Table of Contents

[Análisis preliminar 3](#_Toc183991043)

[Arquitectura sugerida 3](#_Toc183991044)

[Siguientes pasos: 3](#_Toc183991045)

[Esquema detallado de la arquitectura para implementar una arquitectura de Data Lakehouse que gestione los datos: 4](#_Toc183991046)

[Arquitectura propuesta 4](#_Toc183991047)

[Data Lake: 4](#_Toc183991048)

[ETL (Extract, Transform, Load): 4](#_Toc183991049)

[Data Warehouse: 4](#_Toc183991050)

[Herramientas de BI: 5](#_Toc183991051)

[Flujo ETL sugerido 5](#_Toc183991052)

[Esquema de datos resultante 5](#_Toc183991053)

[Dimensiones: 5](#_Toc183991054)

[Hechos: 6](#_Toc183991055)

[Ejemplo ETL en SQL que unifica datos de diferentes plataformas 6](#_Toc183991056)

[Paso 1: Crear una tabla de dimensión unificada 6](#_Toc183991057)

[Paso 2: Crear una tabla de hechos unificada 7](#_Toc183991058)

[Paso 3: Asegurar integridad referencial 8](#_Toc183991059)

[Diagrama visual que representa la arquitectura de un sistema Data Lakehouse para manejar las métricas de redes sociales 8](#_Toc183991060)

El archivo dim\_account.csv contiene información estructurada sobre cuentas con los siguientes campos:

* Account\_ID: Identificador único.
* NAME: Nombre de la cuenta.
* COUNTRY: País asociado (algunos valores son "Unknown").
* TOPIC\_OF\_INFLUENCE: Categoría o tema de influencia.
* PLATFORM: Plataforma asociada (e.g., "THREADS").

El archivo fact\_social\_metrics.csv es una tabla de hechos que complementa a dim\_account.csv, con las siguientes columnas:

* Metric\_ID: Identificador único de la métrica.
* Account\_ID: Relaciona con el ID de cuenta en dim\_account.csv.
* NAME y PLATFORM: Detalles de la cuenta y la plataforma.
* FOLLOWERS: Número de seguidores.
* POTENTIAL\_REACH: Alcance potencial.
* ENGAGEMENT: Métrica de interacción.
* ER (Engagement Rate): Tasa de interacción.

El archivo THREADS.csv duplica, en parte, información de dim\_account.csv y fact\_social\_metrics.csv, pero con un formato diferente. Contiene:

* NAME, COUNTRY, y TOPIC OF INFLUENCE: Información básica de la cuenta.
* FOLLOWERS, POTENTIAL REACH, ER: Métricas de alcance y participación.

Hay pequeñas diferencias en los nombres de columnas y unidades (e.g., 4.9M para seguidores en lugar de un número absoluto).

El archivo INSTAGRAM.csv contiene datos sobre cuentas de Instagram con los siguientes campos:

* NAME: Nombre de la cuenta.
* FOLLOWERS: Seguidores (en millones, e.g., 625.1M).
* ER: Tasa de interacción (puede incluir valores faltantes como "-").
* COUNTRY: País asociado.
* TOPIC OF INFLUENCE: Tema o categoría de influencia.
* POTENTIAL REACH: Alcance potencial (en millones).

# Análisis preliminar

## Arquitectura sugerida

Data Lakehouse: Es la mejor opción, ya que los datos tienen una mezcla de estructura:

* Datos estructurados (dim\_account.csv y fact\_social\_metrics.csv).
* Datos parcialmente estructurados con formatos inconsistentes (THREADS.csv, INSTAGRAM.csv).
  + Puedes almacenar los archivos originales en el lago de datos y transformarlos para análisis.

## Siguientes pasos:

* Diseñar una arquitectura donde:
* El Data Lake almacene los datos originales (THREADS.csv, INSTAGRAM.csv).
* El Data Warehouse contenga datos limpios y estructurados para análisis BI.
  + Implementar un flujo ETL para unificar métricas similares entre plataformas.

# Esquema detallado de la arquitectura para implementar una arquitectura de Data Lakehouse que gestione los datos:

## Arquitectura propuesta

### Data Lake:

* Propósito: Almacena los datos originales sin procesar.
* Contenido:
  + Archivos crudos como THREADS.csv, INSTAGRAM.csv, dim\_account.csv, y fact\_social\_metrics.csv.
  + Estos datos se guardan en su formato original, permitiendo almacenamiento escalable y análisis flexible.

### ETL (Extract, Transform, Load):

* Extracción:
  + Los datos se cargan desde el lago en bruto y se inspeccionan para identificar errores de formato, codificación, y valores nulos.
* Transformación:
  + Unificar nombres de columnas y formatos para integrar métricas similares entre plataformas.
  + Calcular métricas faltantes, como convertir M a números absolutos en FOLLOWERS.
* Carga:
* Los datos limpios se cargan en un Data Warehouse.

