**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: 2021.04.11 3:28 pm

*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. Diana Carolina Sarabia Sierra - CC. 53036766 |
| 2. David Rodriguez Gonzalez - C.C. 1017210951 |
| 3. Diego Leon Zapata Ruiz - C.C. 71367417 |
| 4. Juan Pablo Vasquez Builes - C.C. 1035428558 |
| 5. Diego Armando Vargas Salazar - 9867018 |
| **REPO EN GITHUB:** [**https://github.com/darmandovargas/agd-2021-1s**](https://github.com/darmandovargas/agd-2021-1s) |

# *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)

# **Comprensión del negocio**

## **Descripción del contexto del negocio.**

El conjunto de datos representa los lesionados por accidentes de tránsito en Colombia registrados por Medicina Legal, desde 1 de enero de 2016 hasta el 31 de diciembre de 2018. Se entiende por accidente de tránsito: "*evento, generalmente involuntario, generado al menos por un vehículo en movimiento, que causa daños a personas y bienes involucrados en él, e igualmente afecta la normal circulación de los vehículos que se movilizan por la vía o vías comprendidas en el lugar o dentro de la zona de influencia del hecho*". (Ley 769 de 2002 - Código Nacional de Tránsito).

## **Identificación del problema:**

Identificación de perfiles de riesgo asociados a la accidentalidad vial, segmentado por rango de edad, tipo de vehículo, objeto colisión y zona geográfica.

## **Determinación de objetivos:**

A partir del conjunto de datos se desea identificar perfiles de riesgo que le permita a las autoridades de tránsito, realizar estrategias de prevención focalizadas. Los perfiles hipotéticos pueden estar relacionados con el tipo de vehículo, la edad, la zona geográfica y el día de la semana. Sin embargo, las técnicas de minería de datos y de machine learning, permiten encontrar otros perfiles y asociaciones ocultas.

* Diseñar un modelo de datos
* Realizar un preprocesamiento y preparación de los datos
* Análisis descriptivo de datos.
* Proponer estrategias de intervención para cada perfil encontrado, que permitan disminuir la accidentalidad y los lesionados. Entender que el alcance del ejercicio, no llega hasta ML.

## **Evaluación de la situación actual:**

El conjunto de datos evidencia el reporte de accidentes de tránsito con lesionados en el territorio colombiano en el periodo comprendido entre el 1 de enero de 2016 hasta el 31 de diciembre de 2018, con variables asociadas que ubican cada evento en tiempo y posición geográfica, sumado a variables descriptoras que indican la tipología de la colisión, el tipo de vehículo y el estado de las vías.

Según un informe de la policía nacional, la cantidad de lesionados durante el 2018 fue de 48.998 personas, de las cuales el 63% fueron hombres y 86% adultos. Por otra lado, la dirección de censos y demografía del DANE publicó en sus estadísticas vitales del 2018 que los accidentes de tránsito fueron la octava causa de muerte en Colombia, con 7.155 casos. También indica el informe que ocupan en segundo lugar en defunciones por causa externa con un 24.4%.

# **Comprensión de los datos**

## **recolección de datos**

Los datos a utilizar son cifras consolidadas en los Sistemas de Información del Instituto Nacional de Medicina Legal y Ciencias Forenses. La información es procesada por la agencia Nacional de Seguridad Vial y en ella se encuentra el historial relacionado de personas lesionadas en accidentes de tránsito a nivel nacional desde el año 2016 hasta 2018; dicha información se encuentra distribuida de forma zonal o departamental, de rango por edad entre otros.

La información es recolectada anualmente y para obtenerla se realiza un levantamiento de información del sitio del suceso, con el fin de localizar e identificar las evidencias físicas de las causas de los accidentes. A pesar de que la base tiene variables como el estado de vida, usuario vía, el género y la hora del incidente al momento de trabajar se evidencia que no hay información de calidad así que se ignoran estos datos.

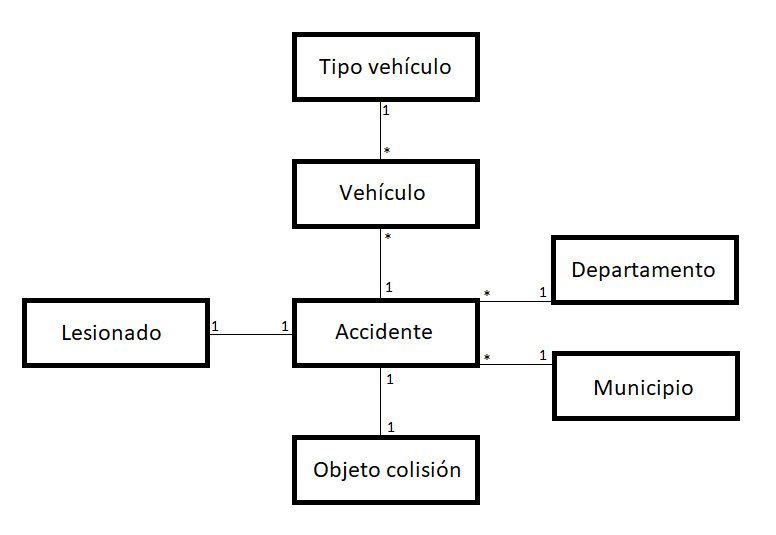
Pueden ver el archivo en el siguiente link [T1.2.1.Datos.zip](https://github.com/darmandovargas/agd-2021-1s/blob/main/T1.2.1.Datos.zip) o descomprimidos en este [link](https://github.com/darmandovargas/agd-2021-1s/tree/main/T1.2.1.Datos).

