**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: 2021.05.22 15:00

*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.Catalina Gallego Yepes - 1.152.451.117 |
| 2.Cesar Augusto Hernandez Rodriguez - 1.152.463.408 |
| 3.Jessica Maria Rivera Jimenez - 1.017.205.993 |
| 4.Jose Julian Mercado Garcia - 1.017.231.636 |
| 5.Julio Cesar Vasquez Bedoya - 1.152.202.499 |
| **REPO EN GITHUB: https://github.com/jcesarvab/Trabajo\_Arq\_Info** |

*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.

El sector financiero en Colombia está conformado por diversas instituciones financieras y sus recursos son administrados por: establecimientos de crédito, sociedades de servicios financieros y otras instituciones financieras.  
Dentro de los establecimientos de crédito, se tiene como principal función canalizar recursos de agentes superavitarios de la economía hacia los deficitarios, mediante captación de fondos para su posterior colocación por medio de préstamos y otras operaciones activas.

Nuestra entidad es un conglomerado de empresas financieras que ofrece un amplio portafolio de productos y servicios financieros, sin embargo este ejercicio se enfocará en el rol de establecimiento de crédito, en el cual se tiene consciencia de que se juega de manera proactiva en un entorno que está en cambio constante y que enfrenta importantes retos económicos, sociales y ambientales, es por ello, que siempre se está buscando la forma de innovar las estrategias desde diferentes aristas para poder mitigar riesgos, mejorar la experiencia de sus clientes y obtener mayor rentabilidad, esto hace tan importante la generación, estructuración y el buen uso de la información para poder llevar a cabo proyectos que satisfagan la misión y visión de la organización.

Nuestro caso de estudio se enfocará en el proceso de otorgamiento, donde deben establecerse para cada uno de los portafolios identificados, las variables que con mayor significancia permitan discriminar los sujetos de crédito que se ajustan al perfil de riesgo de la entidad. La selección de estas variables de discriminación y la importancia relativa que se dé a cada una de ellas debe ser un elemento determinante tanto en el otorgamiento como en el seguimiento de los créditos de cada portafolio.

* 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.

El proceso de originación depende del producto que se esté ofertando, el perfil del cliente y el apetito de riesgo de la compañía, y su ejecución requiere el uso de diferentes herramientas que facilitan la captura y almacenamiento de información no estructurada, sin embargo, estas se encuentran ubicadas en diferentes fuentes que no se comunican entre sí, es decir que no hay una arquitectura común entre los diferentes insumos, lo que implica una mayor dificultad a la hora de consultarlos.

La superintendencia financiera (SFC) mensualmente recolecta de la mayoría de entidades financieras información relacionada a los desembolsos realizados de acuerdo a la modalidad y a ciertas agrupaciones sociodemográficas con el fin de poder analizar el comportamiento de los consumidores colombianos. Dado las restricciones inicialmente mencionadas la generación de este informe es compleja, una base consolidada facilita la oportuna respuesta y evitaría sanciones por parte de la entidad reguladora.

* 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).

Inicialmente se busca tener la información consolidada y homologada con el objetivo de dar respuesta oportuna a los diversos requerimientos legales que recibe la entidad periódicamente, pero además se busca que esta sea la única fuente de consulta para todo el seguimiento que el negocio hace a cada uno de los productos y el comportamiento de sus desembolsos de acuerdo a las diferentes segmentaciones requeridas, y de allí poder identificar posibles alertas tempranas que permitan actuar de manera proactiva ante un posible deterioro de la cartera y no de manera reactiva como se viene haciendo.

* 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.

Para este ejercicio se han seleccionado 3 productos que son desembolsados mediante solicitudes de crédito, esto implica que se debe realizar un estudio previo a la aprobación. De acuerdo al monto del desembolso y al producto, cada uno de estos queda en una fuente y un formato distinto, por lo que a hoy se debe hacer el seguimiento a cada uno de ellos por separado ya que no todas las fuentes capturan la misma información y las consultas e informes no son homologables.

El reporte al ente regulador se debe enviar mensualmente los primeros 5 días calendario del mes y debido a la diversidad de los datos y a pesar de que cada mes se hace lo mismo, este toma tiempo por parte de los analistas encargados de realizarlo, además de los posibles errores dadas las manualidades involucradas que podrían acarrear sanciones económicas.

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

Para obtener los datos para la realización de este ejercicio fue necesario impactar a las áreas de la entidad encargadas de la administración de cada una de las herramientas donde estos se almacenan.

