Proyecto ETL Master Ciencia de Datos

**SQL**

CUNEF

Huerta Sánchez, Carmen

Querejeta Apalategui, Carmen

Olmedilla Bahillo, Nuria

[FASE I: ELECCIÓN DE DATOS 2](#_Toc183985627)

[THREADS.csv 2](#_Toc183985628)

[INSTAGRAM.csv 2](#_Toc183985629)

[FASE II: DOCUMENTACIÓN 4](#_Toc183985630)

[1. Objetivos del Proyecto 4](#_Toc183985631)

[Propósito del proyecto 4](#_Toc183985632)

[Objetivos Específicos 4](#_Toc183985633)

[Justificación de la importancia de realizar un proceso ETL. 4](#_Toc183985634)

[2. Dataset 5](#_Toc183985635)

[Enlace a los datasets utilizados 5](#_Toc183985636)

[Explicación de las fuentes de datos 5](#_Toc183985637)

[Frecuencia de actualización de los datos 7](#_Toc183985638)

[3. Características de los datos 7](#_Toc183985639)

[Descripción de los tipos de datos manejados 7](#_Toc183985640)

[Definición del Modelo de Datos 8](#_Toc183985641)

[4. Calidad de los Datos 8](#_Toc183985642)

[Evaluación de la calidad de los datos 8](#_Toc183985643)

[5. Limpieza de los Datos 10](#_Toc183985644)

[Fase de Limpieza de Datos - Resumen 10](#_Toc183985645)

[6. Problemas y Próximos Pasos 12](#_Toc183985646)

[Problemas 12](#_Toc183985647)

[Próximos Pasos 13](#_Toc183985648)

# FASE I: ELECCIÓN DE DATOS

## THREADS.csv

* Contenido: Este dataset contiene información sobre los usuarios con más seguidores, en la plataforma Threads, con los siguientes campos clave:
  + Name: Nombre del usuario
  + Followers: Número de seguidores
  + ER: Engagement Rate
  + Country: País de origen
  + Topic of influence: Tema de influencia principal de la cuenta
  + Potential Reach: Alcance potencial de la cuenta
* Tamaño: El dataset contiene 100 filas y 6 columnas.
* Uso: La tabla THREADS es una fuente rica en información sobre cuentas de redes sociales en la plataforma Threads, y puede ser utilizada para diversos tipos de análisis y aplicaciones, especialmente en el contexto de análisis de influencers y estudios de mercado.

## INSTAGRAM.csv

* Contenido: La tabla INSTAGRAM contiene datos sobre cuentas de Instagram, con información como el número de seguidores (FOLLOWERS), la tasa de interacción (ER), el alcance potencial (POTENTIAL\_REACH), el país (COUNTRY) y el tema de influencia (TOPIC\_OF\_INFLUENCE).
  + NAME: Nombre de la cuenta en Instagram.
  + FOLLOWERS: Número de seguidores de la cuenta.
  + ER: Tasa de interacción en porcentaje.
  + COUNTRY: País asociado a la cuenta.
  + TOPIC\_OF\_INFLUENCE: Tema principal de influencia (como Moda o Deportes).
  + POTENTIAL\_REACH: Alcance estimado en número de personas.
* Tamaño: El dataset contiene 100 filas y 6 columnas.
* Uso: Es útil para analizar influencers, segmentar audiencias por región o temática, y comparar métricas con otras plataformas.

# FASE II: DOCUMENTACIÓN

## 1. Objetivos del Proyecto

### Propósito del proyecto

Desarrollar un modelo ETL que permita integrar y analizar datos de redes sociales (Threads e Instagram) para identificar patrones, evaluar el impacto de influencers y apoyar estrategias de marketing.

### Objetivos Específicos

1. Extracción de Datos: Obtener y estructurar información relevante de las tablas THREADS e INSTAGRAM.
2. Transformación: Limpiar, estandarizar y enriquecer los datos para garantizar su calidad y consistencia.
3. Carga y Modelado: Crear un modelo dimensional que facilite el análisis de métricas clave y su relación con atributos descriptivos.
4. Análisis Comparativo: Evaluar y comparar métricas de rendimiento entre ambas plataformas.
5. Aplicación Práctica: Generar insights sobre tendencias, segmentos y oportunidades de mejora en estrategias de marketing digital.

### Justificación de la importancia de realizar un proceso ETL.

Un proceso ETL es fundamental para garantizar la integración confiable de datos provenientes de múltiples plataformas. La extracción, transformación y carga de los datos permitirá:

* Optimizar la toma de decisiones mediante información limpia y estructurada.
* Reducir errores y redundancias en el análisis.
* Proveer una base sólida para evaluar el impacto de la interacción multicanal en estrategias de contenido.

