**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: 2021.05.29 10:00

*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.Luisa Fernanda Arias Giron-1053827318** |
| **2.Juan Sebastián Correa Acevedo -1020479653** |
| **3.Gloria Stephany Gomez Gomez-1053859173** |
| **4.Carolina Taborda Rua-43265755** |
| **5.Mario Varas González - 712484** |
| **REPO EN GITHUB:** |

*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**
  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.

WOM es una empresa prestadora de servicio de telefonía móvil en Colombia, para poder dar un buen soporte al servicio y para administrar la actividad comercial, la empresa genera información tanto en elementos de red como en el OSS (Operations Support Systems) que básicamente es el sistema que soporta la operación del negocio en telecomunicaciones; Del OSS podemos obtener DB de facturación, recargas, CRM, consumos de datos, sms, voz y roaming.

El proveedor fullsatck de la empresa es Whale cloud y los elementos de red son Huawei, para nuestro trabajo, emplearemos las DB de provistas por el OSS únicamente

* 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.

Las empresas de telecomunicaciones se ven altamente afectadas por el fraude, ya sea al momento de la venta, así como también en el mal uso del servicio, con fines de no pago y/o beneficio particular al margen de la ley, tales como, suplantación de identidad, reoriginamiento, altos consumos internacionales, desviación de llamadas, entre otros

* 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).

* Diseñar una solución que genere alarmas preventivas
* Generar información para realizar bloqueos
* Alimentar DB con patrones de posibles casos de fraudes
* Utilizar la DB de posibles fraudes como insumo para trabajar en soluciones de AI
  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 Wom por ser una empresa nueva en el mercado colombiano, no cuenta aún con controles, alarmas, ni demás estrategias que le ayuden a prevenir y controlar los posibles casos de fraudes en consumo y suscripción

* **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 imprímalo(s). Llame el archivo T1.2.1.Datos.zip

Tomamos tres bases de datos del datawarehouse, por medio de conexión a base de datos, la primera db trae información de los consumos, la segunda de clientes y la tercera de facturación

Tuvimos como inconvenientes que la empresa aún lleva muy poco en el mercado y aún no hay un conocimiento de toda la información que arrojan los sistemas y plataformas por lo que toco emplear más tiempo de lo esperado para obtener la información esperada

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

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

**FACTURACION**

|  |  |  |
| --- | --- | --- |
| Nombre del atributo / variable | Formato o Tipo de Dato | Descripcion |
| FACTURA\_CODIGO | Entero | Código interno dado a cada factura al momento de generarse |
| FACTURA\_FECHA | 3/04/2021 | Fecha en la que se genera la factura |
| FACTURA\_TIPO | Varchar | Motivo de generar factura |
| SUBSCRIBER\_ID | Entero | Código interno del servicio del cliente |
| CUENTA\_CODIGO | Entero | Id de cliente en el sistema |
| FACTURA\_MONTO | Boleado | Valor factura |
| FACTURA\_FECHA\_VENCIMIENTO | 05-Apr-21 | Fecha límite de pago |
| CONSULTOR\_CODIGO | varchar | Nombre de quien genera la factura en caso de ser en punto de atención |

