Introducción a la Ciencia de Datos

Maestría en Ciencia de Datos y Aprendizaje Automático

Facultad de Ingeniería, UdelaR

11/07/2024

Tarea 3

Joana Auriello, Pablo Molina

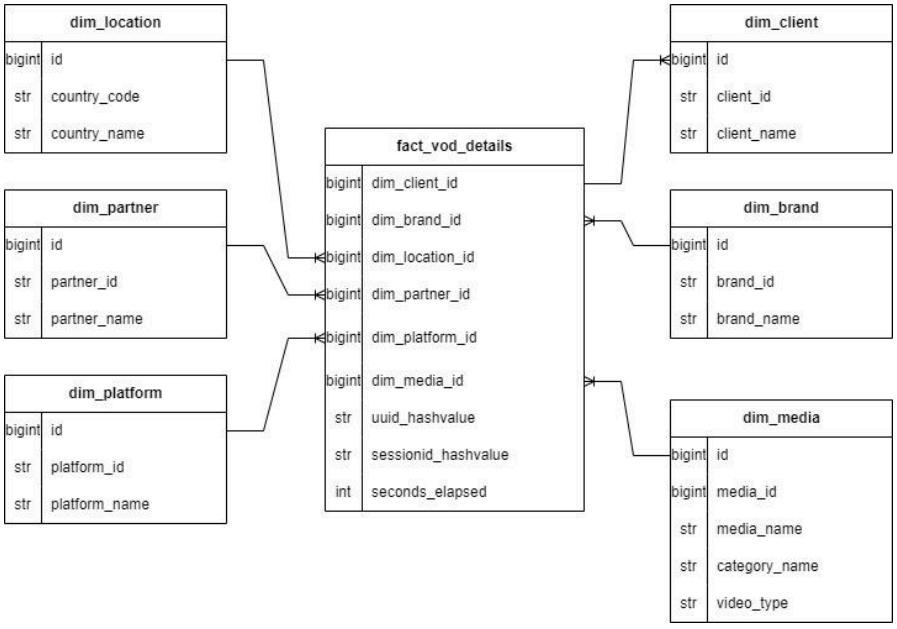
Datos: descripción y calidad

Los datos a analizar surgen de los logs de aplicaciones de servicios de streaming. De estos logs se extraen datos vinculados a los dispositivos, sistemas operativos, plataformas, ubicaciones geográficas, entre otros, de los usuarios que están consumiendo el contenido disponible en estas apps de streaming. Estos logs son procesados a través de pipelines ETL en la nube y se cargan a un data warehouse periódicamente.

Por razones de granularidad de los datos y privacidad, a pesar de que mantenemos una muestra representativa, en este trabajo se presenta solamente una parte del esquema completo con el que se trabaja usualmente.

A modo de ejemplo y para demostrar la dimensión de los datos utilizados, un solo día de datos representado a nivel de títulos individuales por usuario tiene aproximadamente más de 2 millones de registros.

Modelo de datos



En este modelo, la tabla fact_vod_details actúa como tabla de hechos y contiene los datos principales de visualización, incluyendo uuid_hashvalue(llave primaria), seconds_elapsed y sessionid_hashvalue. Las llaves foráneas en esta tabla (dim_client_id, dim_brand_id, dim_location_id, dim_partner_id, dim_platform_id, dim_media_id) se relacionan con las tablas dimensionales correspondientes. La tabla dim_client incluye id (llave primaria), client_id y client_name. La tabla dim_brand contiene id (llave primaria), brand_id y brand_name. La tabla dim_location tiene id (llave primaria), country_code y country_name. La tabla dim_partner incluye id (llave primaria), partner_id y partner_name. La tabla dim_platform contiene id (llave primaria), platform_id y platform_name. Finalmente, la tabla dim_media incluye id (llave primaria), media_id, media_name, category_name y video_type. Estas relaciones permiten la integración de datos detallados sobre clientes, marcas, ubicaciones, sistema operativo, plataformas y títulos

A partir de este modelo se genera un dataset que contiene las siguientes columnas:

- client_id, client_name, brand_id, brand_name, country_code, country_name, partner_id, partner_name, platform_id, platform_name, media_id, media_name, category_name, video_type, UUID_HASHVALUE, seconds_elapsed, sessions (número total de sesiones, definido como count distinct de sessionid_hashvalue).

Calidad

- **Inconsistencia en las Categorías:** Dado que cada cliente define sus propias categorías en su aplicación, es posible encontrar categorías similares con nombres diferentes, como "Comedia" y "Comedy". Esto puede dificultar el análisis agregado y la comparación entre usuarios.
- **Datos Faltantes o Incompletos:** Puede haber **registros con valores faltantes** en campos clave como client_name, brand_name, country_name, etc. Esto afectaría la precisión de los análisis y reportes.
- **Duplicación de Datos:** Podrían existir **duplicados** en la tabla fact_vod_uu_details o en las tablas dimensionales, lo que inflaría los resultados de conteos y sumas.
- **Valores Nulos o Incorrectos en Llaves Foráneas:** Las llaves foráneas en fact_vod_uu_details pueden tener valores nulos o incorrectos, impidiendo la correcta relación con las tablas dimensionales y afectando la integridad referencial.
- **Problemas de Integridad Referencial:** Falta de correspondencia entre las llaves primarias de las tablas dimensionales y las llaves foráneas en la tabla de hechos puede llevar a registros huérfanos o a la imposibilidad de realizar uniones correctas.
- **Formato de Datos Inconsistente:** Diferencias en el formato de datos, como fechas en diferentes formatos o códigos de país inconsistentes (USA vs US), pueden complicar la consolidación y análisis de datos.