## 2. Dataset

### Enlace a los datasets utilizados

* THREADS.csv: [kaggle.com/socialMediaInfluencers](https://www.kaggle.com/datasets/bhavyadhingra00020/top-100-social-media-influencers-2024-countrywise)
* INSTAGRAM.csv: [kaggle.com/socialMediaInfluencers](https://www.kaggle.com/datasets/bhavyadhingra00020/top-100-social-media-influencers-2024-countrywise)

### Explicación de las fuentes de datos

Las fuentes de datos THREADS e INSTAGRAM contienen información sobre cuentas de redes sociales, incluyendo seguidores, tasas de interacción, alcance potencial, país y temas de influencia, específicos para cada plataforma

Data Catalog / Data Dictionary

THREADS

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Description |
| NAME | VARCHAR | Nombre de la cuenta en Threads. Identifica de forma única a cada usuario dentro de esta tabla. |
| FOLLOWERS | DECIMAL | Número de seguidores de la cuenta, usualmente en formato como 1.5M o 20K. |
| ER | DECIMAL | Tasa de interacción de la cuenta, expresada como porcentaje (3.5%). |
| COUNTRY | VARCHAR | País asociado con la cuenta. Puede contener valores como Unknown si el país no está especificado. |
| |  | | --- | | TOPIC\_OF\_INFLUENCE |  |  | | --- | |  | | VARCHAR | Tema de influencia principal de la cuenta (e.g.,"Moda", "Tecnología") |
| POTENTIAL\_REACH | DECIMAL | Número estimado de personas alcanzadas, en formato como 1M o 500K. |

INSTAGRAM

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Description |
| NAME | VARCHAR | Nombre de la cuenta en Instagram. Identifica de forma única a cada usuario dentro de esta tabla. |
| FOLLOWERS | DECIMAL | Número de seguidores de la cuenta, usualmente en formato como 1.5M o 20K. |
| ER | DECIMAL | Tasa de interacción de la cuenta, expresada como porcentaje (3.5%) |
| COUNTRY | VARCHAR | País asociado con la cuenta. Puede contener valores como Unknown si el país no está especificado. |
| TOPIC\_OF\_INFLUENCE | VARCHAR | Tema de influencia principal de la cuenta (e.g., "Moda", "Tecnología"). |
| POTENTIAL\_REACH | DECIMAL | Número estimado de personas alcanzadas, en formato como 1M o 500K. |

### Frecuencia de actualización de los datos

Ambos datasets son estáticos y reflejan una captura puntual del momento en que fueron recopilados. No cuentan con actualizaciones periódicas.

## 3. Características de los datos

### Descripción de los tipos de datos manejados

Los datos utilizados en este proyecto provienen de dos datasets principales: THREADS.csv e INSTAGRAM.csv. Cada dataset contiene información estructurada sobre cuentas en plataformas de redes sociales.

Columnas clave y tipos de datos:

i. NAME (VARCHAR): Nombre de la cuenta. Es el identificador único dentro de cada dataset.

ii. FOLLOWERS (DECIMAL): Número de seguidores de la cuenta, convertido de cadenas de texto (e.g., "1.5M", "20K") a valores decimales para cálculos precisos.

iii. ER (Engagement Rate, DECIMAL): Tasa de interacción, transformada de porcentajes (e.g., "3.5%") a valores decimales (e.g., 0.035) para análisis cuantitativo.

iv. COUNTRY (VARCHAR): País asociado con cada cuenta, que puede incluir valores "Unknown" si no está especificado.

v. TOPIC\_OF\_INFLUENCE (VARCHAR): Tema principal de influencia de la cuenta, categorizado en áreas como "Moda", "Tecnología", etc.

vi. POTENTIAL\_REACH (DECIMAL): Alcance potencial estimado, estandarizado a valores numéricos desde formatos como "1M" o "500K".

Propiedades de los datos manejados:

* Los datos son semi-estructurados y requieren transformaciones para garantizar su calidad y consistencia.
* Incluyen tanto valores categóricos (e.g., COUNTRY, TOPIC\_OF\_INFLUENCE) como métricas numéricas (e.g., FOLLOWERS, ER, POTENTIAL\_REACH).
* En su estado original, presentan valores faltantes y formatos heterogéneos que se resolvieron durante la fase de transformación.