## **Descripción de datos (diccionario):**

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

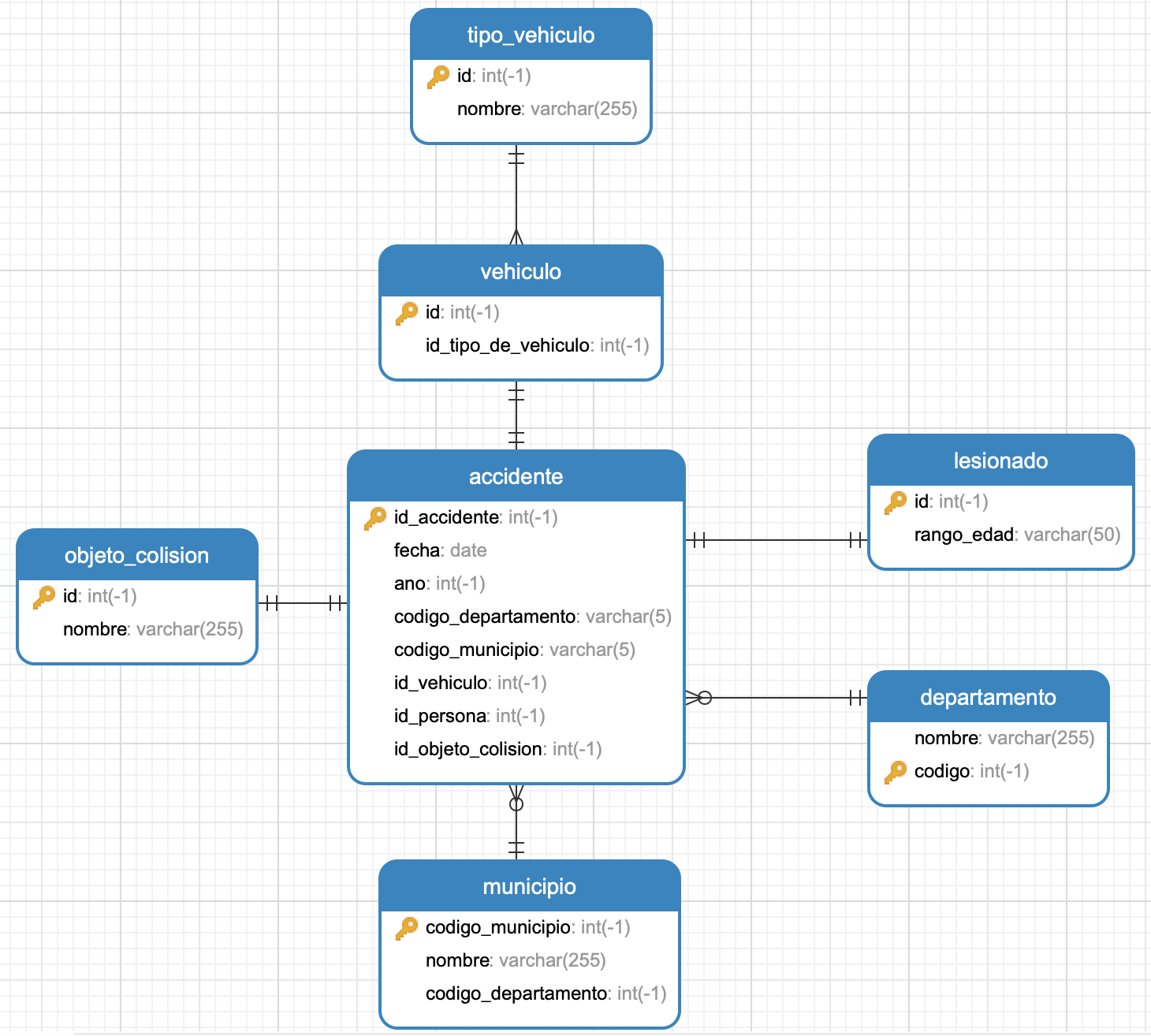
|  |  |  |
| --- | --- | --- |
| Nombre del atributo / variable | Formato o Tipo de Dato | Descripción |
| Id\_accidente | Número | Identificador auto incremental de los accidentes reportados |
| Fecha | Fecha | Fecha en que ocurrió el hecho |
| Ano | Número | Año en que ocurrió el hecho |
| Id\_departamento | Número | Identificador del número de departamento |
| Codigo\_departamento | Número | Código departamento de Colombia donde ocurrió el hecho (DIVIPOLA) |
| Departamento | Texto simple | Nombre Departamento de Colombia donde ocurrió el hecho |
| Id\_municipio | Número | Identificador de Municipio de Colombia donde ocurrió el hecho |
| Codigo\_municipio | Número | Código del municipio de Colombia donde ocurrió el hecho (DIVIPOLA) |
| Municipio | Texto simple | Nombre Municipio de Colombia donde ocurrió el hecho |
| Id\_vehiculo | Número | Identificación del tipo de vehículo accidentado |
| Vehículo | Texto simple | Nombre del vehículo accidentado |
| Id\_grupo | Número | Identificación del grupo de vehículo accidentado |
| grupo\_vehiculo | Texto simple | Nombre del grupo de tipo de vehículo accidentado |
| Id\_objeto | Número | Identificación del objeto de colisión |
| Objeto\_colisión | Texto simple | Descripción del objeto de colisión del accidente |
| Id\_rango\_edad | Número | Identificación de la edad de la persona lesionada (0 a 100) |
| Rango\_edad | Texto simple | Rango de edad de la persona lesionada (0 a 100, cada cinco años) |
| Id\_tipo\_lesionado | Número | Identificación del tipo de lesión |
| Lesionado | Texto simple | Peatón, usuario de bicicleta , usuario de moto, usuario de vehículo, otros |