Para abordar estos problemas, se podría implementar procesos de limpieza y estandarización de datos, como la normalización de categorías, la validación de llaves foráneas, la deduplicación de registros y la implementación de controles de calidad de datos durante la recolección y carga de datos.

Problemática a resolver con herramientas del curso

Problema/Pregunta a Resolver

¿Cómo podemos segmentar a los usuarios basándonos en sus patrones de visualización y características demográficas para mejorar la personalización de contenido y las estrategias de retención?

Descripción de la implementación de la solución

Métodos

Para resolver este problema, se puede seguir un proceso estructurado que involucre varias etapas de la ciencia de datos, aplicando conceptos y herramientas presentadas en el curso.

1. Recolección y Exploración de Datos

- **Recolección**: Usar la consulta proporcionada para recolectar datos relevantes de las tablas fact_vod_uu_details, dim_client, dim_brand, dim_location, dim_partner, dim_platform y dim_media.
- **Exploración**: Realizar un análisis exploratorio de datos (EDA) para entender las distribuciones, valores atípicos y patrones generales. Visualizar datos usando histogramas, gráficos de barras y matrices de correlación para identificar relaciones importantes entre variables.

2. Limpieza y Calidad de Datos

- **Limpieza**: Tratar los valores faltantes y duplicados. Normalizar las categorías de medios para evitar inconsistencias. Validar y corregir las llaves foráneas para mantener la integridad referencial.
- **Calidad de Datos**: Implementar controles de calidad para asegurar que los datos sean precisos, completos y consistentes.

3. Segmentación y Aprendizaje Automático

- **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.

4. Visualización e Interpretación de Resultados

- **Visualización de Resultados**: Crear gráficos de dispersión, diagramas de cajas y gráficos de radar para visualizar los segmentos de usuarios. Usar mapas de calor para visualizar la correlación entre las variables y los clusters formados.
- **Interpretación**: Interpretar los resultados para entender las características y comportamientos de cada segmento. Identificar patrones comunes y diferencias clave entre los segmentos.

