**Analítica de Grandes Datos**

**Departamento de Ciencias de la Computación y la Decisión**

**Facultad de Minas**

**Universidad Nacional de Colombia**

Trabajo Nociones de Arquitectura de la Información

Versión: 2021.05.23 05:00

*Observación: Cada vez que agregue nuevos elementos al documento, o que modifique algún componente del informe, revise la coherencia y consistencia con los otros elementos que hacen parte del mismo.*

**Responsables**

|  |
| --- |
| Nombre Completo – Documento de Identificación |
| 1. Andrea Stephanya Pino Villegas - **C.C 1.144.158.067** |
| 1. Daniela Rivera Vargas - **C.C. 1.037.636.542** |
| 1. Daniel Escobar Ruiz - **C.C. 1.040.747.009** |
| 1. Edward Darío Mercado Gómez - **C.C. 1.038.131.004** |
| 1. Simon Pineda Agudelo - **C.C. 1.152.220.029** |
| **REPO EN GITHUB: https://github.com/skailuxspa-unal/agd-trabajo-1** |

*Realiza este trabajo considerando los datos que generan los sistemas transaccionales e información no estructurada de tu dominio (si trabajas por ejemplo para TCC tu dominio es la mensajería; también puedes explorar en la página* [*https://www.kaggle.com/datasets*](https://www.kaggle.com/datasets) *o* [*https://arxiv.org/*](https://arxiv.org/)*). Considera tener acceso a esta información, de al menos 10 MB (puede ser uno o varios archivos de texto), y* ***tener al menos cuatro clases conceptuales. Este documento también debe almacenarse en el REPO. Plazo Máximo de Entrega 23 de Mayo, NO SE recibirá por correo electrónico, envío por*** [***https://forms.gle/h7ty3yZykaUq5m7y6***](https://forms.gle/h7ty3yZykaUq5m7y6)

1. **Comprensión del negocio**
   1. **Descripción del contexto del negocio.**

Describa en máximo 250 palabras el contexto en el cual se generan los datos y cuál es el proceso que los genera.

Somos una agencia de consultoría que ofrece soluciones tecnológicas basadas en el análisis de datos e información para la toma de decisiones inteligentes. En este proyecto en particular, vamos a estar apoyando a un cliente que se encuentra en la industria de la comercialización de artículos electrónicos. Este se encuentra interesado en realizar pautas publicitarias en la red social youtube con target en los Estados Unidos de América. Por lo tanto, los datos se obtienen de la plataforma de streaming de videos teniendo en cuenta el número de vistas, likes, dislikes, temática, comentarios y categoría del video. Los datos usados fueron extraídos en parte de Kaggle y fueron complementados utilizando la API v3 de YouTube.

Categorías: <https://www.kaggle.com/datasnaek/youtube-new?select=US_category_id.json>

Datos: <https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv>

* 1. **Identificación del problema:**

Delimite en máximo 150 palabras la problemática, así como identificar los requisitos, supuestos, restricciones y beneficios de la solución de este.

Tomando como referencia las restricciones presupuestales y recursos tecnológicos de nuestro clientes se debe identificar, cuales son los videos con mas reproducciones y sus respectivas categorías para definir cuales son los videos en donde las pautas publicitarias de nuestro cliente traerian mas beneficio a través de su canal de ventas e-commerce, introducción y posicionamiento de marca.

* 1. **Determinación de objetivos:**

Describa en máximo 150 palabras las metas a lograr al proponer una solución basada en un modelo de datos o de analítica (cómo y qué tipo de ventaja competitiva se ganará).

Mediante la estructuración, definición y estudio del modelo de datos de la presente base de datos se logrará ofrecer al cliente una clara estrategia publicitaria a través de la plataforma Youtube, que permitirá obtener mejores resultados. Estos a su vez y como finalidad, se traducirán en mayores ventas para nuestro cliente en específico.

* 1. **Evaluación de la situación actual:**

Describa en máximo 150 palabras el estado actual antes de implementar la solución de analítica, a fin de tener un punto de comparación que permita medir el grado de éxito de la solución.

Actualmente, contamos con una base de datos de 20MB. Los datos no se encuentran estructurados, ni tampoco se encuentran en un orden establecido. Por lo tanto, se desconoce cuál sería la mejor estrategia de publicidad visual en la plataforma youtube para nuestro cliente.

1. **Comprensión de los datos**
   1. **recolección de datos**

Describa en máximo 150 palabras los datos a utilizar identificando las fuentes, las técnicas empleadas en su recolección, los problemas encontrados en su obtención y la forma como se resolvieron los mismos. Además, adjunte los datos (archivos de texto, etc.) agréguelos en el github **(REPO EN GITHUB)** en un solo archivo, por favor comprímalo(s). Llame el archivo T1.2.1.Datos.zip

Los datos recopilados representan videos de YouTube que estuvieron en trending en 2018. Esto incluye títulos, categorías, número de likes, dislikes, entre otros. Dichos datos fueron obtenidos de Kaggle, sin embargo se decidió complementarlos utilizando el API de YouTube para descargar 20 comentarios por cada video, esto con el fin de entender, grosso modo, qué opiniones tiene la gente de cada video.

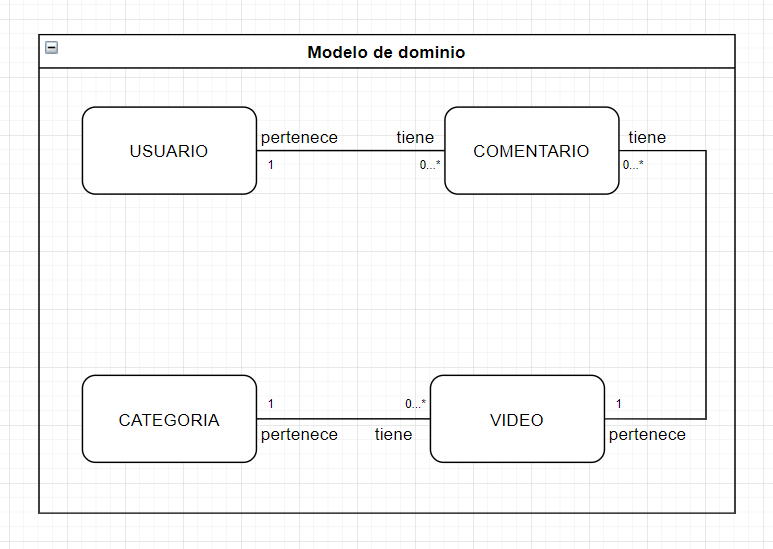
* 1. **Descripción de datos (diccionario):**

Diligencia la siguiente tabla, puede agregar otra columna si lo considera necesario.

|  |  |  |  |
| --- | --- | --- | --- |
| Nombre del atributo / variable | Clase conceptual | Formato o Tipo de Dato | Descripción |
| category\_id | categoría  video | Int | Identificador numérico de la categoría. |
| category | categoría | string | Nombre de la categoría. |
| viewer\_id | usuario  comentario | string | Identificador alfanumérico del canal dueño del video. |
| viewer\_display\_name | usuario | string | Nombre del autor de un comentario. |
| video\_id | comentario  video | string | Identificador alfanumérico del vídeo. |
| comment\_id | comentario |  |  |
| comment\_text | comentario | string | Texto del comentario. |
| like\_count | comentario | Int | Número de pulgares arriba de un comentario. |
| reply\_count | comentario | Int | Número de respuestas que tiene un comentario. |
| publish\_time | video | date | Tiempo de publicación del video.< |
| channel\_title | video | string | Nombre del canal al que pertenece el vídeo. |
| Title | video | string | Título del vídeo. |
| Description | video | string | Descripción del vídeo. |
| trending\_date | video | date | Fecha en la cual el video entró a trending. |
| thumbnail\_link | video | string | Link a la imagen de la miniatura del video. |
| comments\_disabled | video | bool | Indica si el video tiene los comentarios deshabilitados. |
| ratings\_disabled | video | bool | Indica si el video tiene las calificaciones desactivadas. |
| video\_error\_or\_removed | video | bool | Indica si el video presenta error o ha sido removido. |
| Views | video | Int | Número de veces que ha sido visto el vídeo. |
| Likes | video | Int | Número de pulgares arriba del vídeo. |
| Dislikes | video | Int | Número de pulgares abajo del vídeo. |
| comment\_count | video | Int | Número de comentarios que tiene el video. |
| Tags | video | string | Los tags que tiene el video. |

