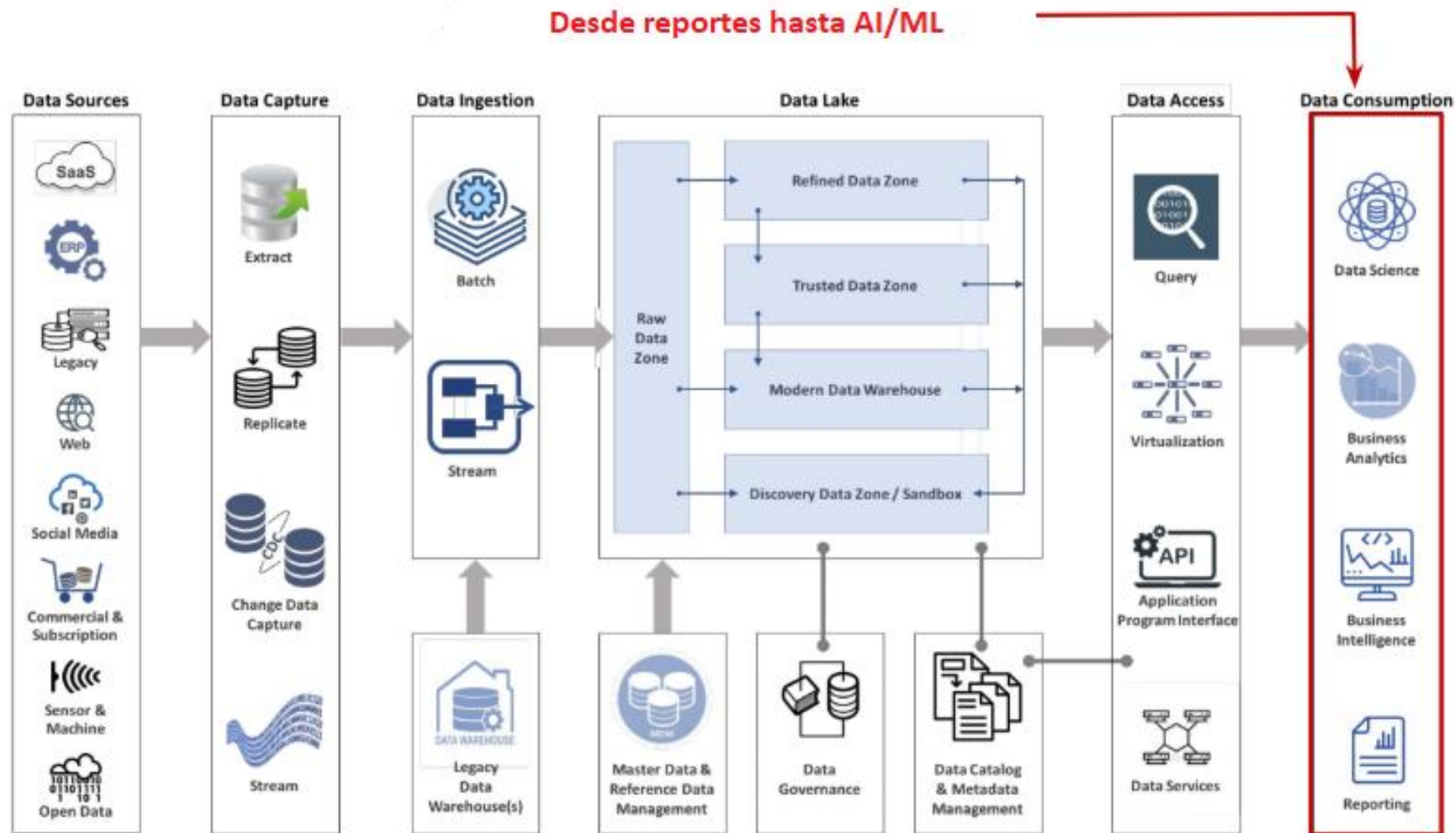
La rueda de DAMA



Arquitectura de integración de datos moderna