## **Modelo del domin**i**o**



# **Modelo Entidad-Relación**

## **Toma de pantalla del modelo E-R**



## **Sentencia o consulta de creación del tabla(s)**

Las sentencias de consulta están en el siguiente Github:

[T1.3.2.Creacion\_Tablas](https://github.com/darmandovargas/agd-2021-1s/blob/main/T1.3.2.Creacion_Tablas.sql)

## **Sentencias para Insertar datos (Sentencias** de todos los datos. Se adjunta aparte**)**

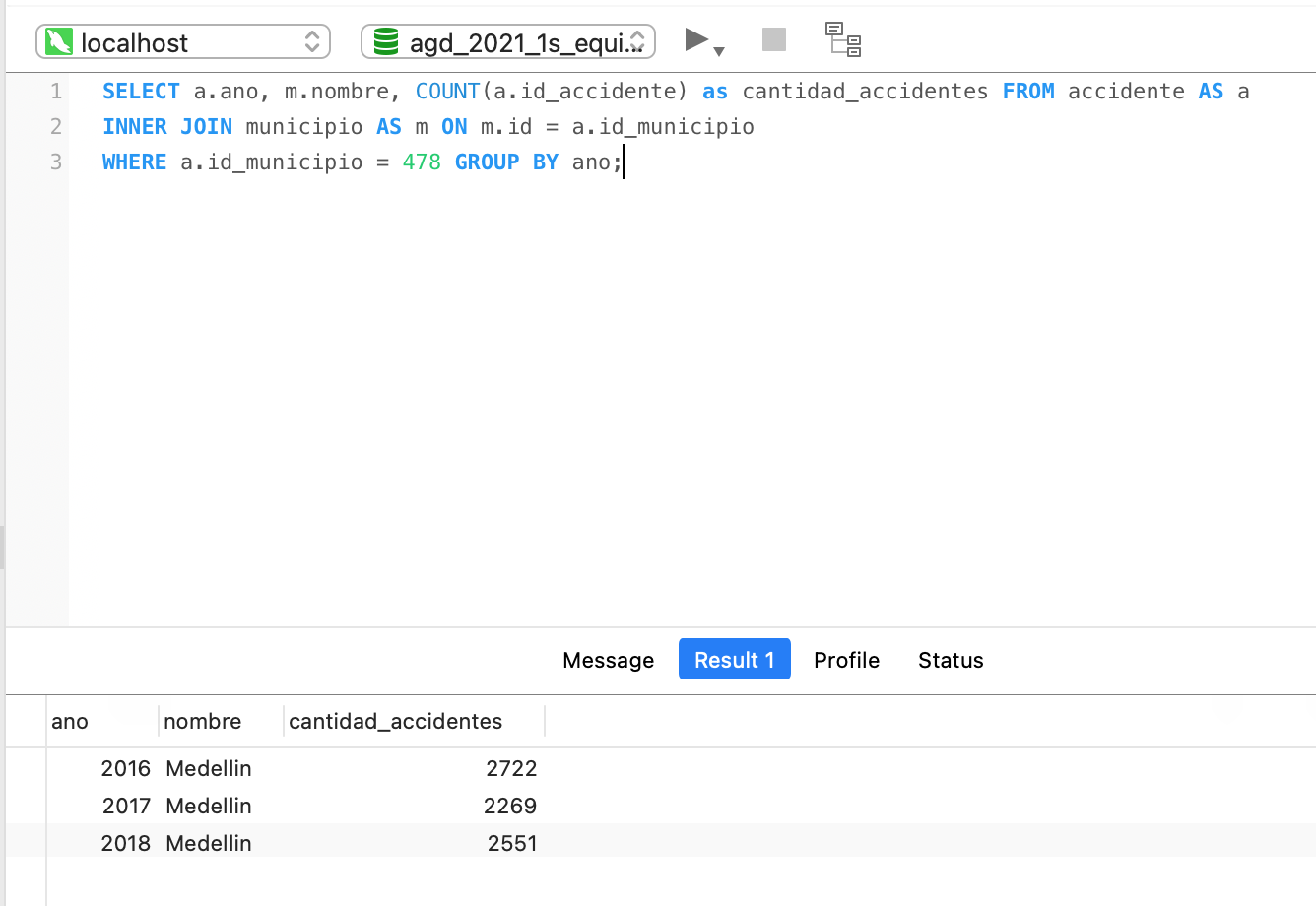
[*T1.3.3.Insertar\_Datos.sql*](https://github.com/darmandovargas/agd-2021-1s/blob/main/T1.3.3.Insertar_Datos.sql)