Se solicitó una muestra de los datos del año 2019 y se recibieron en formato excel y en archivos separado por comas, dado que cada archivo tenía una fuente y estructura diferente fue necesario definir un estándar para los tipos de datos, identificar los campos valiosos para este ejercicio y realizar el enmascaramiento de los datos sensibles de los clientes. Todo lo anterior se realizó con Microsoft Access y se aplicó a cada uno de los archivos recibidos.

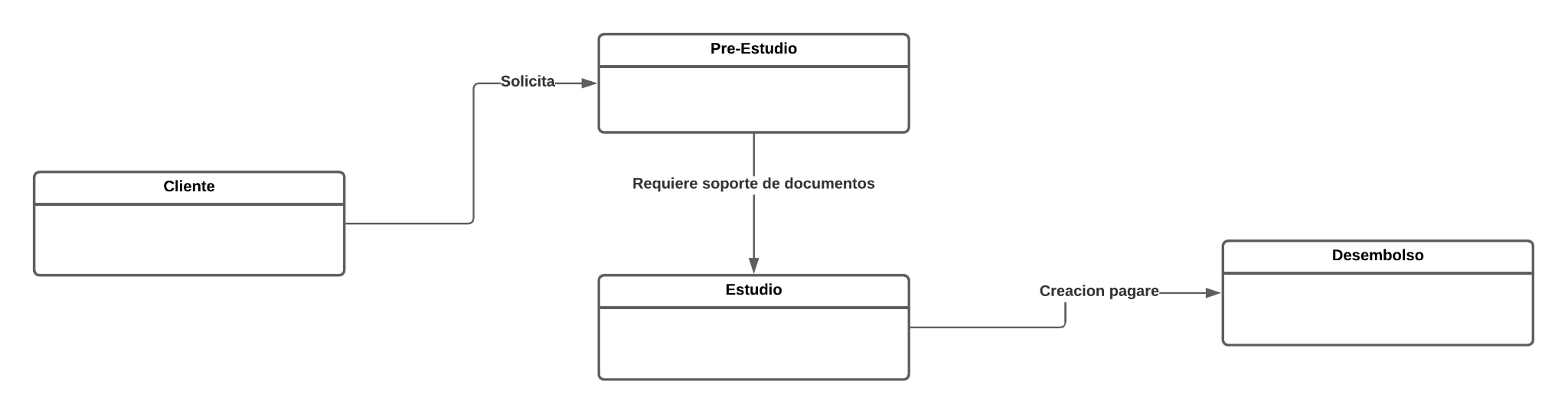
* 1. **Descripción de datos (diccionario):**