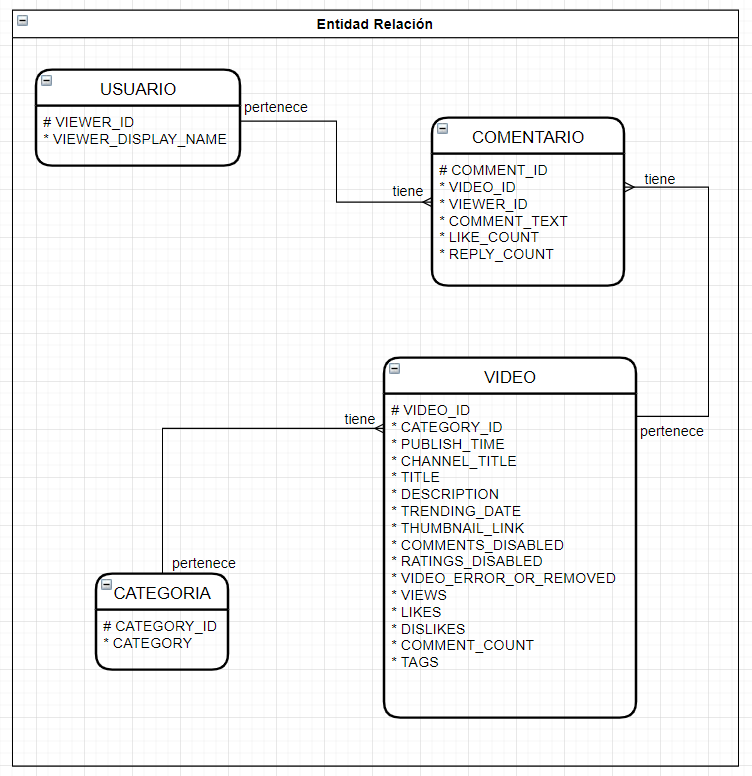
* 1. **Modelo del dominio**

***Observación****: Incluya el gráfico del modelo del dominio que representa la estructura de datos de su problema.*

**

1. **Modelo Entidad-Relación**
   1. **Toma de pantalla del modelo E-R**

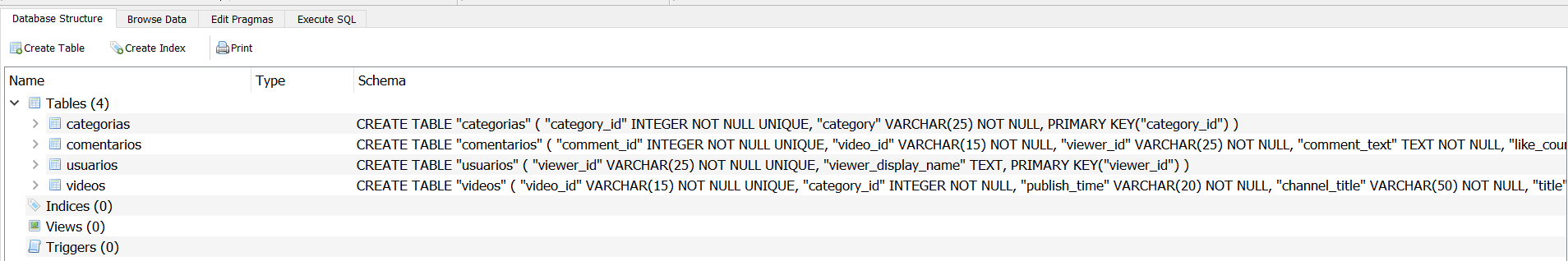
***Observación****: lo que se pide, puede usar https://draw.io o Microsoft Visio® y modele usando la notación de Barker.*

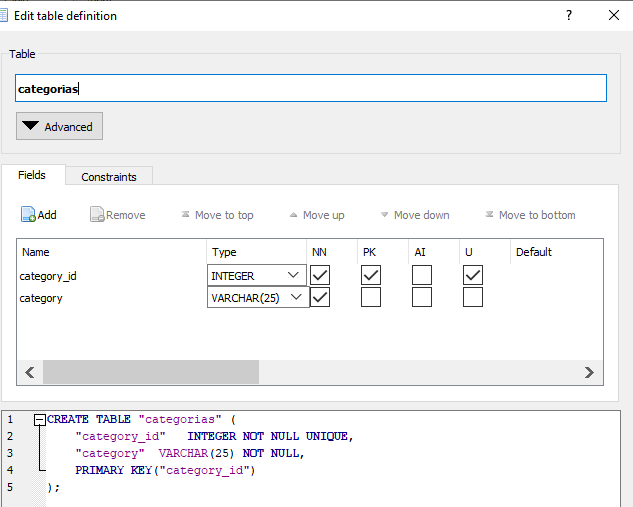
**

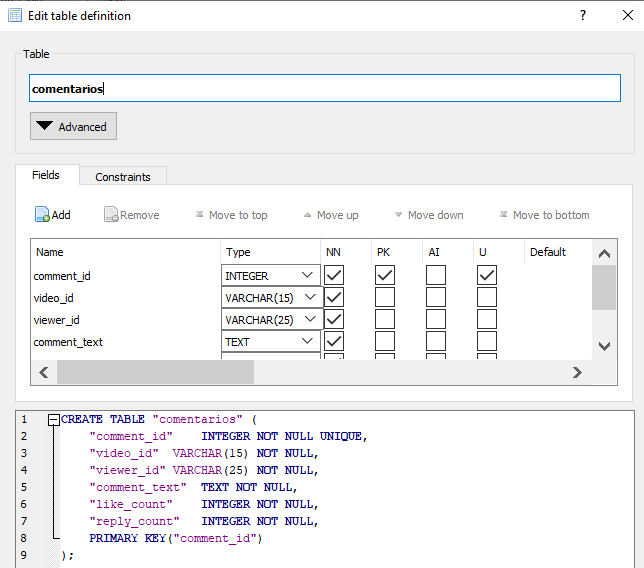
* 1. **Sentencia o consulta de creación del tabla(s)**

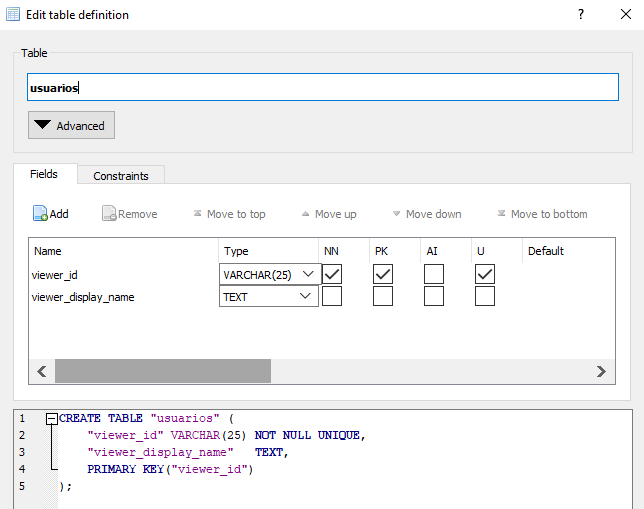
***Observación****: Escriba el código en el Sistema de Gestión de Bases de Datos Relacionales de su elección (se recomienda SQLite por simplicidad, mediante https://sqlitebrowser.org/) para crear las tablas que corresponda con su conjunto de datos específico. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.2.Creacion\_Tablas.sql*