## **Sentencia de consulta**

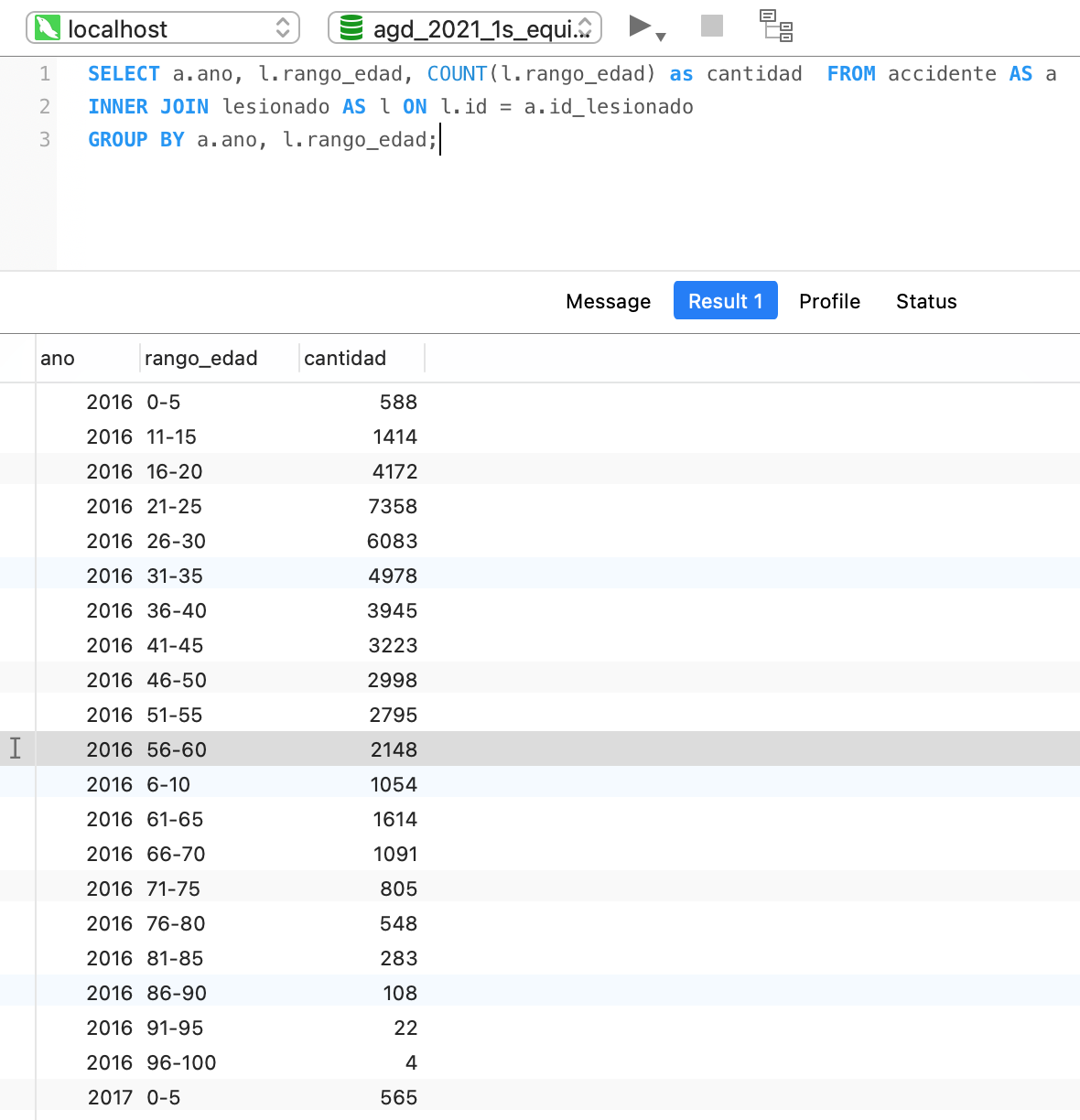
[*T1.3.4.Consultar\_Datos.sql*](https://github.com/darmandovargas/agd-2021-1s/blob/main/T1.3.4.Consultar_Datos.sql)

A continuación se explican las consultas realizadas con el fin de inspeccionar la información.

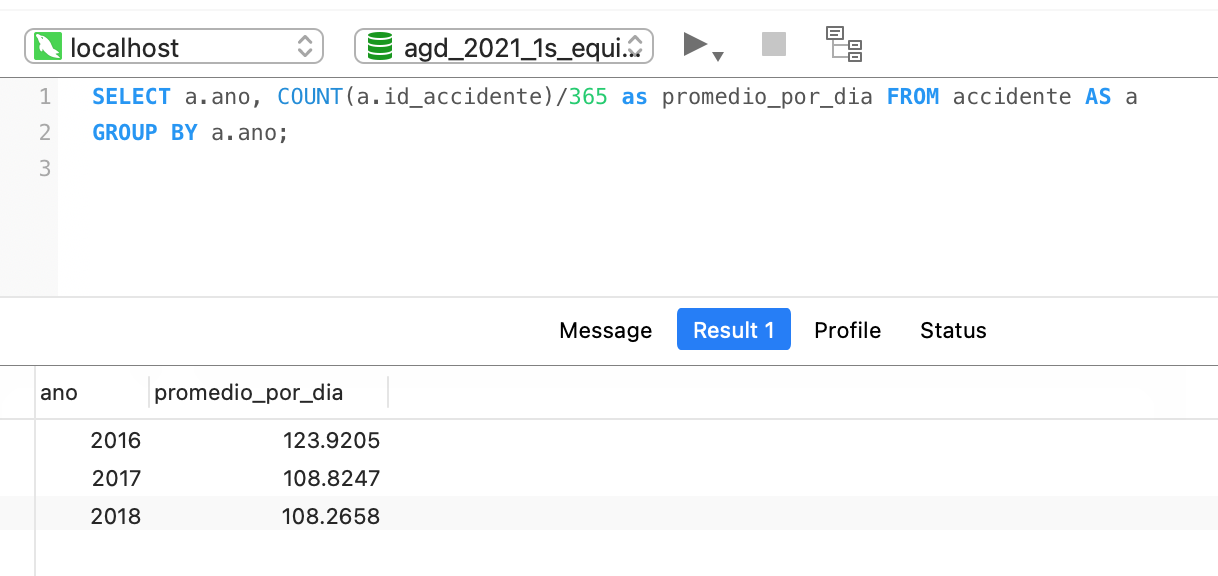
**3.4.1** Consulta de accidentes por año en la ciudad de Medellín, donde se evidencia una reducción significativa en el año 2017 y paulatinamente en los años de estudio.