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

|  |  |  |  |
| --- | --- | --- | --- |
| Tabla | Nombre del atributo / variable | Formato o Tipo de Dato | Descripción |
| Saldos | id | string | Identifica el documento del cliente |
| Saldos | tid | double | Se identifica el tipo de documento del cliente |
| Saldos | apl | string | aplicativo al que corresponde la información de los diferentes productos |
| Saldos | obl341 | string | número correspondiente a la obligación y es considerado una llave de cruce |
| Saldos | clf | string | Modalidad |
| Saldos | altmora | double | Días de mora de la obligación |
| Saldos | sk | double | Saldo Capital |
| Saldos | cv1 | double | Cartera vencida |
| Saldos | vdesem | double | Valor desembolsado |
| Saldos | fdesem | timestamp | Fecha del desembolso |
| Clientes | num\_doc | string | Identifica el documento del cliente |
| Clientes | cod\_tipo\_doc | string | Llave correspondiente al tipo de documento |
| Clientes | genero\_cli | string | Corresponde al género del cliente |
| Clientes | ano\_nac\_cli | int | Corresponde al año de nacimiento del cliente |
| Clientes | cod\_segm | string | Código del segmento del cliente |
| Clientes | dpto\_of | string | Departamento del cliente |
| Clientes | nivel\_academico | string | Nivel académico del cliente |
| Clientes | ing\_mes | double | Corresponde a los ingresos mensuales del cliente |
| Clientes | estado\_civil | string | Corresponde al estado civil del cliente |
| Pre\_estudio | nro\_solicitud | string | Llave generada por la herramienta para identificar la solicitud |
| Pre\_estudio | nro\_doc | string | Numero de documento del cliente |
| Pre\_estudio | tipo\_de\_documento | string | Tipo de documento del cliente |
| Pre\_estudio | relacion\_ppte | string | Relación del solicitante (deudor, codeudor) |
| Pre\_estudio | fecha | timestamp | Fecha de radicación de la solicitud |
| Pre\_estudio | marca | bigint | Marca de la solicitud |
| Pre\_estudio | decision\_final | string | Resultado del procesamiento de la solicitud que entrega la herramienta |
| Pre\_estudio | canal\_venta | int | Canal de venta del producto |
| Pre\_estudio | pla\_max | int | Plazo máximo para la financiación |
| Pre\_estudio | plazo\_solicitado | int | Plazo solicitado por el cliente |
| Pre\_estudio | productcategory | string | Categoría del producto |
| Pre\_estudio | pdto | string | Producto solicitado |
| Pre\_estudio | val\_financiar\_total | double | Valor a financiar |
| Pre\_estudio | monto\_solicitado | double | Monto solicitado por el cliente |
| Pre\_estudio | monto\_aprobado | double | Monto aprobado por la herramienta |
| Pre\_estudio | tipo\_de\_participante | string | Tipo de participante |
| Pre\_estudio | tipo\_vivienda | string | Tipo de vivienda donde habita el cliente |
| Pre\_estudio | personas\_a\_cargo | int | Cantidad de personas a cargo que tiene el cliente |
| Pre\_estudio | tipo\_contrato | string | Tipo de contrato laboral que tiene el cliente |
| Pre\_estudio | ocupacion | string | Ocupacion del cliente |
| Pre\_estudio | pago\_arrendmto | double | Valor reportado por concepto de pago de arrendamiento |
| Pre\_estudio | gtos\_fmliares | double | Valor reportado por concepto de gastos familiares |
| Pre\_estudio | gastos\_fam\_calcu | double | Valor calculado por la herramienta de gastos familiares |
| Pre\_estudio | egresos\_solicitante | double | Total de egresos mensuales del cliente |
| Estudio | num\_solicitud\_radicado | string | Llave generada por la herramienta para identificar la solicitud |
| Estudio | num\_radicado\_exp | string | Llave foranea de cruce con la información de Pre Estudio |
| Estudio | tipo\_doc | string | Tipo de documento del cliente |
| Estudio | num\_doc | string | Numero de documento del cliente |
| Estudio | estado\_solicitud\_actual | string | Estado |
| Estudio | tipo\_solicitante | string | Tipo de solicitante |
| Estudio | tipo\_proceso | string | Tipo de proceso interno |
| Estudio | f\_radm | int | Fecha de radicación |
| Estudio | f | int | Fecha de aprobación |
| Estudio | Obl | String | num obligacion |
| Estudio | estrategia | string | Identifica el producto |