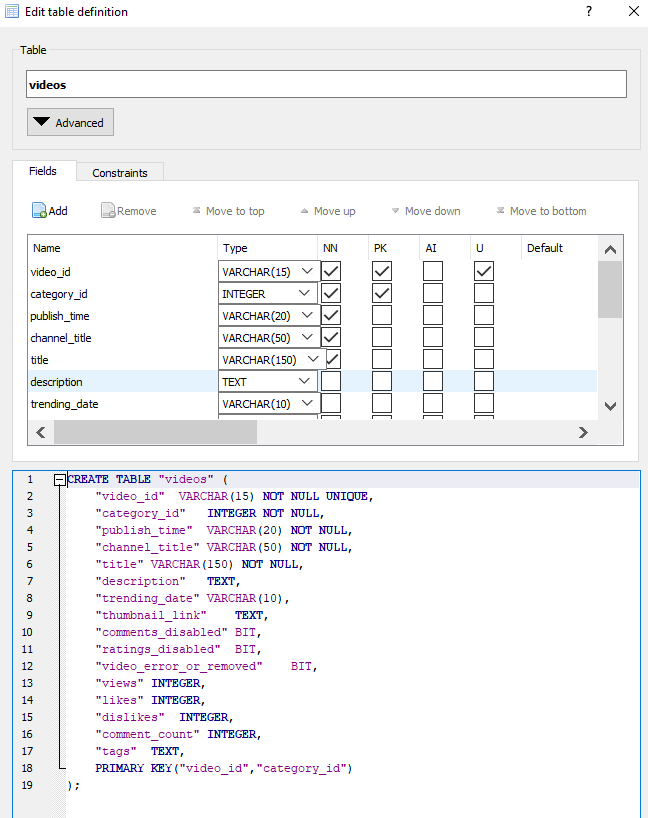
***Se recomienda repasar SQL en*** [***https://www.w3schools.com/sql/default.asp***](https://www.w3schools.com/sql/default.asp)

******

******

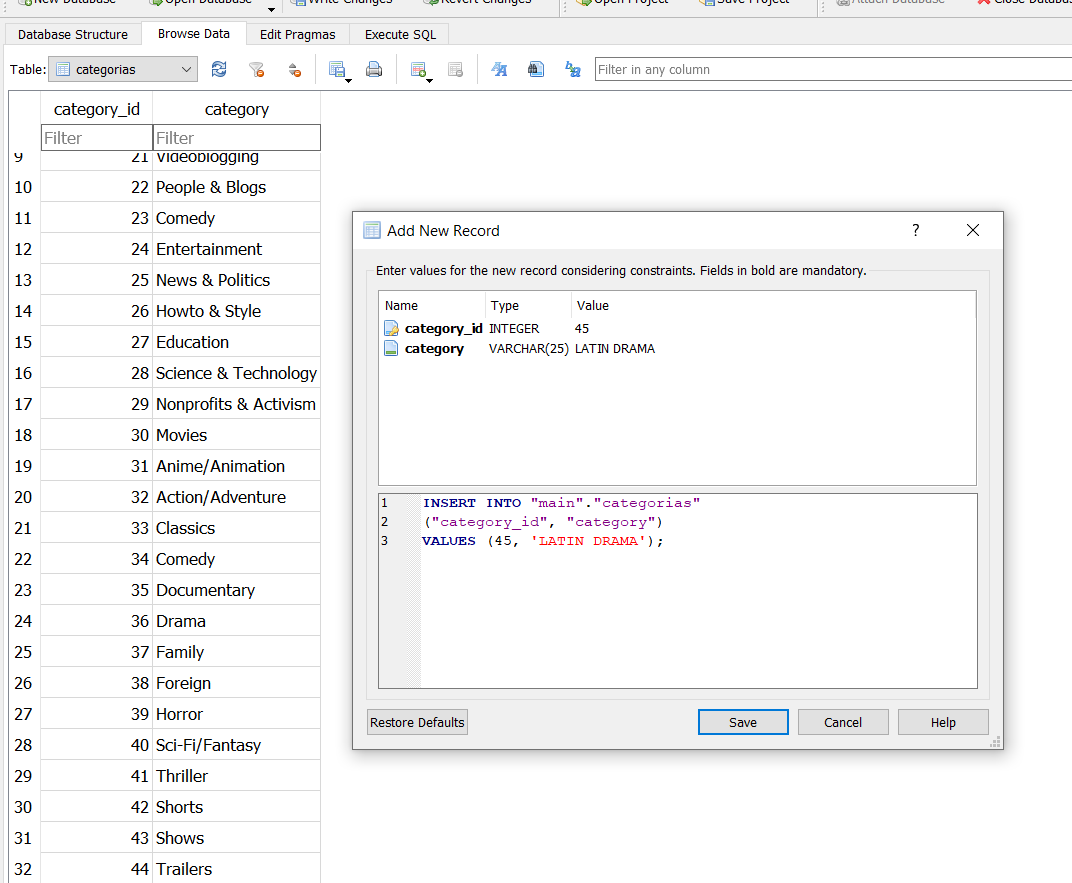
******

******

******

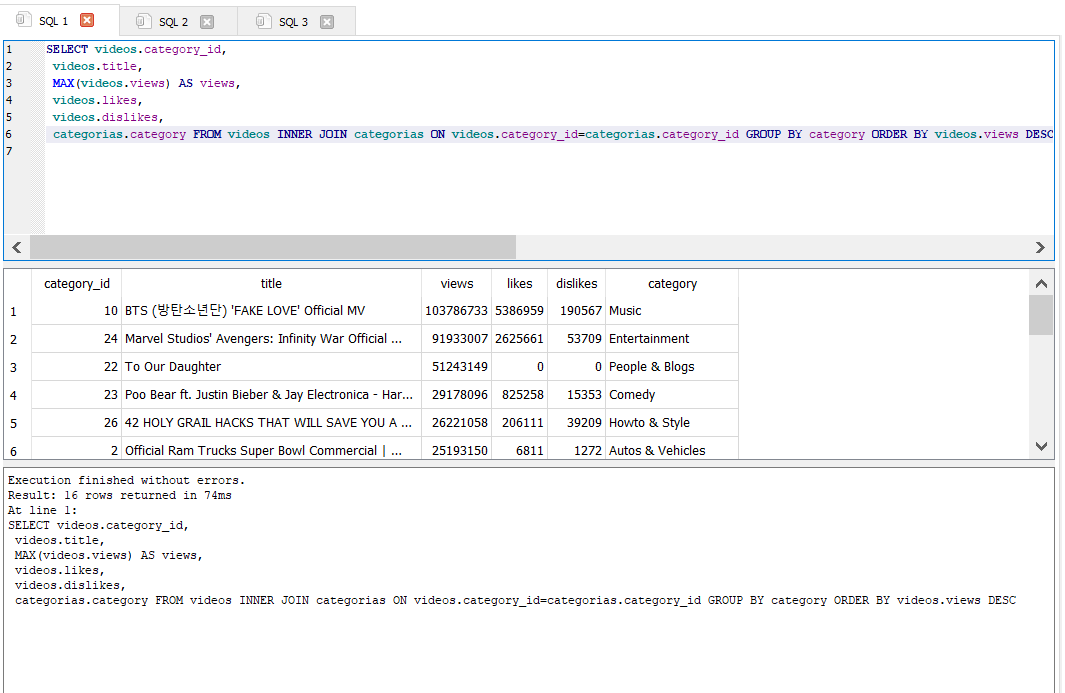
* 1. **Sentencias para Insertar datos**

***Observación****: Escriba el código para insertar los datos en cada una de las tablas creadas. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.3.Insertar\_Datos.sql*