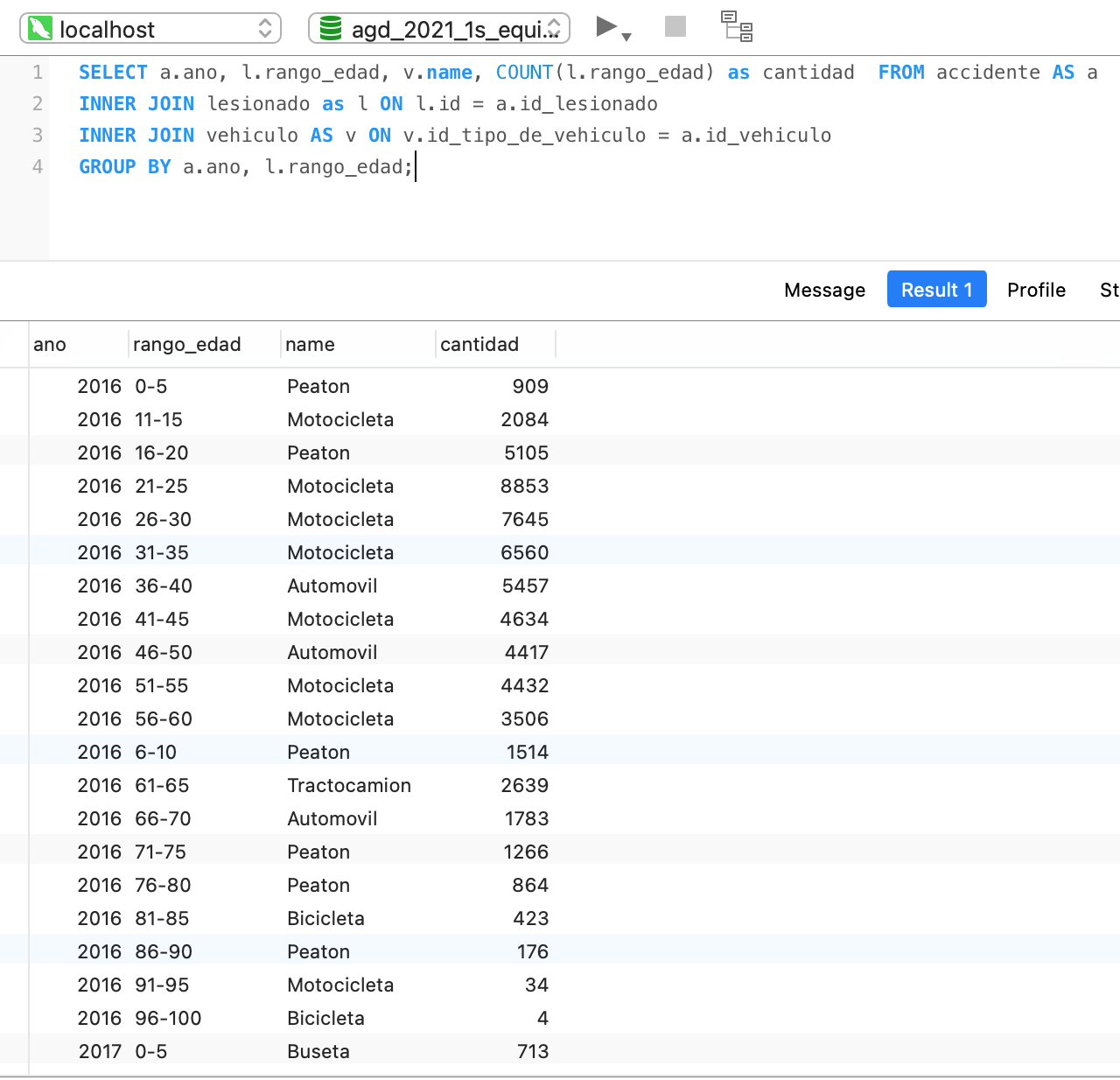


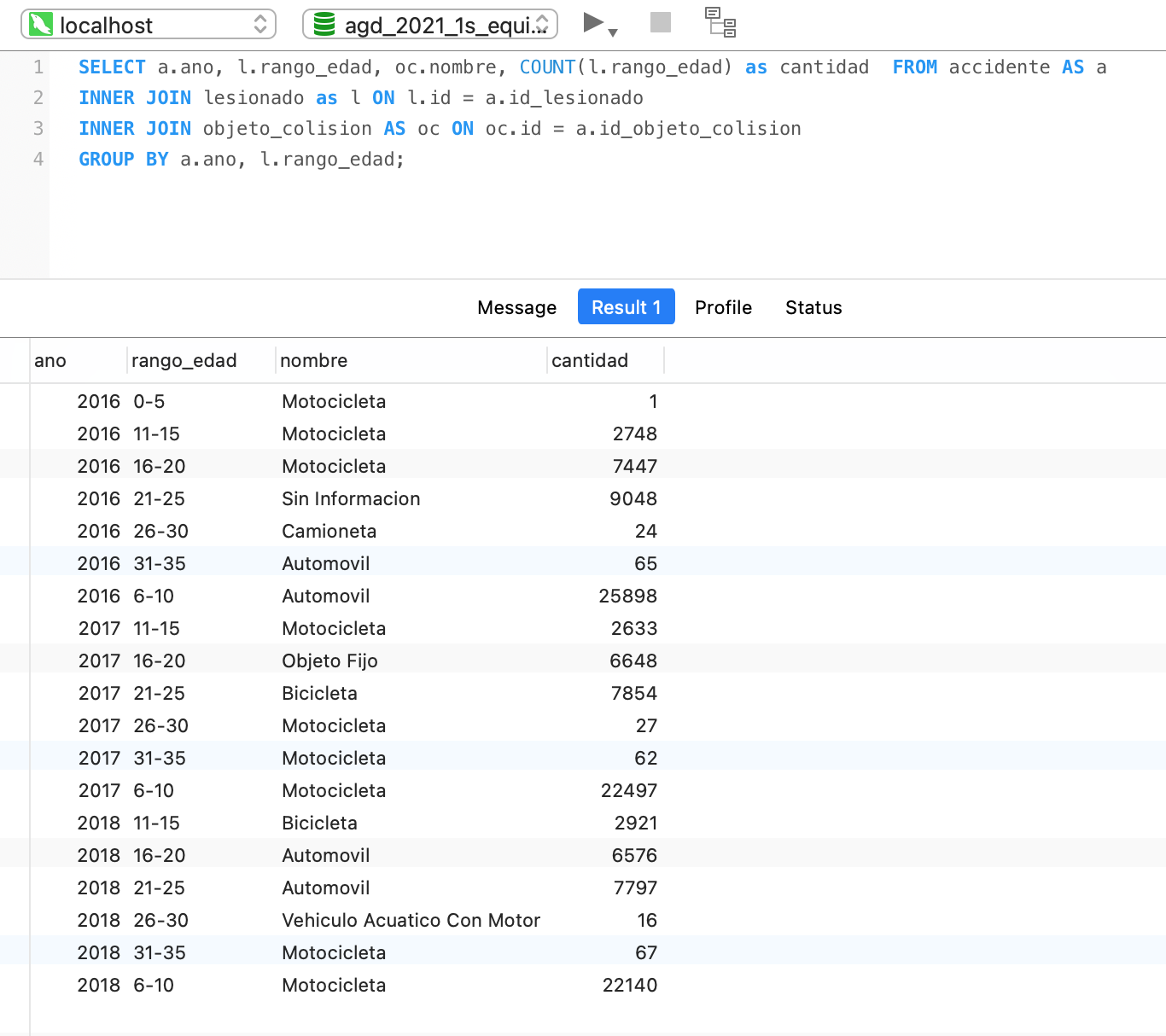
**3.4.2** Consulta de accidentes por año y rango de edad, los cuales nos permite evidenciar un común denominador donde las personas de 21 a 30 años son los más vulnerables en los accidentes año tras año, por lo cual se recomiendan políticas públicas que busquen disminuir estas cifras enfocado en este grupo poblacional.



**3.4.3** Consulta de promedio de accidentes por día agrupados por año, donde se evidencia una reducción en el año 2017 y 2018, más leve en el último.



**3.4.4** Consulta de vehículos involucrados en el accidente del lesionado, notamos una relación con los grupos de mayor riesgo de 21 a 25 seguido por el grupo de 26 a 30 años con la motocicleta en 2016, 2017 y 2018 como vemos a continuación:  


**3.4.5** Consulta de lesionados por años según rango de edad y objeto de colisión, donde se evidencia que la población más vulnerable se encuentran entre los 21 y 35 años colisión principalmente con motocicletas, esto para toda la serie de datos.