Los datos preparados y estandarizados permiten un análisis más profundo sobre el impacto y la segmentación de los influencers en redes sociales, así como la comparación de métricas entre las plataformas THREADS e INSTAGRAM.

### Definición del Modelo de Datos

NAME:

* Es el identificador principal de la cuenta.
* Se mantiene como VARCHAR para admitir nombres alfanuméricos.

FOLLOWERS y POTENTIAL\_REACH:

* Convertidos a DECIMAL para permitir cálculos precisos.
* Usa una precisión alta (DECIMAL (12, 2)) para manejar valores grandes (e.g., millones).

ER (Engagement Rate):

* Almacenado como un valor decimal entre 0 y 1 (e.g., 0.025 para 2.5%).
* Usa DECIMAL (5, 4) para alta precisión.

COUNTRY y TOPIC\_OF\_INFLUENCE:

* Almacenados como VARCHAR para admitir texto descriptivo.

## 4. Calidad de los Datos

### Evaluación de la calidad de los datos

La calidad de los datos se evalúa considerando métricas clave que aseguran que los datos sean útiles, confiables y adecuados para su análisis. A continuación, se presenta una evaluación de calidad para las tablas THREADS e INSTAGRAM, basada en las siguientes dimensiones:

Dimensiones de Evaluación

Precisión: Evalúa si los datos son exactos y cumplen con los formatos esperados.

* Ejemplo:
  + FOLLOWERS y POTENTIAL\_REACH deben estar en formato numérico o contener sufijos válidos (K, M).
  + ER debe ser un porcentaje dentro del rango de 0% a 100%.

Estructura: Verifica que los datos estén correctamente organizados y cumplan con los tipos de datos definidos.

* Ejemplo:
  + Columnas críticas como NAME, FOLLOWERS, y ER no deben contener valores nulos o vacíos.
  + Las columnas deben seguir los formatos predefinidos para números y porcentajes.

Completitud: Analiza el porcentaje de datos presentes en comparación con el total esperado.

* Ejemplo:
  + Identificación de valores nulos o vacíos en columnas clave como NAME, COUNTRY, y TOPIC\_OF\_INFLUENCE.

Consistencia: Comprueba la coherencia entre datos relacionados dentro y entre tablas.

* Ejemplo:
  + Los valores de FOLLOWERS y ER deben coincidir entre THREADS e INSTAGRAM para cuentas con el mismo NAME.
  + Se permite una pequeña tolerancia en los valores numéricos (e.g., 5% de diferencia).

Semántica: Valida si los datos se interpretan correctamente según su contexto.

* Ejemplo:
  + COUNTRY debe contener nombres de países válidos.
  + FOLLOWERS y POTENTIAL\_REACH deben representar valores lógicos y realistas.

Razonabilidad: Evalúa si los valores tienen sentido lógico y cumplen con reglas definidas.

* Ejemplo:
  + FOLLOWERS no debe ser negativo y debe estar en un rango razonable (< 1,000 millones).
  + ER debe reflejar tasas lógicas para cuentas de redes sociales (e.g., generalmente < 10%).

Problemas comunes entre los datasets:

Valores nulos: Hemos encontrado valores nulos en las columnas de COUNTRY, ER y TOPIC OF INFLUENCE

## 5. Limpieza de los Datos

### Fase de Limpieza de Datos - Resumen

La limpieza y transformación de los datos ha incluido varias actividades organizadas en pasos claros para garantizar la calidad, consistencia y usabilidad de las tablas THREADS e INSTAGRAM. A continuación, se presenta un resumen ordenado de las acciones realizadas:

#### i. Evaluación de Calidad

Se han analizado las tablas para determinar problemas de calidad en las siguientes dimensiones:

* Precisión: Validación de valores y formatos correctos.
* Estructura: Verificación de tipos de datos y columnas no nulas.
* Completitud: Análisis de valores nulos o vacíos.
* Consistencia: Comparación entre columnas similares en ambas tablas.
* Semántica: Verificación de que los datos cumplen con los estándares establecidos.
* Razonabilidad: Confirmación de que los valores son lógicos y están dentro de rangos aceptables.

#### ii. Transformación de Datos

##### ii.i Estandarización

* Conversión de Formatos: Se han transformado las columnas FOLLOWERS y POTENTIAL\_REACH de texto con sufijos (M, K) a valores numéricos.
* Normalización de Porcentajes: La columna ER se ha normalizado convirtiendo los porcentajes (e.g., 2.5%) a valores decimales (0.025).