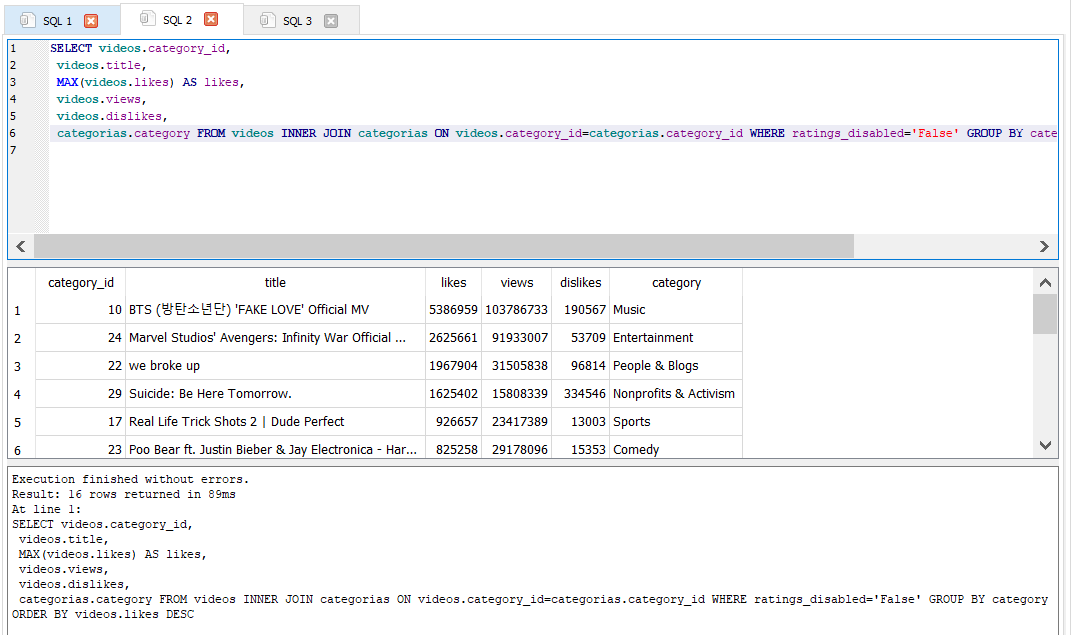
**

* 1. **Sentencia de consulta**

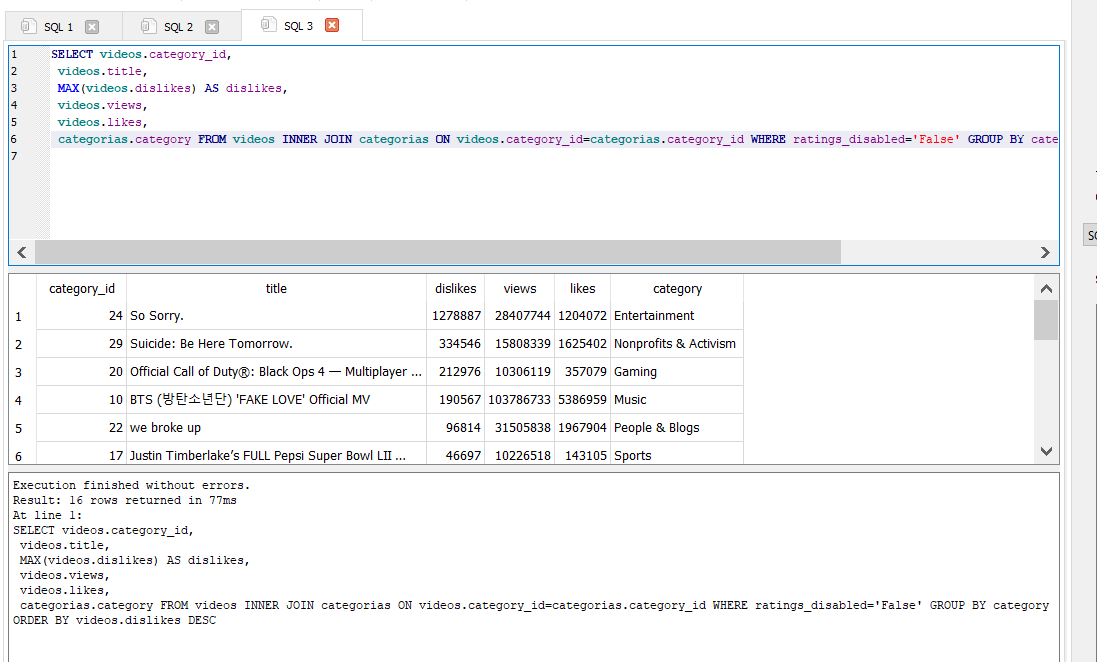
***Observación****: realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de pantalla donde evidencie la sentencia SQL y su correspondiente ejecución. Además, *Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.4.Consultar\_Datos.sql*

*Videos con máximo de views por categoría.*

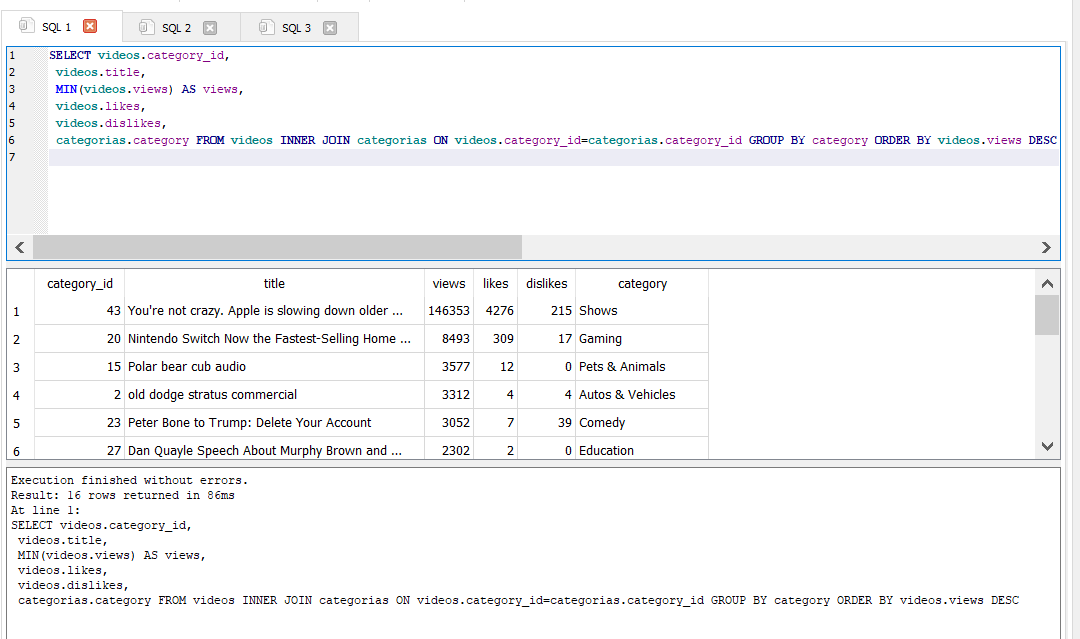
*Videos con máximo de likes por categoría, excluyendo los que tienen ‘rating disable’ = True.*