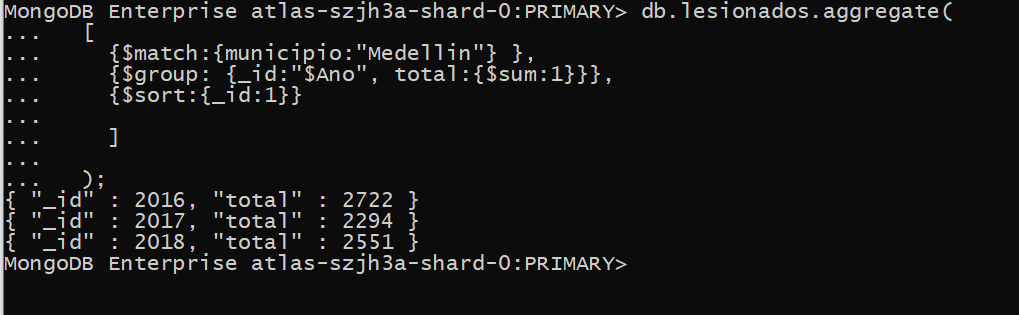
# **MongoDB**

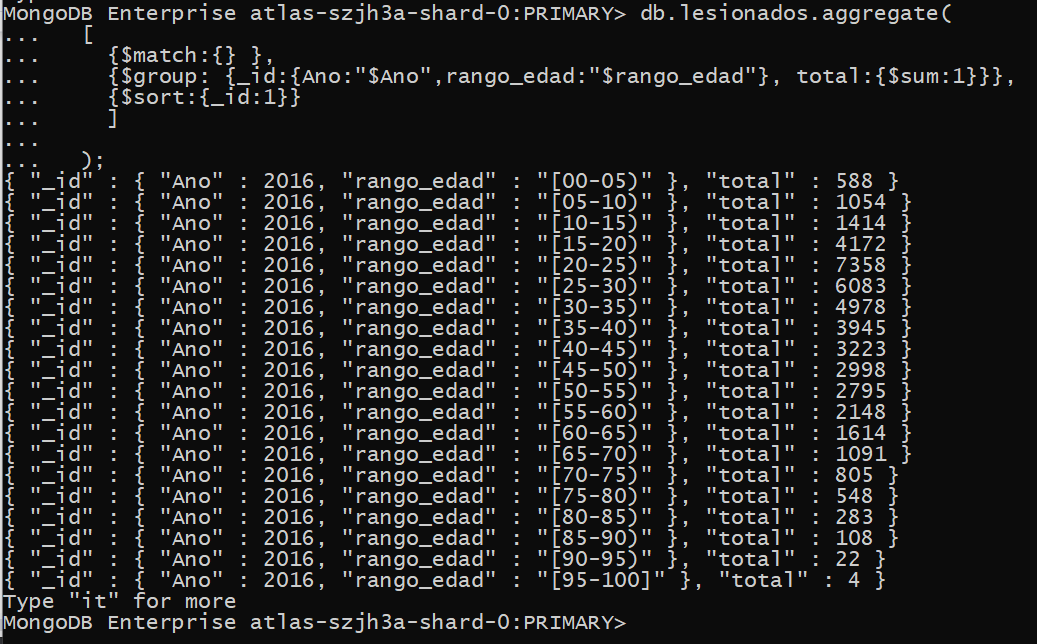
## **Sentencia o consulta de creación del documento(s)**

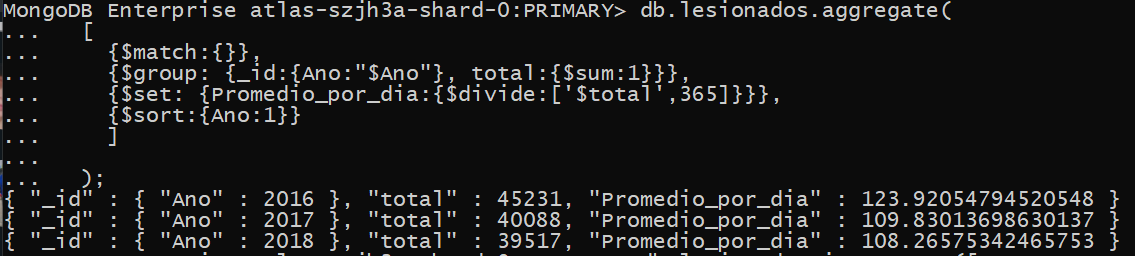
[*T1.4.1.Creacion\_Documentos.txt*](https://github.com/darmandovargas/agd-2021-1s/blob/main/T1.4.1.Creacion_Documentos.txt)

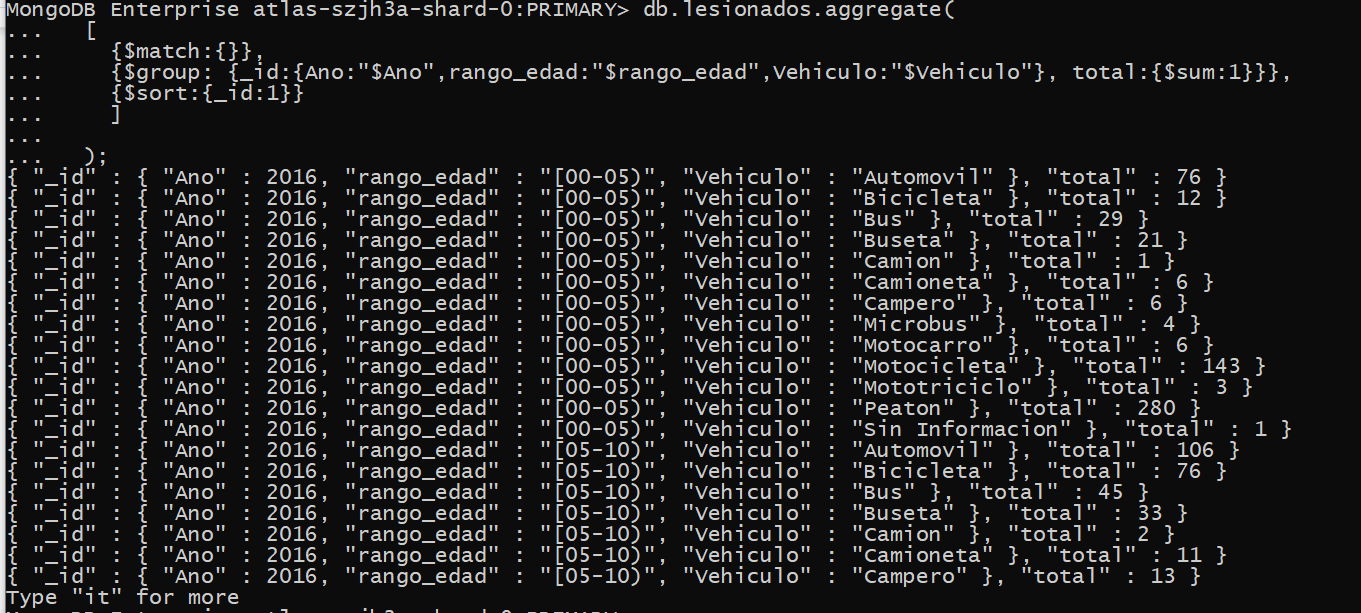
## **Sentencia de consulta**

[*T1.4.2.Consultar\_Datos.txt*](https://github.com/darmandovargas/agd-2021-1s/blob/main/T1.4.2.Consultar_Datos.txt)