##### ii.ii Manejo de Valores Nulos

* Los valores nulos o vacíos en la columna COUNTRY y TOPIC\_OF\_INFLUENCE se han reemplazado con 'Unknown'.

##### ii.iii Enriquecimiento de Datos

* Cálculo de ENGAGEMENT: Se ha añadido una nueva columna para calcular la interacción diaria de las cuentas como ENGAGEMENT = FOLLOWERS \* ER.
* Clasificación de Cuentas: Se ha agregado una columna SIZE\_CATEGORY para clasificar las cuentas en categorías (Micro, Small, Medium, Large) según el número de seguidores.

#### iii. Consistencia entre Tablas

* Se han verificado y comparado los pares de registros entre las tablas THREADS e INSTAGRAM con los siguientes criterios:
  + Diferencia de seguidores (FOLLOWERS) no mayor a 10,000.
  + Diferencia en la tasa de interacción (ER) no mayor a 0.01.
* Se ha añadido una columna MATCH\_WITH\_INSTAGRAM en THREADS para indicar si las cuentas coinciden en ambas plataformas.

#### iv. Validaciones Clave Realizadas

##### i. Verificación de Formatos:

* Se ha validado que FOLLOWERS y POTENTIAL\_REACH cumplan con los formatos esperados (e.g., números o sufijos M, K).
* Se ha comprobado que ER esté en formato porcentaje y dentro del rango lógico [0%, 100%].

##### ii. Valores Nulos:

* Se han identificado y contabilizado los registros nulos en columnas críticas como NAME, COUNTRY, FOLLOWERS, POTENTIAL\_REACH, y ER.

##### iii. Consistencia de Datos:

* Comparación entre tablas para NAME, FOLLOWERS, ER, y COUNTRY para asegurar concordancia.

##### iv. Semántica:

* Validación de que los países pertenezcan a una lista definida y los valores de FOLLOWERS y ER sigan reglas lógicas.

##### v. Razonabilidad:

* Confirmación de que los valores de FOLLOWERS no son negativos y están dentro de un rango aceptable (<1B).

## 6. Problemas y Próximos Pasos

### Problemas

Durante la evaluación y limpieza de los datos en las tablas THREADS e INSTAGRAM, se han identificado los siguientes problemas:

#### i. Valores Nulos o Vacíos:

* Columnas clave como COUNTRY y TOPIC\_OF\_INFLUENCE contenían valores nulos o vacíos, afectando el análisis geográfico y temático.

#### ii. Inconsistencias en Formatos:

* FOLLOWERS y POTENTIAL\_REACH estaban en formatos mixtos, incluyendo números enteros, decimales, y sufijos (M, K), lo que impedía el cálculo directo.
* ER estaba representado como porcentajes (3.5%) y requería conversión a valores decimales (0.035).

#### iii. Falta de Consistencia Entre Tablas:

* Para cuentas compartidas (NAME) entre THREADS e INSTAGRAM, los valores de FOLLOWERS, ER y POTENTIAL\_REACH no coincidían completamente, incluso considerando tolerancias razonables.

#### iv. Problemas Semánticos:

* Algunos valores en COUNTRY no correspondían a países válidos, afectando la segmentación geográfica.
* TOPIC\_OF\_INFLUENCE contenía valores categóricos ambiguos o inconsistentes.

### Próximos Pasos

Para mejorar aún más la calidad y utilidad de los datos, proponemos los siguientes pasos:

#### i. Validación y Corrección de Datos

* Implementar validaciones automáticas para identificar y corregir valores fuera de rango o formatos incorrectos.
* Refinar las reglas de consistencia entre tablas (THREADS e INSTAGRAM), especialmente para métricas como FOLLOWERS y ER.

#### ii. Enriquecimiento Adicional

* Incorporar datos externos para completar información faltante, como listas estandarizadas de países para validar la columna COUNTRY.
* Ampliar las categorías en TOPIC\_OF\_INFLUENCE con taxonomías más completas y precisas.

#### iii. Creación de Informes Automatizados

* Desarrollar reportes periódicos que analicen la calidad de los datos en dimensiones clave (e.g., completitud, consistencia).
* Generar alertas cuando se detecten problemas recurrentes, como valores nulos o inconsistencias entre tablas.

#### iv. Optimización del Modelo

* Convertir columnas estandarizadas (FOLLOWERS, ER, POTENTIAL\_REACH) a tipos de datos numéricos permanentes en la base de datos para mejorar el rendimiento.
* Consolidar las tablas THREADS e INSTAGRAM en un modelo dimensional para análisis más avanzado.