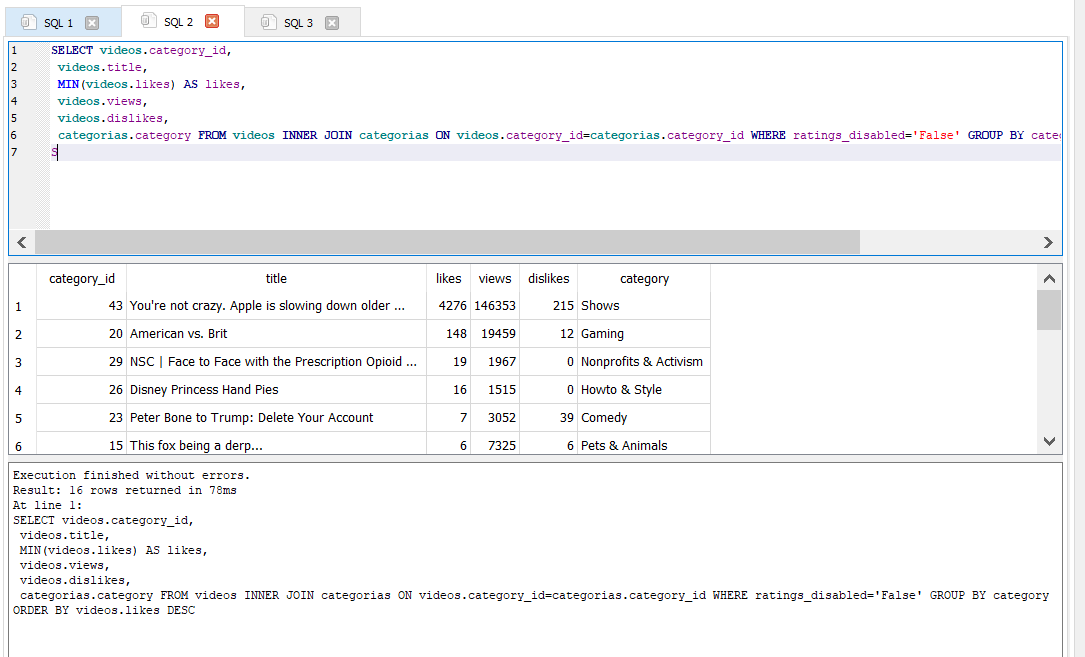
**

*Videos con máximo de dislikes por categoría, excluyendo los que tienen ‘rating disable’ = True.*

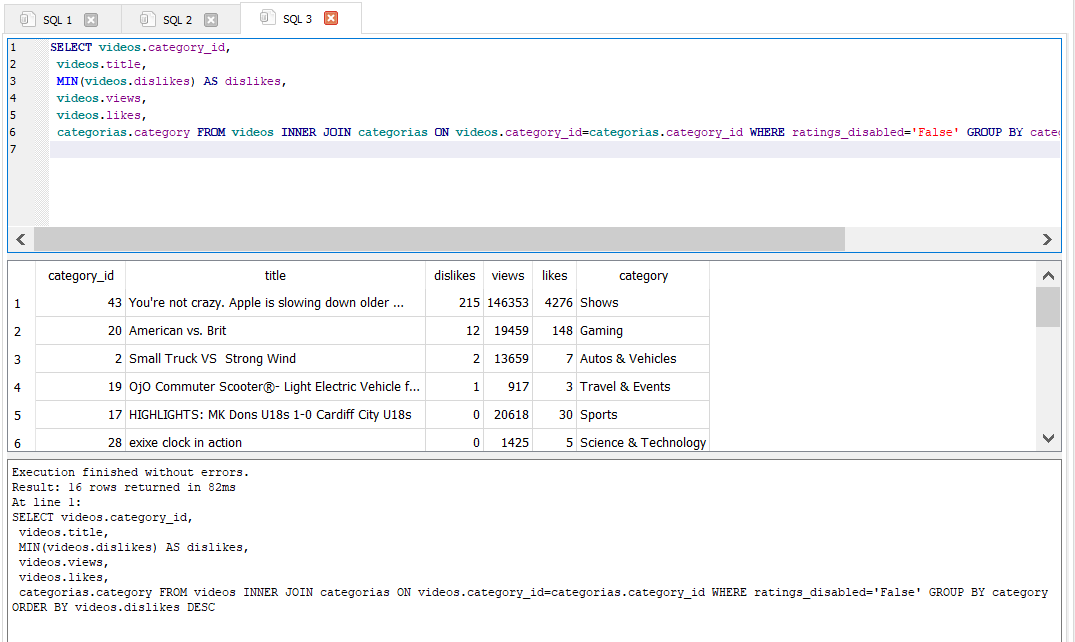
**

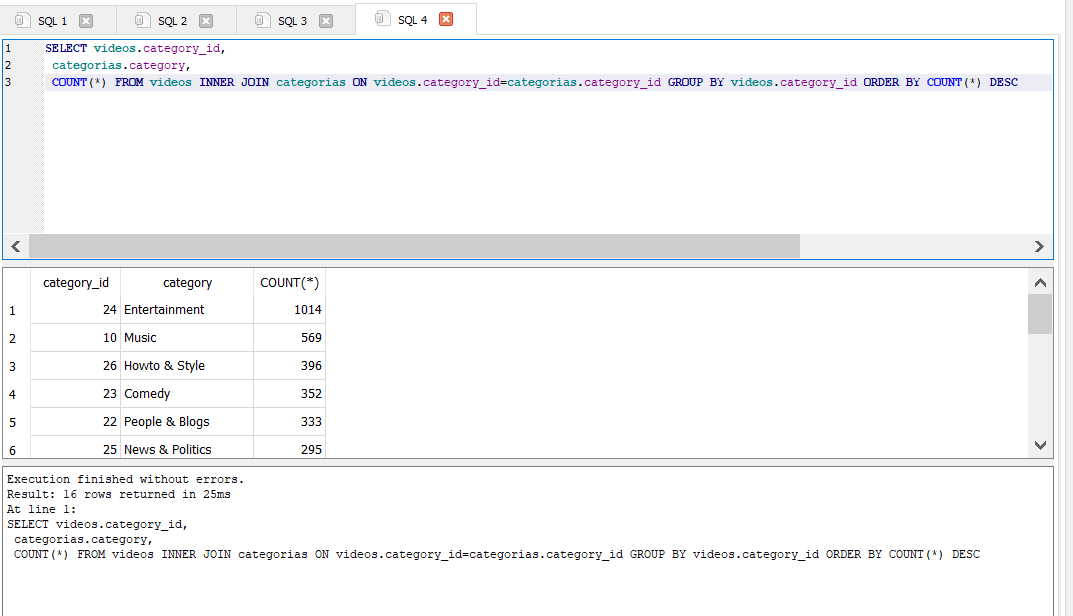
*Videos con mínimo de views por categoría.*