**CONSUMO**

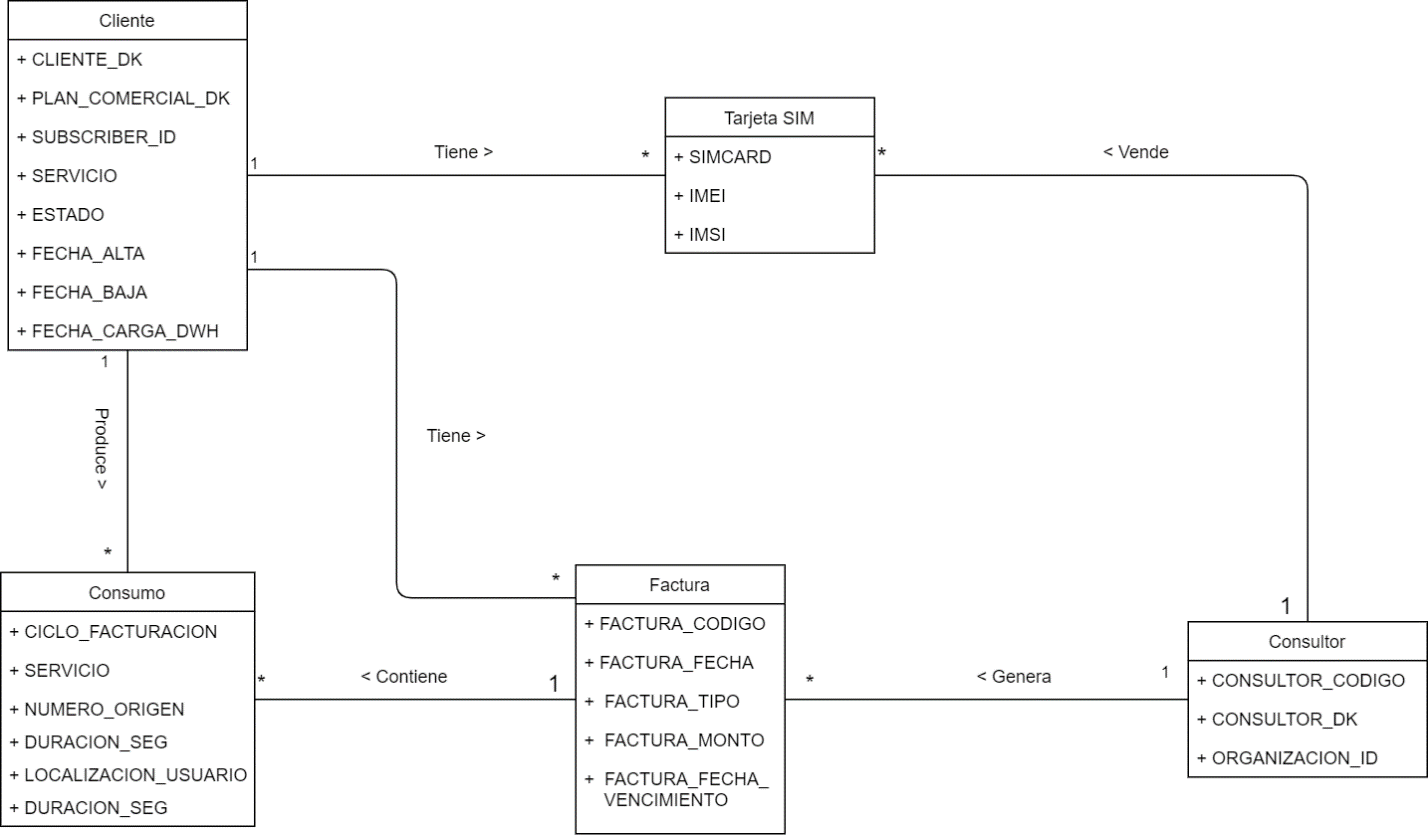
|  |  |  |
| --- | --- | --- |
| Nombre del atributo / variable | Formato o Tipo de Dato | Descripción |
| PERIODO\_PROCESO\_CODIGO | 20210425 | Fecha en que genera factura |
| CLIENTE\_DK | Entero | Id de cliente en el sistema |
| SUBSCRIBER\_ID | 1130266408 | Código interno del servicio del cliente |
| CICLO\_FACTURACION | Entero | Código de periodo en que factura el cliente |
| SERVICIO | varchar | Prepago o pospago |
| NUMERO\_ORIGEN | 3127753935 | teléfono del cliente |
| DURACION\_SEG | 1888 | Tiempo en segundos |
| LOCALIZACION\_USUARIO | 0137322114152bf4 | coordenadas |
| PLMNID\_SGSN | Entero | Antena1 |
| PLMNID\_GGSN | Entero | Antena2 |
| UP\_DATA\_BYTES | Booleano | Velocidad subida |
| DOWN\_DATA\_BYTES | Booleano | Velocidad bajada |
| VALOR1 | Booleano | Valor consumo en bolsa 1 |
| VALOR2 | Booleano | Valor consumo en bolsa 2 |
| BYTES1 | Booleano | Total, consumo en bolsa 1 |
| BYTES2 | Booleano | Total, consumo en bolsa 2 |

**CIENTES**

|  |  |  |
| --- | --- | --- |
| Nombre del atributo / variable | Formato o Tipo de Dato | Descripción |
| CLIENTE\_DK | Entero | Id de cliente en el sistema |
| PLAN\_COMERCIAL\_DK | Entero | Código interno dado a cada plan |
| CONSULTOR\_DK | Entero | Id vendedor |
| SUBSCRIBER\_ID | Entero | Código interno del servicio del cliente |
| ORGANIZACION\_ID | Entero | Código interno de la empresa del vendedor |
| STAFF\_ID | Entero | Código interno del lugar de venta |
| MSISDN | Entero | Numero uno de |
| SIMCARD | Entero | Código de la tarjeta sim |
| IMEI | Entero | Numero de la tarjeta sim |
| IMSI | Entero | Numero telefónico asignado por la empresa |
| CICLO\_FACTURACION | Entero | Fecha de factura |
| SERVICIO | Varchar | Prepago o pospago |
| ESTADO | Varchar | Estado del servicio activo, suspendido retirado |
| MOVIMIENTO\_NOMBRE | Varchar | Tipo transacción, activación, retiro portabilidad |
| MOVIMIENTO\_TIPO\_NOMBRE | Varchar | Cliente nuevo o portado |
| OPERADOR\_PORTA\_ORIGEN | Varchar | operador si es portado |
| OPERADOR\_PORTA\_DESTINO | Varchar | operador si es portado |
| COMISIONABLE | entero | 0 No 1 Si, se paga por venta |
| FECHA\_ALTA | 22/04/2021 | Fecha activación |
| FECHA\_BAJA |  | Fecha retiro |
| FECHA\_CARGA\_DWH | 27/04/2021 | Fecha que se carga al sistema |

