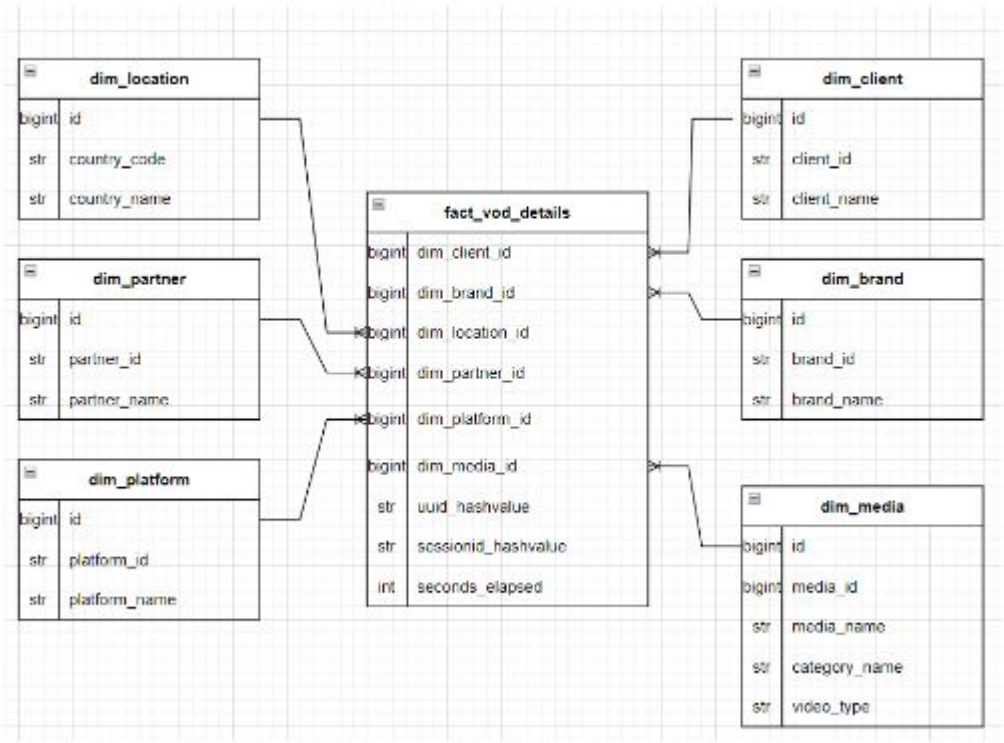
Introducción a la ciencia de datos - Tarea 3

Grupo 20 Auriello - Molina
11/07/2024

Datos


dim_location
bigint id
str country_code
str country_name
dim_partner
bigint id
str partner_id
str partner_name
dim_platform
bigint id
str platform_id
str platform_name
fact_vod_details
bigint dim_client_id
bigint dim_brand_id
bigint dim_location_id
bigint dim_partner_id
bigint dim_platform_id
bigint dim_media_id
str uuid_hashvalue
str sessionid_hashvalue
int seconds_elapsed
dim_client
bigint id
str client_id
str client_name
dim_brand
bigint id
str brand_id
str brand_name
dim_media
bigint id
bigint media_id
str media_name
str category_name
str video_type

Calidad de datos

Problemas resueltos:

- **Inconsistencia en las Categorías**
- **Datos Faltantes o Incompletos**
- **Duplicación de Datos**
- **Valores Nulos o Incorrectos en Llaves Foráneas**
- **Problemas de Integridad Referencial (**Falta de correspondencia entre las llaves primarias de las tablas dimensionales)
- **Formato de Datos Inconsistente (US vs USA)**

Problemática a resolver

Basándonos en sus patrones de visualización y características demográficas, surgen los siguientes problemas que pueden ser resueltos con datos:

- ¿Cómo podemos segmentar a los usuarios para mejorar la personalización de contenido y las estrategias de retención?
- ¿Podemos predecir el tiempo que permanecerán en la aplicación?
- ¿Es posible automatizar el proceso de mapeo entre las categorías estándar y las utilizadas por cada aplicación?

Soluciones propuestas

- **Segmentación de Clientes**: Utilizar métodos de aprendizaje no supervisado como K-means clustering para segmentar a los usuarios en grupos basados en sus patrones de visualización y características demográficas.
- **Predicción de duración de la sesión:** Utilizando Random Forest, este algoritmo se puede utilizar tanto para problemas de clasificación como de regresión, su faceta de regresión será la que nos permita a partir de las características del usuario estimar el tiempo que este permanecerá en la aplicación.
- **Automappeo de categorias**: En este caso es posible la utilización de medidas de distancia que junto con un algoritmo de predicción nos permita comparar el parecido entre nuestras categorías y el estándar, definiendo un threshold de certeza podemos automatizar que se remplacen aquellas categorías que tienen una distancia menos o un mayor parecido con las categorías del sistema de clasificación estándar.