**

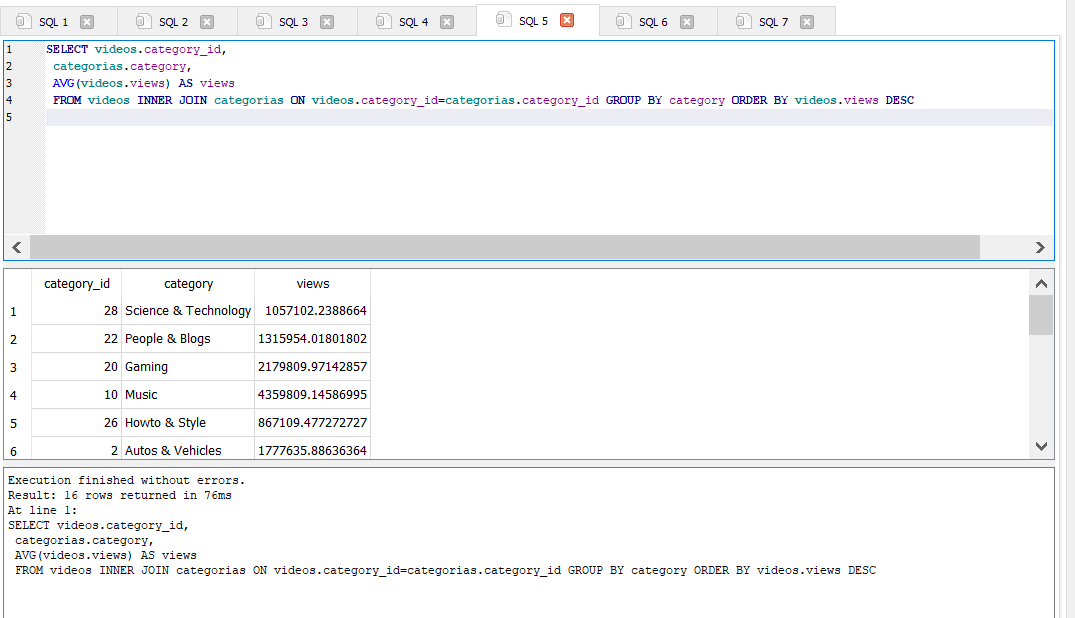
*Videos con mínimo de likes por categoría, excluyendo los que tienen ‘rating disable’ = True.*

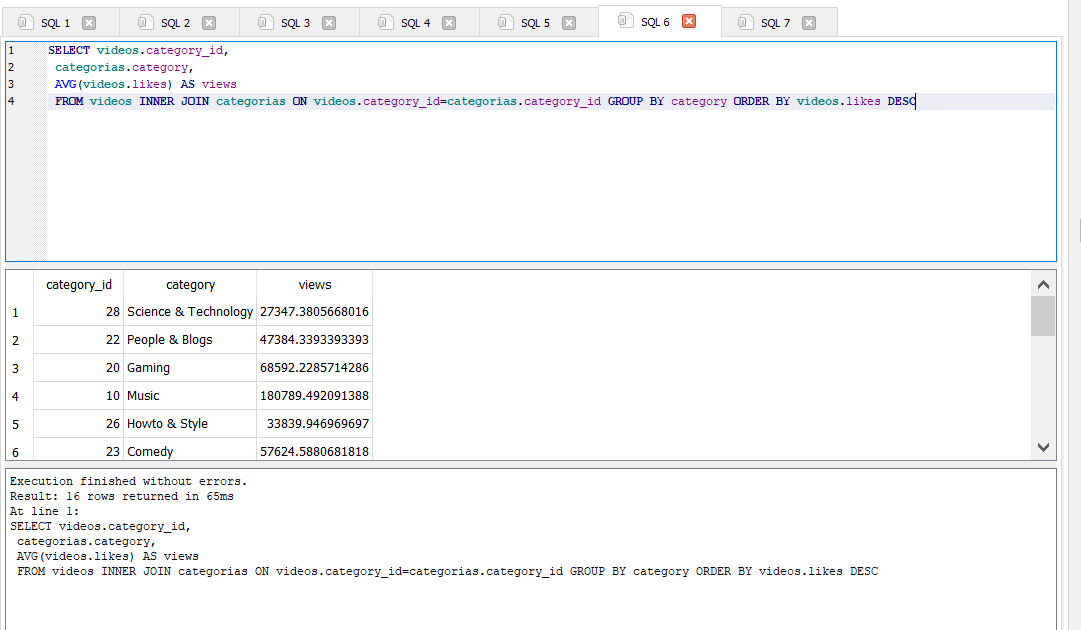
*Videos con mínimo de dislikes por categoría, excluyendo los que tienen ‘rating disable’ = True.*