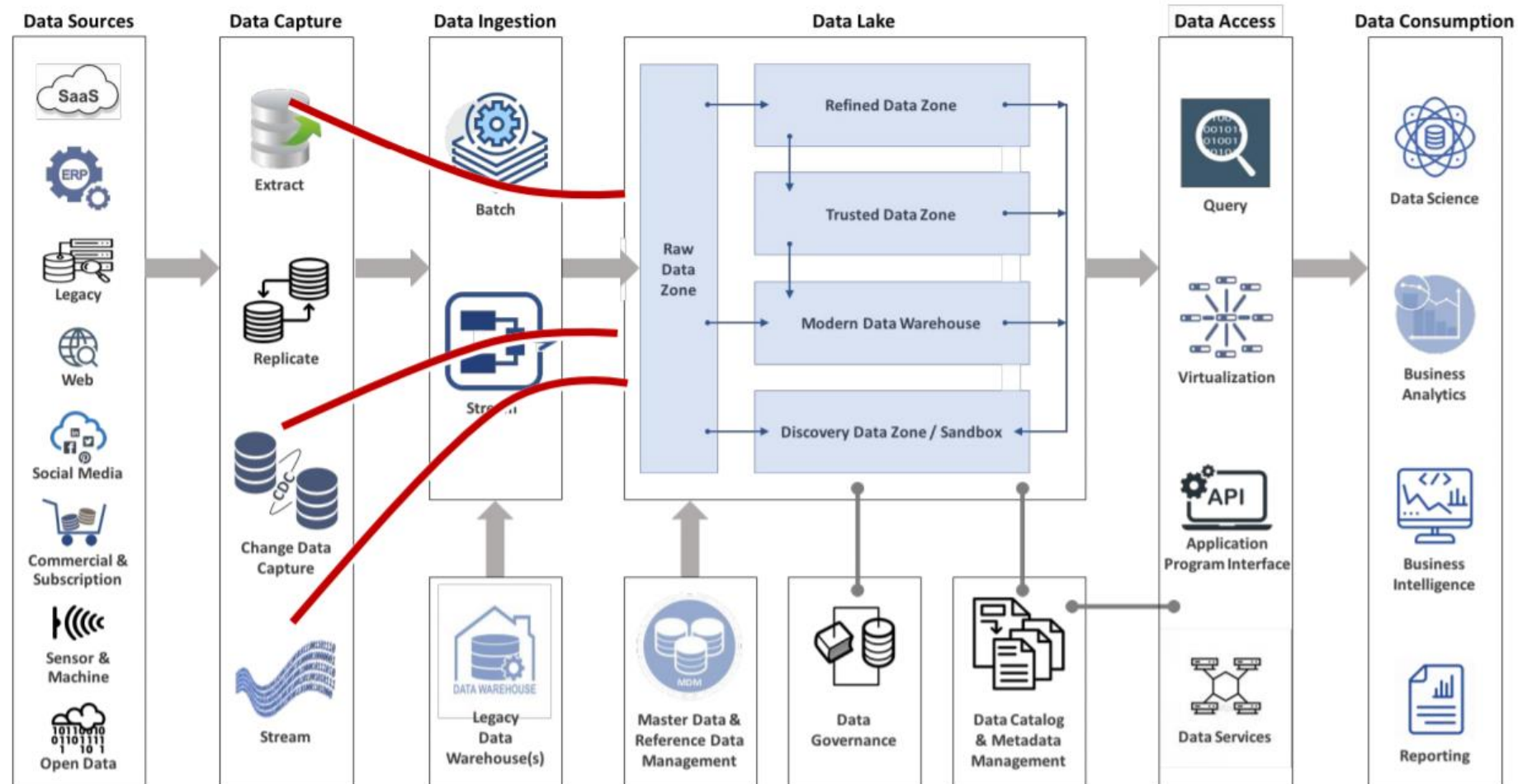


Soporta todos los usos de datos

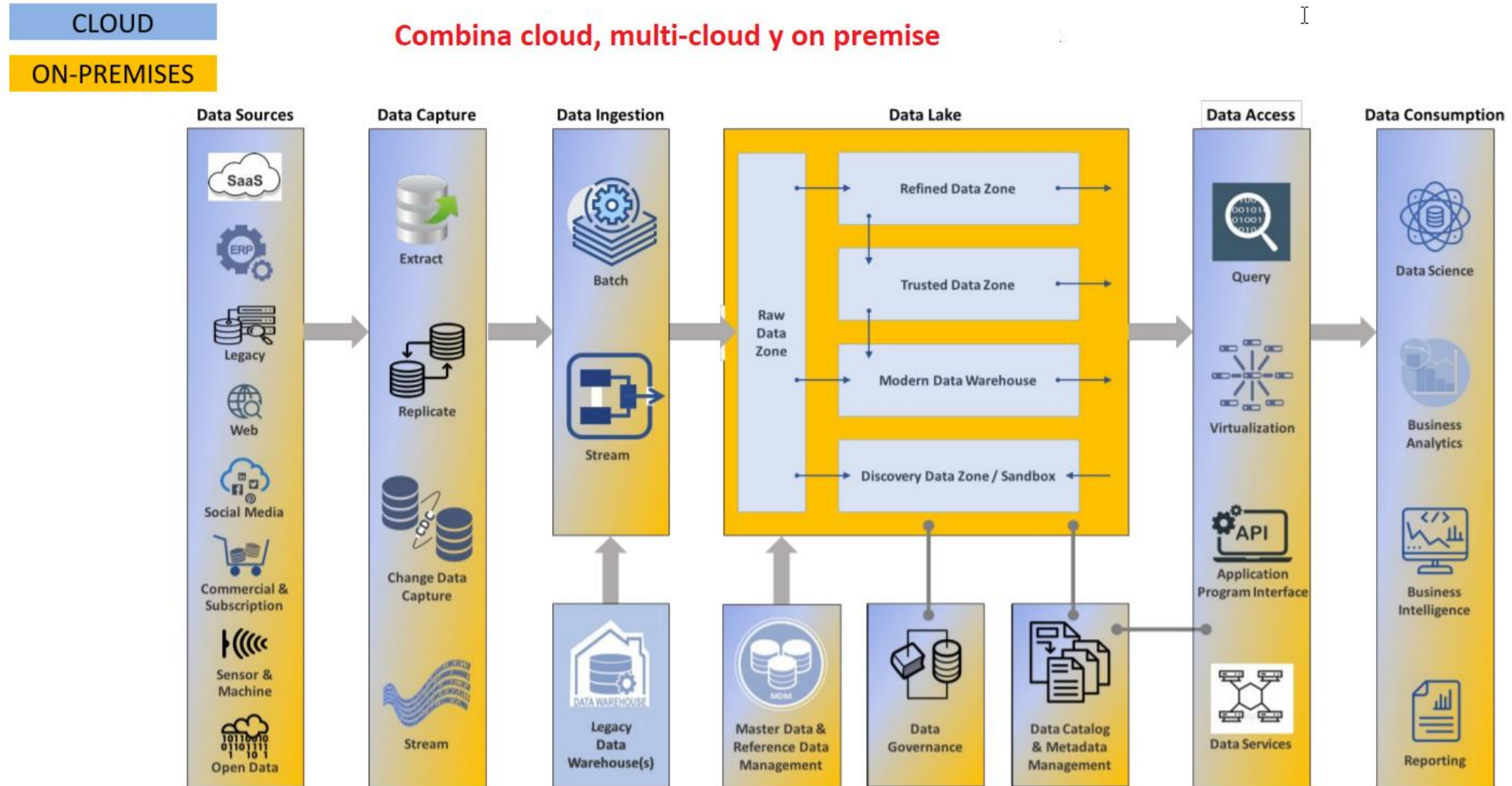