### Data Warehouse:

* Propósito: Optimizado para consultas y análisis.
* Tablas:
  + - Tabla de dimensión: dim\_account (Información de las cuentas).
    - Tabla de hechos: fact\_social\_metrics (Métricas clave como seguidores, alcance, engagement).
* Modelo: Estrella o copo de nieve, dependiendo del nivel de detalle requerido.

### Herramientas de BI:

* Conecta el Data Warehouse a herramientas como Power BI o Tableau para análisis visual.

# Flujo ETL sugerido

* Unificación de columnas:
  + FOLLOWERS de INSTAGRAM.csv y THREADS.csv deben estandarizarse a números absolutos.
  + ER debe estandarizarse a porcentajes.
* Transformación de datos faltantes:
  + Reemplazar valores faltantes en COUNTRY o ER con categorías como Unknown.
* Creación de métricas adicionales:
  + Combinar POTENTIAL REACH y FOLLOWERS para calcular métricas de impacto agregado.

# Esquema de datos resultante

## Dimensiones:

1. dim\_account:
   * Account\_ID: Identificador único.
   * NAME: Nombre de la cuenta.
   * COUNTRY: País asociado.
   * TOPIC\_OF\_INFLUENCE: Categoría o tema.

## Hechos:

1. fact\_social\_metrics:

* Metric\_ID: Identificador único.
* Account\_ID: Relación con dim\_account.
* PLATFORM: Plataforma (e.g., THREADS, INSTAGRAM).
* FOLLOWERS: Seguidores.
* POTENTIAL\_REACH: Alcance potencial.
* ENGAGEMENT: Interacciones totales.
* ER: Tasa de interacción

# Ejemplo ETL en SQL que unifica datos de diferentes plataformas

## Paso 1: Crear una tabla de dimensión unificada

CREATE TABLE dim\_account AS

SELECT DISTINCT

Account\_ID,

NAME,

COALESCE(COUNTRY, 'Unknown') AS COUNTRY,

TOPIC\_OF\_INFLUENCE

FROM (

SELECT Account\_ID, NAME, COUNTRY, TOPIC\_OF\_INFLUENCE

FROM dim\_account\_raw

UNION ALL

SELECT NULL AS Account\_ID, NAME, COUNTRY, TOPIC\_OF\_INFLUENCE

FROM threads\_raw

UNION ALL

SELECT NULL AS Account\_ID, NAME, COUNTRY, TOPIC\_OF\_INFLUENCE

FROM instagram\_raw

);

## Paso 2: Crear una tabla de hechos unificada

CREATE TABLE fact\_social\_metrics AS

SELECT

ROW\_NUMBER() OVER () AS Metric\_ID,

Account\_ID,

PLATFORM,

CAST(REPLACE(FOLLOWERS, 'M', '') AS FLOAT) \* 1000000 AS FOLLOWERS,

CAST(REPLACE(POTENTIAL\_REACH, 'M', '') AS FLOAT) \* 1000000 AS POTENTIAL\_REACH,

ENGAGEMENT,

CAST(REPLACE(ER, '%', '') AS FLOAT) / 100 AS ER

FROM (

SELECT Account\_ID, PLATFORM, FOLLOWERS, POTENTIAL\_REACH, ENGAGEMENT, ER

FROM fact\_social\_metrics\_raw

UNION ALL

SELECT NULL AS Account\_ID, 'THREADS' AS PLATFORM, FOLLOWERS, POTENTIAL\_REACH, NULL AS ENGAGEMENT, ER

FROM threads\_raw

UNION ALL

SELECT NULL AS Account\_ID, 'INSTAGRAM' AS PLATFORM, FOLLOWERS, POTENTIAL\_REACH, NULL AS ENGAGEMENT, ER

FROM instagram\_raw

);

## Paso 3: Asegurar integridad referencial

ALTER TABLE dim\_account ADD CONSTRAINT pk\_account PRIMARY KEY (Account\_ID);

ALTER TABLE fact\_social\_metrics ADD CONSTRAINT fk\_account FOREIGN KEY (Account\_ID) REFERENCES dim\_account(Account\_ID);

# Diagrama visual que representa la arquitectura de un sistema Data Lakehouse para manejar las métricas de redes sociales

A diagram of data storage

Description automatically generated

Incluye:

1. Data Lake: Almacenamiento de archivos crudos como Threads e Instagram.
2. ETL: Proceso de transformación y limpieza.
3. Data Warehouse: Base de datos optimizada para consultas.
4. BI Tools: Herramientas de análisis.