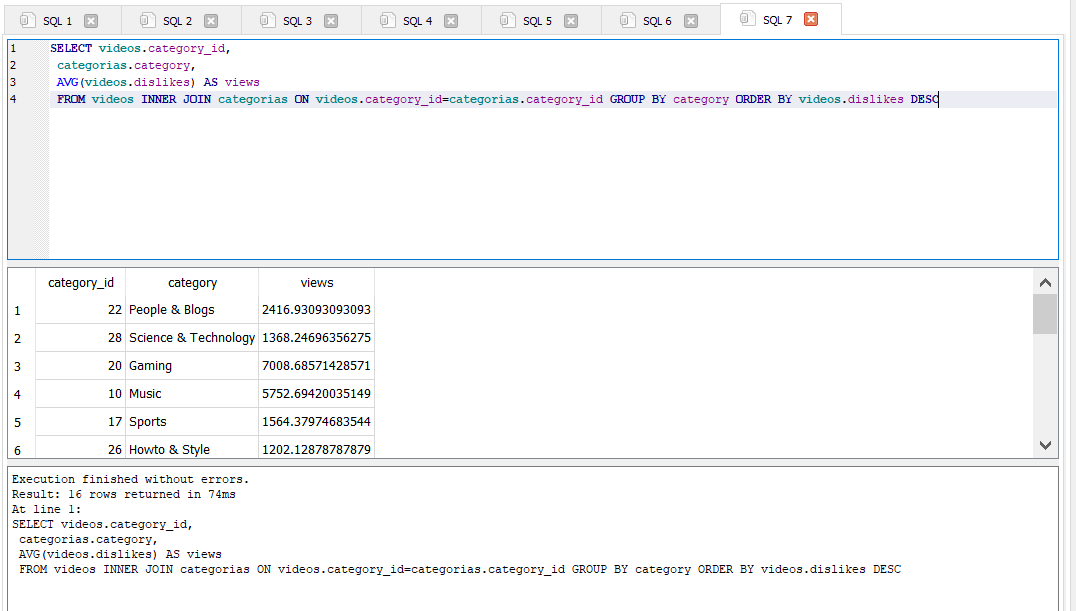
**

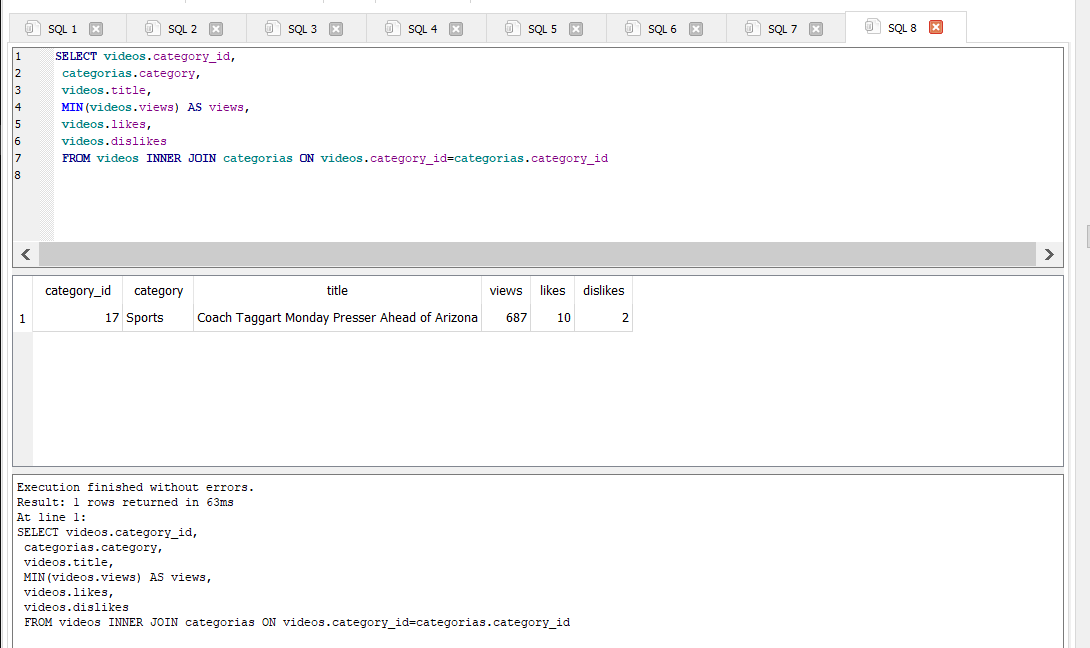
*Cuenta de videos por categoría*

*Promedio de views por por categoría.*

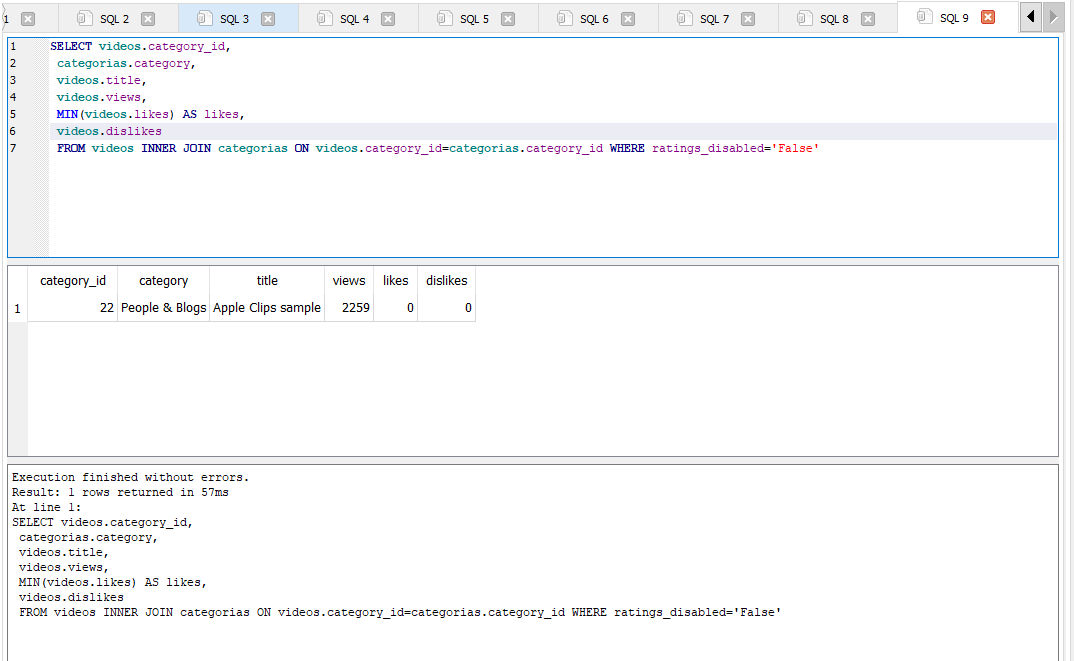
**

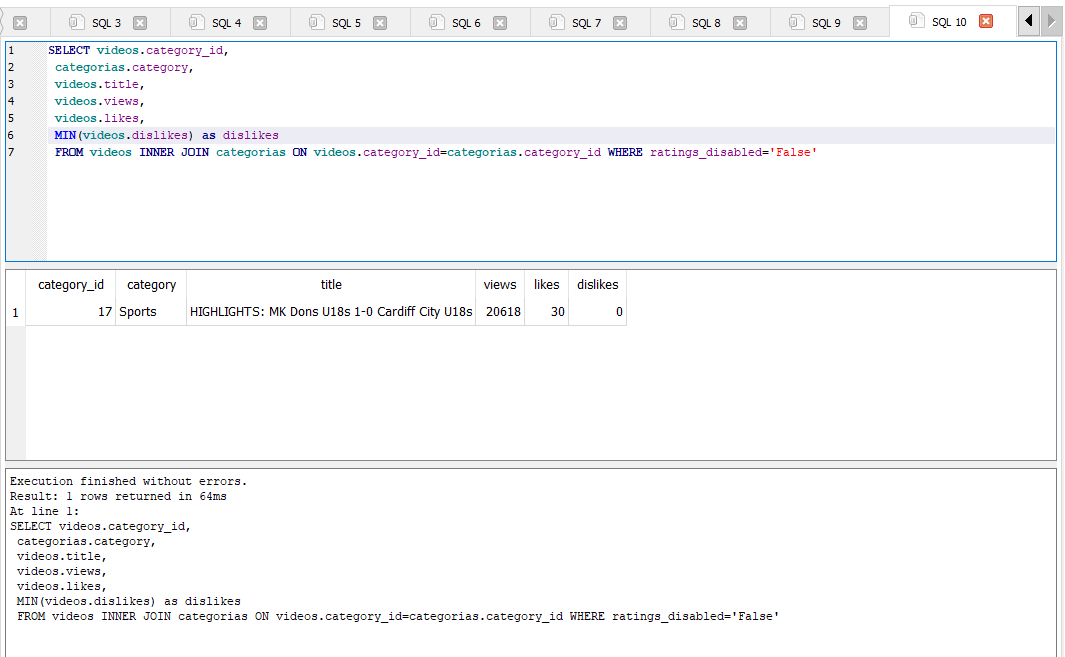
*Promedio de likes por por categoría, excluyendo los que tienen ‘rating disable’ = True.*

*Promedio de dislikes por por categoría, excluyendo los que tienen ‘rating disable’ = True.*

*Video con el mínimo de views.*

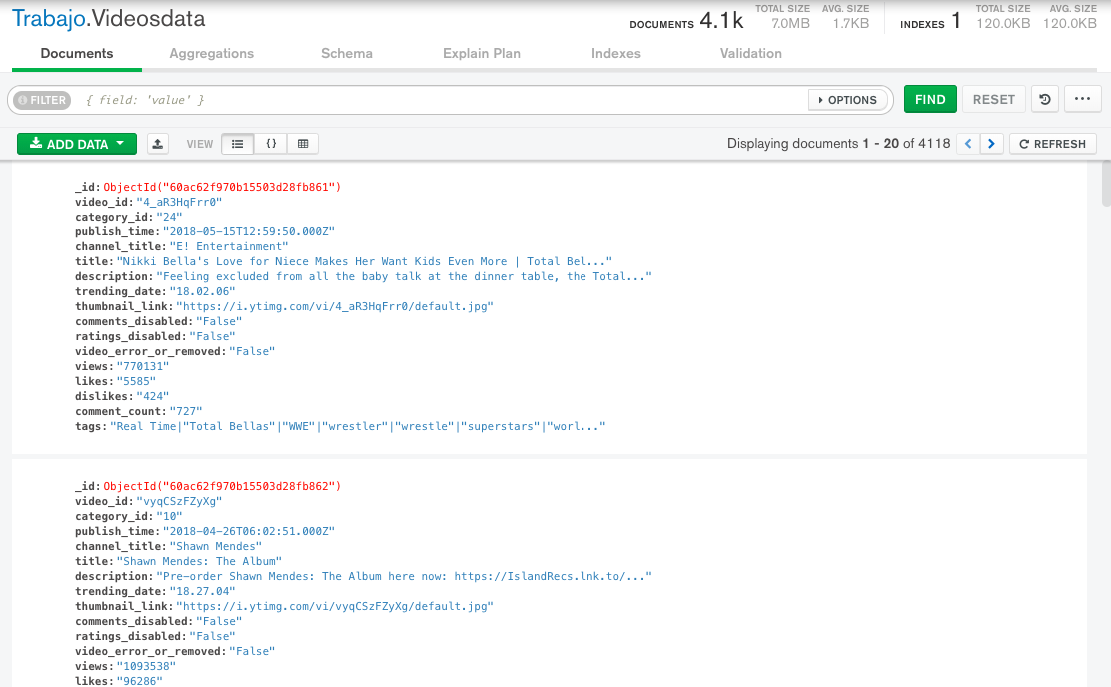
*Video con el mínimo de likes, excluyendo los que tienen ‘rating disable’ = True.*