Soporta todos las latencias de datos

De procesamiento en batch a streaming

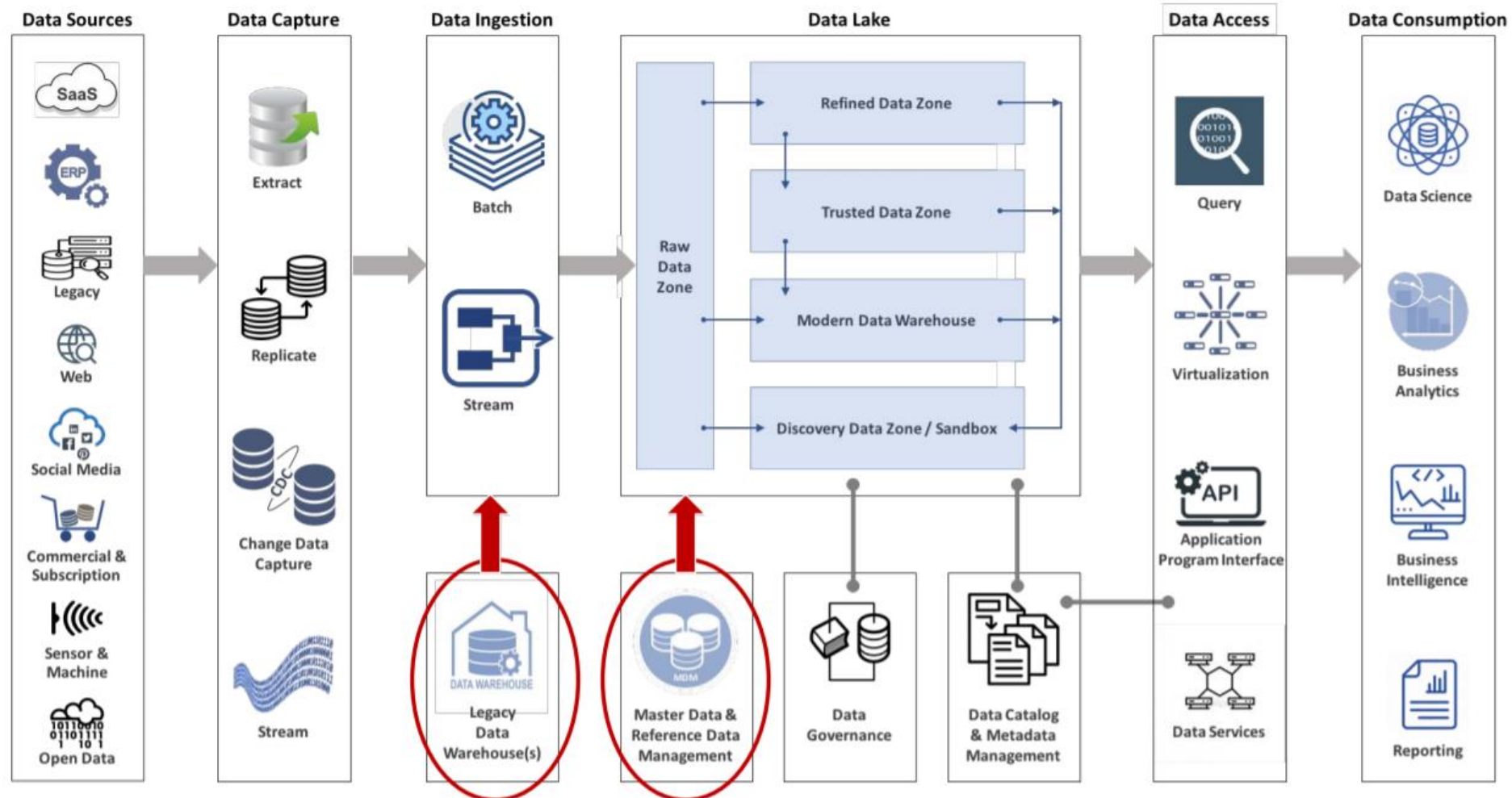