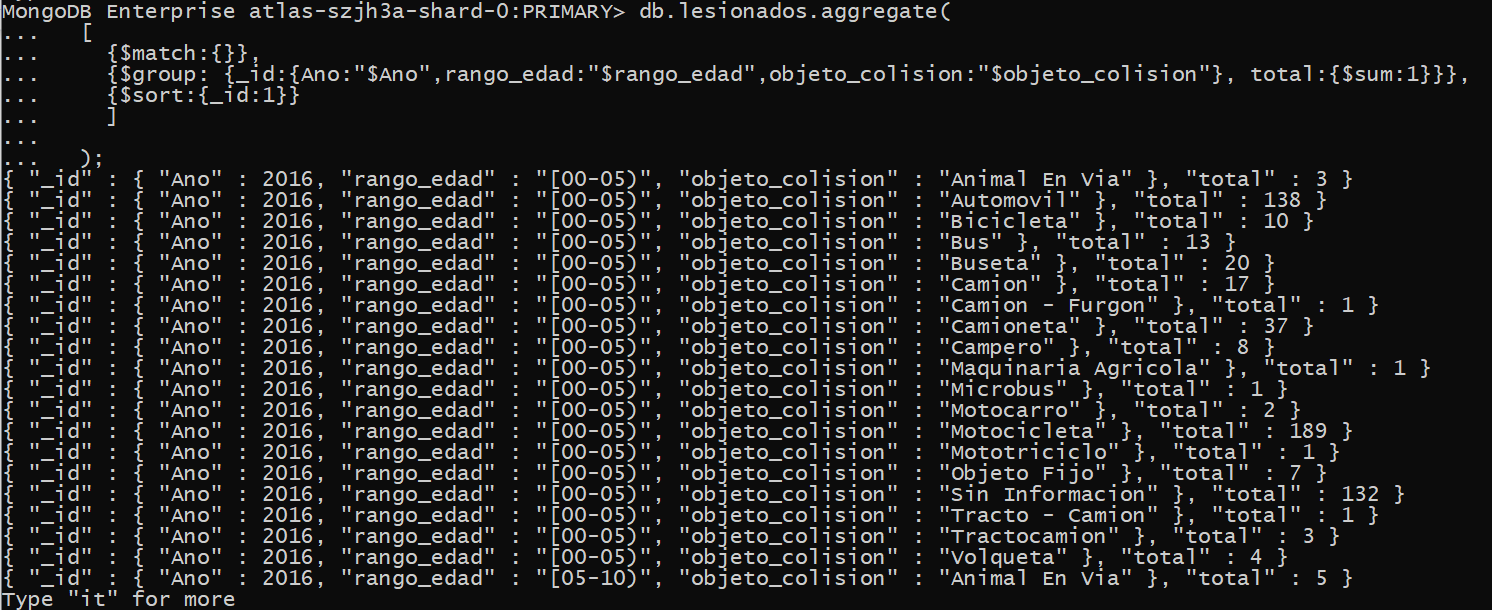
***4.2.1*** Consulta de accidentes por año en la ciudad de Medellín******

***4.2.2*** Consulta de accidentes por año y rango de edad******

***4.2.3*** Consulta de promedio de accidentes por día agrupados por año.

***4.2.4*** Consulta de vehículos involucrados en el accidente del lesionado.******

***4.2.5*** Consulta de lesionados por años según rango de edad y objeto de colisión.

******

# **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

* El modelo relacional tradicional trata, en lo posible, de evadir redundancia de la información y además tiene un esquema rígido. Los modelos NoSQL son contrarios y permiten mucha más flexibilidad. Cuando se intenta hacer una migración, es importante mantener el nuevo modelo semánticamente idéntico al original o lo más cercano posible, para evitar pérdidas o distorsión de la información.
* La interacción de la aplicación que se conecta a la nueva base de datos es afectada significativamente, por lo que se pudieran presentar costos en los cambios de código, necesarios para la adaptación al nuevo sistema de bases de datos.

1. ¿Cuáles limitaciones adicionales se deben considerar, a parte de las mencionadas en el artículo?

A pesar de que las bases de datos NoSQL permiten transacciones para interactuar con las aplicaciones, es altamente limitado e inestable, lo que puede llegar a ser una solución inadecuada para aplicaciones transaccionales complejas.

1. ¿Cuáles son las razones (criterios) que se deben considerar para migrar un conjunto de datos relacionados a NoSQL?

* Cuando hay un alto volumen de datos, el sistema relacional ya no satisface las expectativas de disponibilidad y escalabilidad. Las bases de datos NoSQL tiene un mejor desempeño que las bases de datos relacionales y, debido a esto se pueden usar máquinas de menor rendimiento.
* Factores como velocidad, disponibilidad y escalabilidad son imperativos.

Este es el artículo que les comenté: [A Framework for Migrating Relational Datasets to NoSQL - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S1877050915011758)