**

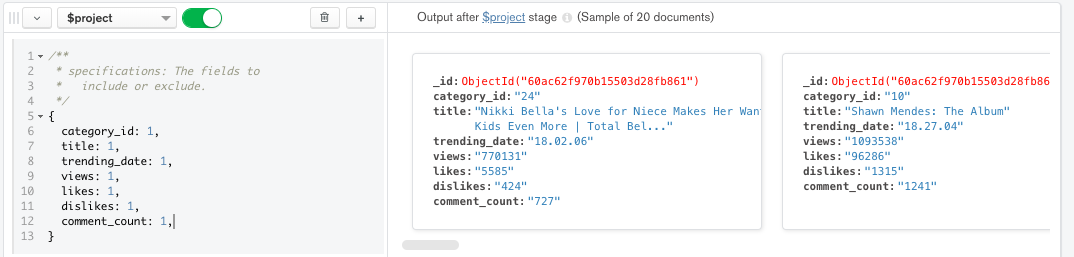
*Video con el mínimo de dislikes, excluyendo los que tienen ‘rating disable’ = True.*

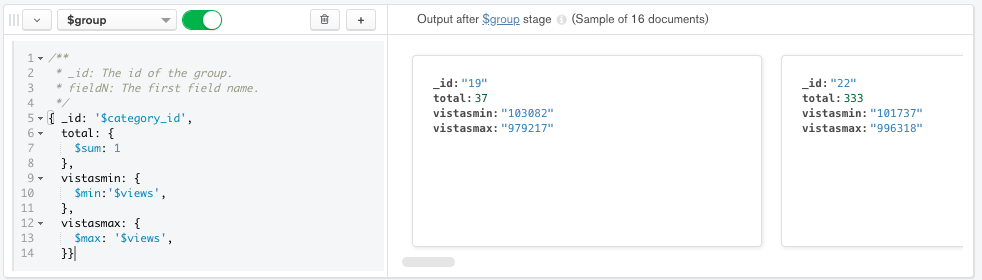
1. **MongoDB**
   1. **Sentencia o consulta de creación del documento(s)**

***Observación****: Escriba el código en MongoDB para crear al menos 20 documentos que correspondan a su conjunto de datos específico. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.4.1.Creacion\_Documentos.sql*

**

* 1. **Sentencia de consulta**

***Observación****: Realice la exploración básica de los datos, conteos totales y por categorías, máximos y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Anexe las tomas de pantalla donde evidencie la sentencia SQL y su correspondiente ejecución. Además, *Almacene en el repositorio* **(REPO EN GITHUB)** *script con el nombre de T1.4.2.Consultar\_Datos.sqlRefinamiento de la base de datos.*

**

*Maximos y minimos por vistas de videos por categoría. Adicionalmente, videos totales por categoría.*

*Resultados para las primeras 3 categorías:*

1. ***\_id****:"1"*
2. ***total****:238*
3. ***vistasmin****:"1004815"*
4. ***vistasmax****:"99653"*
5. ***\_id****:"2"*
6. ***total****:44*
7. ***vistasmin****:”25193150”*
8. ***vistasmax****:"95509"*
9. ***\_id****:"10"*
10. ***total****:569*
11. ***vistasmin****:"103786733”*
12. ***vistasmax****:"9998724"*

**

*Máximos y mínimos tanto para likes como dislikes por categoría. Ademas de videos totales por categoría.*

*Resultados para las primeras 3 categorías:*

1. ***\_id****:"1"*
2. ***total****:238*
3. ***Likesmin****:"0"*
4. ***Likesmax****:"9978"*
5. ***Dislikesmax****:"9917"*
6. ***Dislikesmin****:"0"*
7. ***\_id****:"2"*
8. ***total****:44*
9. ***Likesmin****:"0"*
10. ***Likesmax****:"9826"*
11. ***Dislikesmax****:"93"*
12. ***Dislikesmin****:"2"*
13. ***\_id****:"10"*
14. ***total****:569*
15. ***Likesmin****:"0"*
16. ***Likesmax****:"5386959"*
17. ***Dislikesmax****:"190567"*
18. ***Dislikesmin****:"0"*
19. **Análisis de lectura**

***Observación****: Considerando el artículo: “*The Definitive Guide to Graph Databases for the RDBMS Developer*” de Neo4J. Compartido en las carpeta de lecturas recomendadas. Analice y responda cada pregunta en máximo 150 palabras:*

1. ¿Cuáles son las limitaciones, que se pueden inferir de la lectura, para migrar los conjuntos de datos relacionales a NoSQL?

Las nuevas tecnologías generan diferentes retos como resistencias al cambio tanto de las empresas como en los equipos de trabajo. A su vez, en la actualidad se encuentran pocos especialistas en estas nuevas tecnologías, por lo tanto es un campo inexplorado con pocos recursos para autogestión del personal técnico.

Por otro lado, el tratamiento de los conjuntos de datos relacionales es totalmente diferente a los no relacionales. Así, manipular y migrar los mismos como no relacionales podrían acarrear problemas, ya que se tendrían propiedades que no podrían migrar de la misma manera.

1. ¿Cuáles limitaciones adicionales se deben considerar, a parte de las mencionadas en el artículo?
2. Dada la misma novedad de la tecnología, la comunidad, Bibliografía, librerías y apoyos online de la que se puede disponer es relativamente escasa en comparación a las existentes para las tecnologías tradicionales.
3. Las bases de datos relacionales tienen estándares bien definidos para la creación, actualización y realización de consultas, cosa que puede no ocurrir en muchas ocasiones con un desarrollo reciente como lo son las bases no relacionales.
4. Migrar el modelo de almacenamiento implica un costo de inversión. Educar a los programadores, adquisición de nuevas licencias y costos adicionales en capital humano.
5. ¿Cuáles son las razones (criterios) que se deben considerar para migrar un conjunto de datos relacionados a NoSQL?
   1. Un criterio de costo beneficio. Como se expresó anteriormente, migrar a esta nueva tecnología implica un costo que no en todas la empresas se traduce en un beneficio extra. En definitiva, hay un beneficio adicional al migrar a las BD NoSQL si la compañía requiere procesamiento en tiempo real o gestionar contenidos de datos de tipo audiovisuales, pdf, entre otros archivos que las RBDM no son capaces de almacenar.
   2. Tener muy en claro el objetivo por el cual se está migrando la información y el tipo de consultas a realizar en la misma que empuje la necesidad de cambiar la base relacional a NoSQL.
   3. El tiempo y consumo de recursos computacionales para la migración. Adicionalmente, tener en cuenta que si se piensa migrar una base de datos relacional con muchos joints a NoSQL esto puede causar un detrimento en eficiencia. Por lo tanto, la características mismas de la base como volumen de la información deben ser tenidas en cuenta.