Soporta ecosistemas híbridos de datos



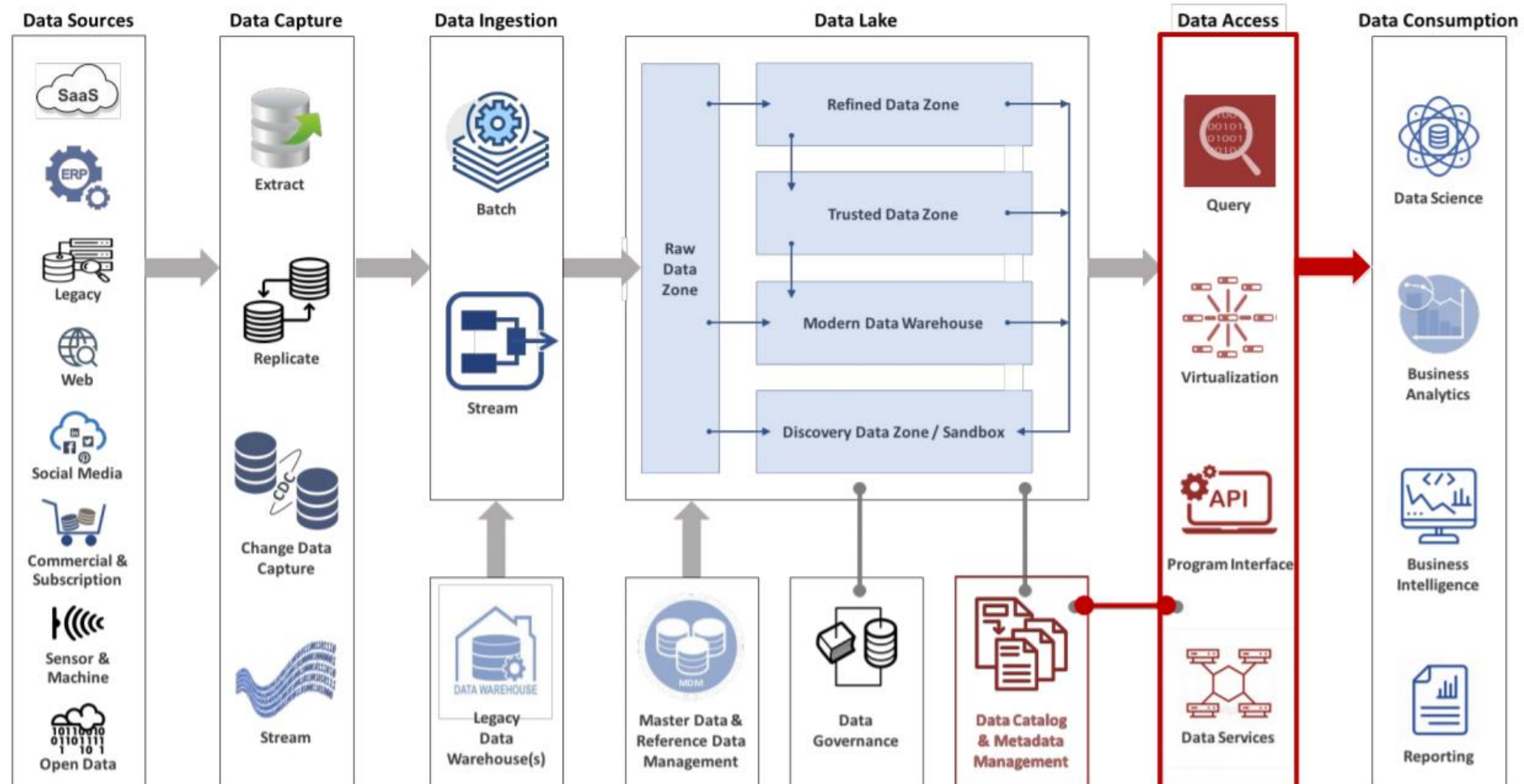
Soporta ecosistemas híbridos de datos