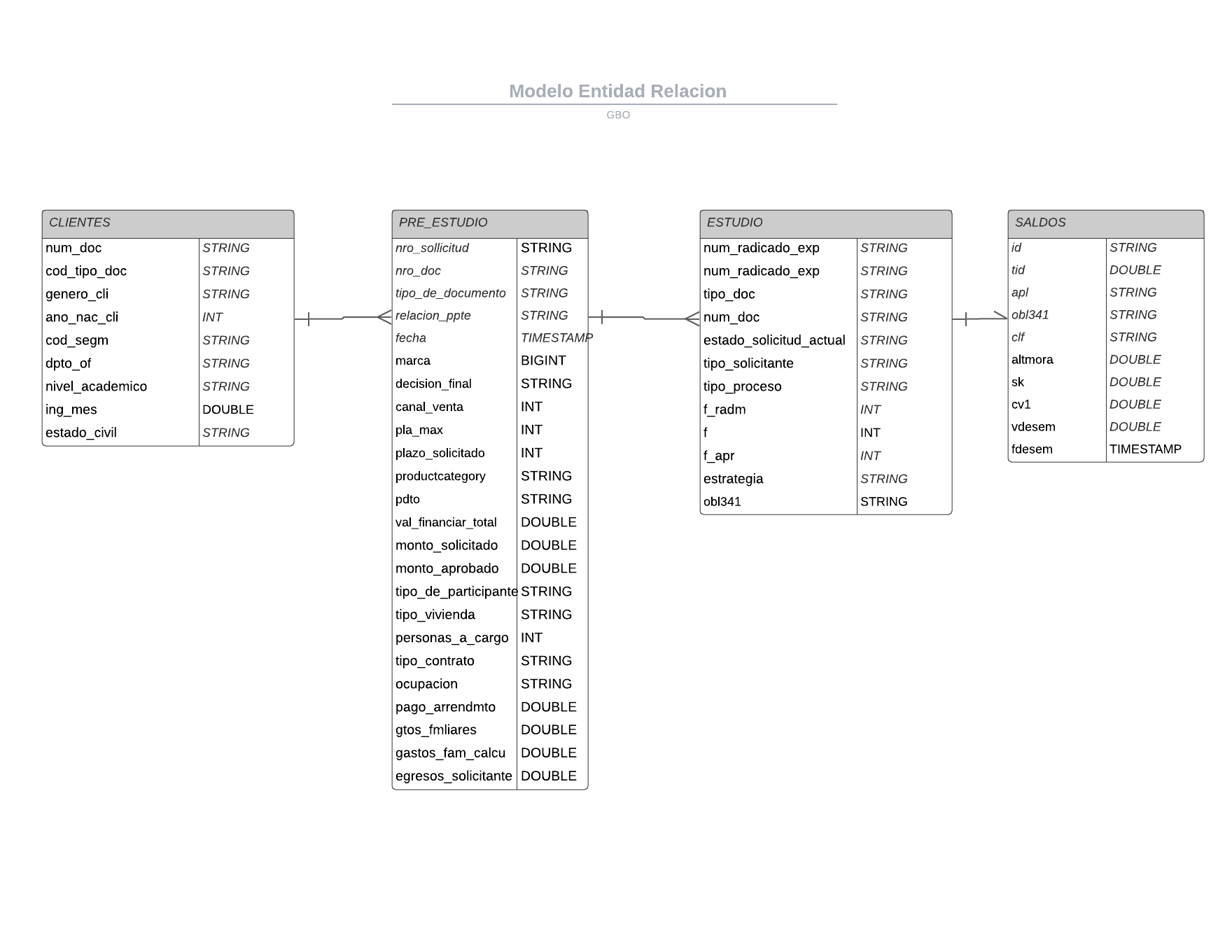
* 1. **Modelo del domino**

***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***

* 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*

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*

El código incluido T1.4.1.Creacion\_Documentos.txt correspondiente a la creación de las 4 tablas, incluye la inclusión de 20 documentos por tabla, sin embargo para realizar el análisis del punto siguiente, se cargaron los archivos completos mediante la herramienta de MongoDB Compass.

* 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 T1.4.2.Consultar\_Datos.sql*

1. **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?

R/ Las limitaciones que se infieren de la lectura son:

* La realización de consultas frecuentes se dificulta pues a diferencia de una base de datos relacional, en una *NoSQL* no se posee un índice por el cual consultar de manera ágil.
* La migración de un sistema relacional a uno *NoSQL*, tal como *Graph Databases*, implica la gestión de un lenguaje de consulta diferente, que aunque tiene un origen *SQL*, debe ser aprendido y ejecutado.
* Dado que muchos conjuntos de datos relacionales no están diseñados u optimizados para exportar grandes cantidades de datos dentro de períodos cortos de tiempo, la migración directa de un *RDBMS* a *Graph* puede implicar retrasos en el proceso.

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

R/ Algunas limitaciones que deben ser consideradas de migrar de datos relacionales a NoSQl son:

* Los sistemas NoSQL tienen oportunidades en la Atomicidad, propiedad que asegura que en las transacciones, se ejecuten todas las actividades o no se ejecute ninguna y no ocurran ejecuciones parciales en la base.

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

R/ Las principales razones son:

- Los sistemas de datos relacionados no están diseñados para almacenar robustamente las relaciones entre diferentes elementos, lo que implica que deban ser utilizados múltiples *JOINs*, agregando un nivel de complejidad y costo computacional e incrementando los tiempos de respuesta del modelo.

Lo anterior se agrava por el hecho de que en la actualidad la información es altamente conectada y está creciendo en volumen, velocidad y variedad, y por lo tanto las relaciones están creciendo incluso con mayor velocidad.

- La gestión de datos relacionados se ve impactada por el hecho de que el negocio y/o los usuarios están constantemente realizando requerimientos para modificar la arquitectura de los datos, sin embargo estos cambios son problemáticos para los desarrolladores, implican un alto costo de mantenimiento y afectan la agilidad del proceso, ya que el rediseño de los esquemas puede ser lento y costoso.