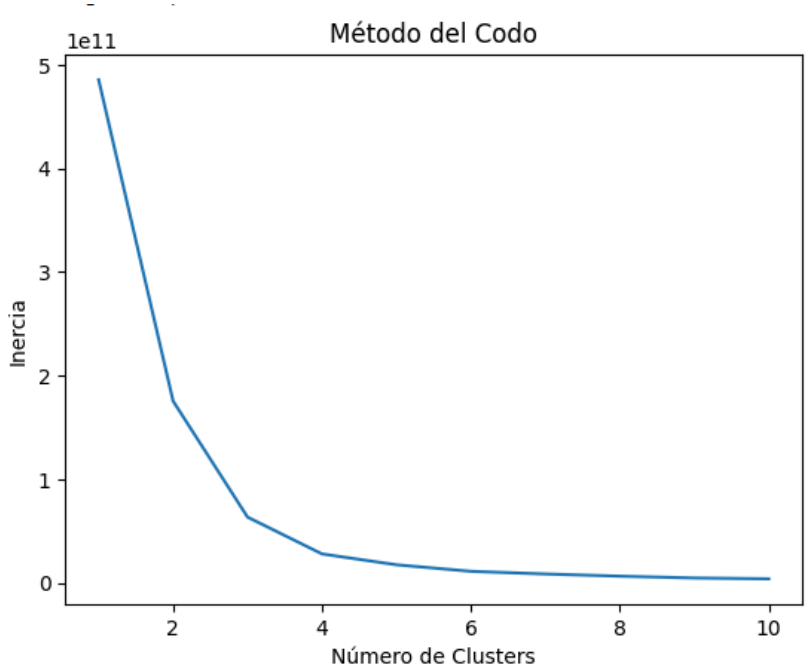
5. Aplicaciones Prácticas

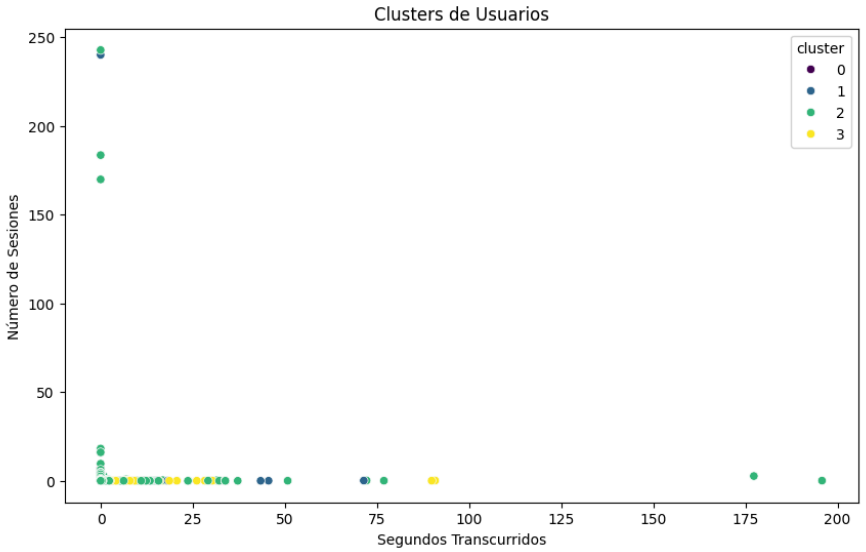
- **Personalización de Contenido**: Utilizar los segmentos de clientes para desarrollar recomendaciones personalizadas de contenido.

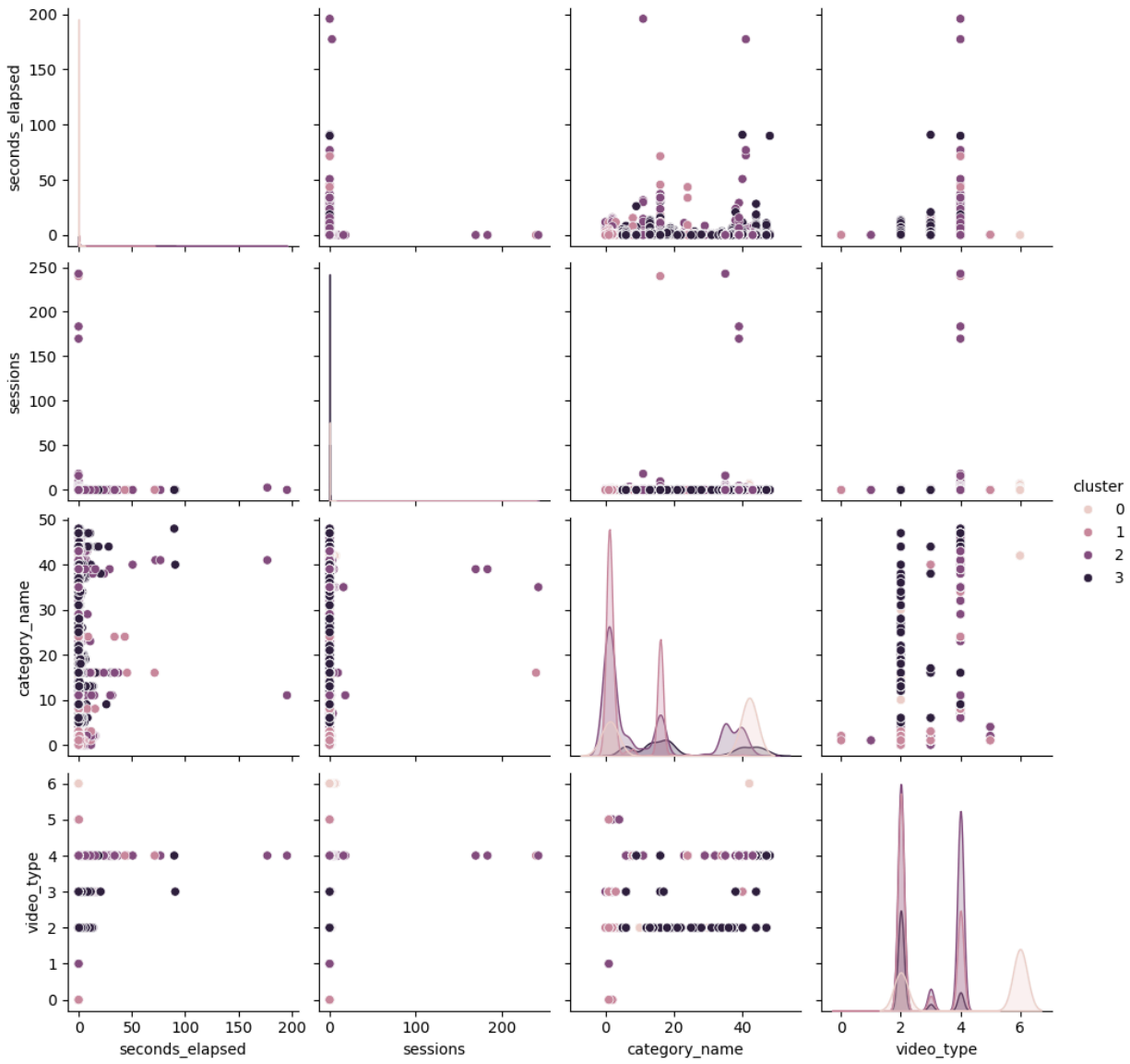
- **Estrategias de Retención**: Diseñar estrategias de retención específicas para cada segmento basado en sus características y comportamientos.

Implementación

Visualizaciones






seconds_elapsed
sessions
category_name
video_type
cluster
0
1
2
3