Integra almacenes antiguos



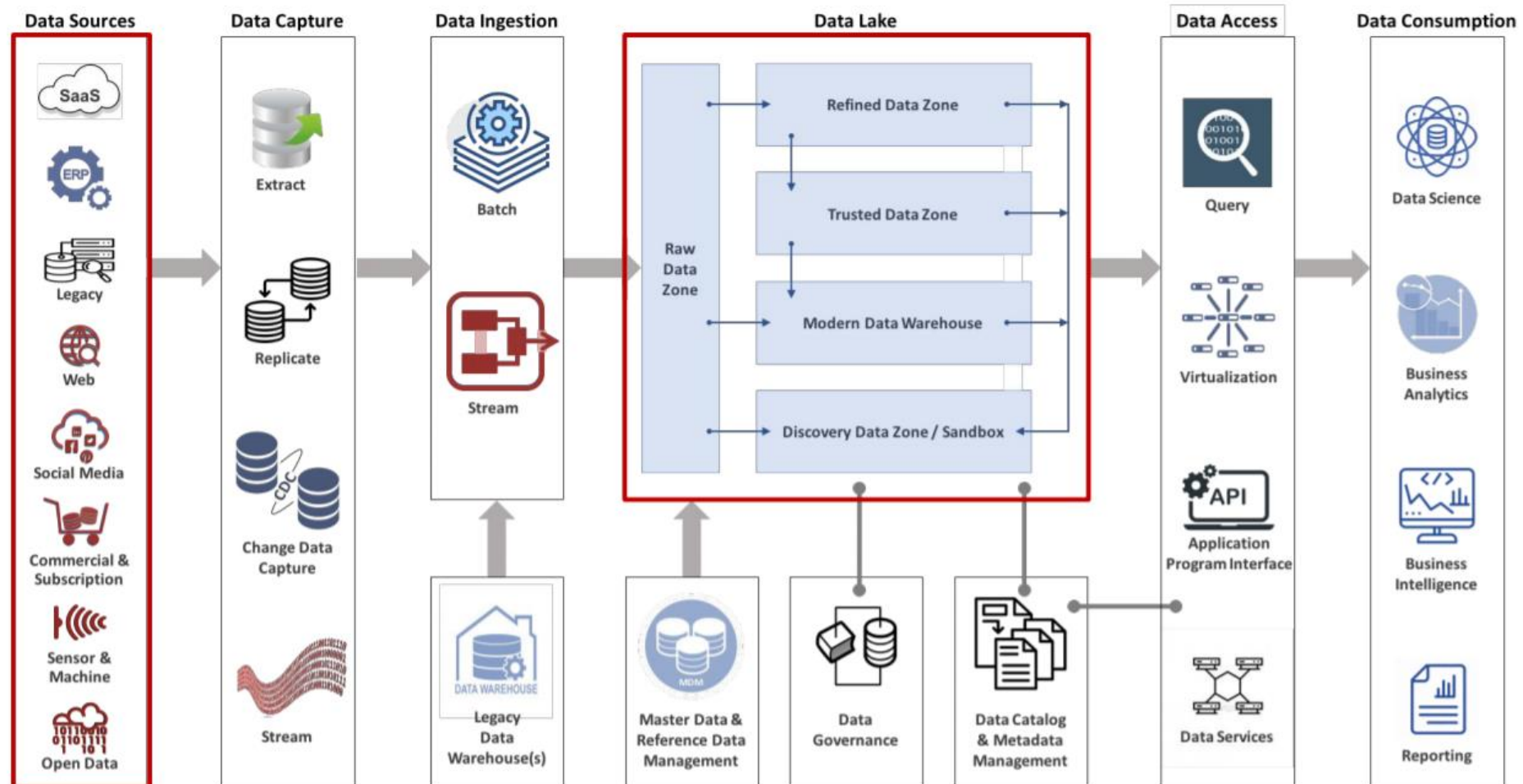
Selfservice

Fácil acceso para todos los consumidores



Big Data

Capturar, procesar y almacenar todo tipo de datos



Sostenimiento. Data-ops y monitoreo