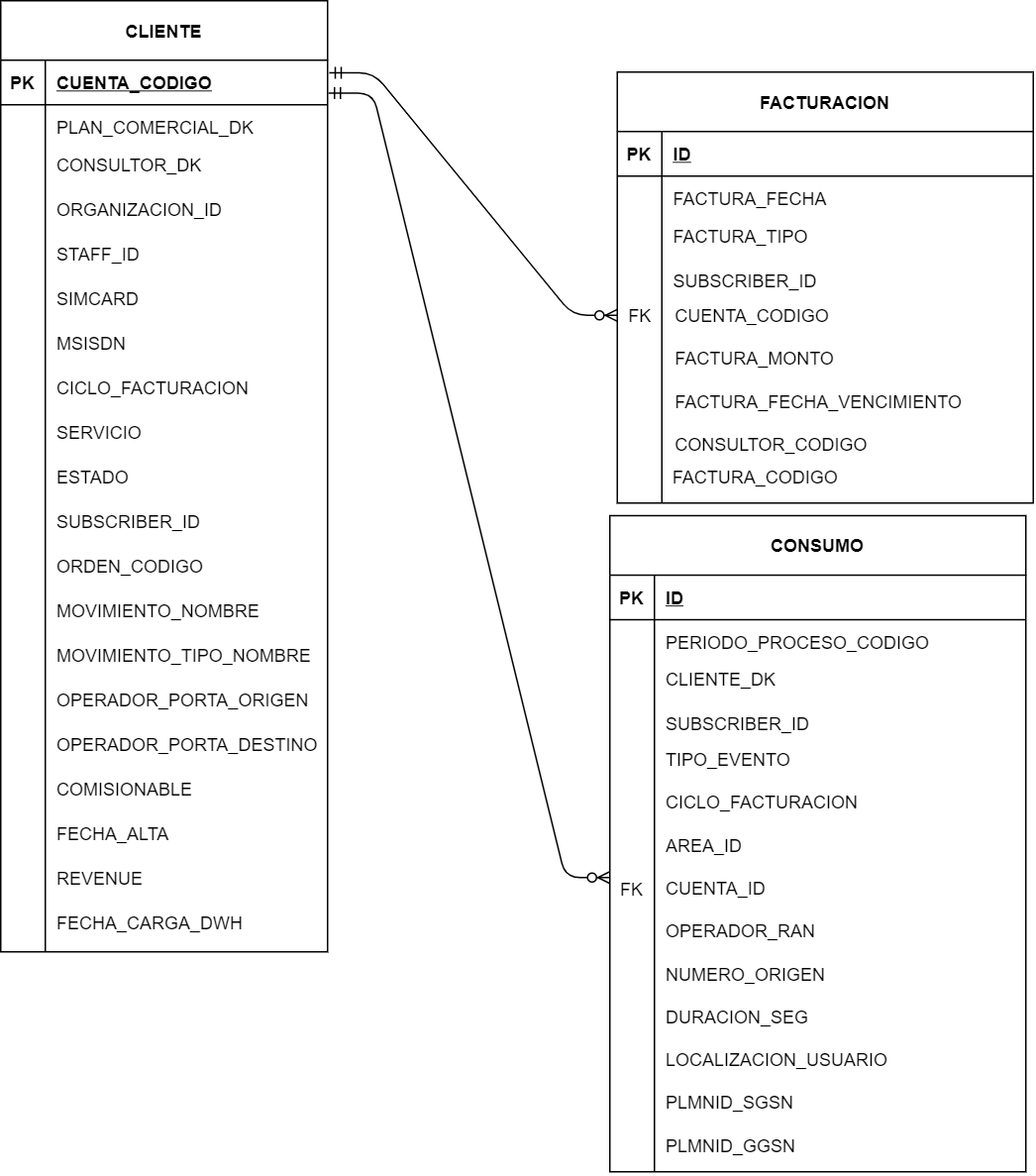
* 1. **Modelo del domino**

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



* **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**

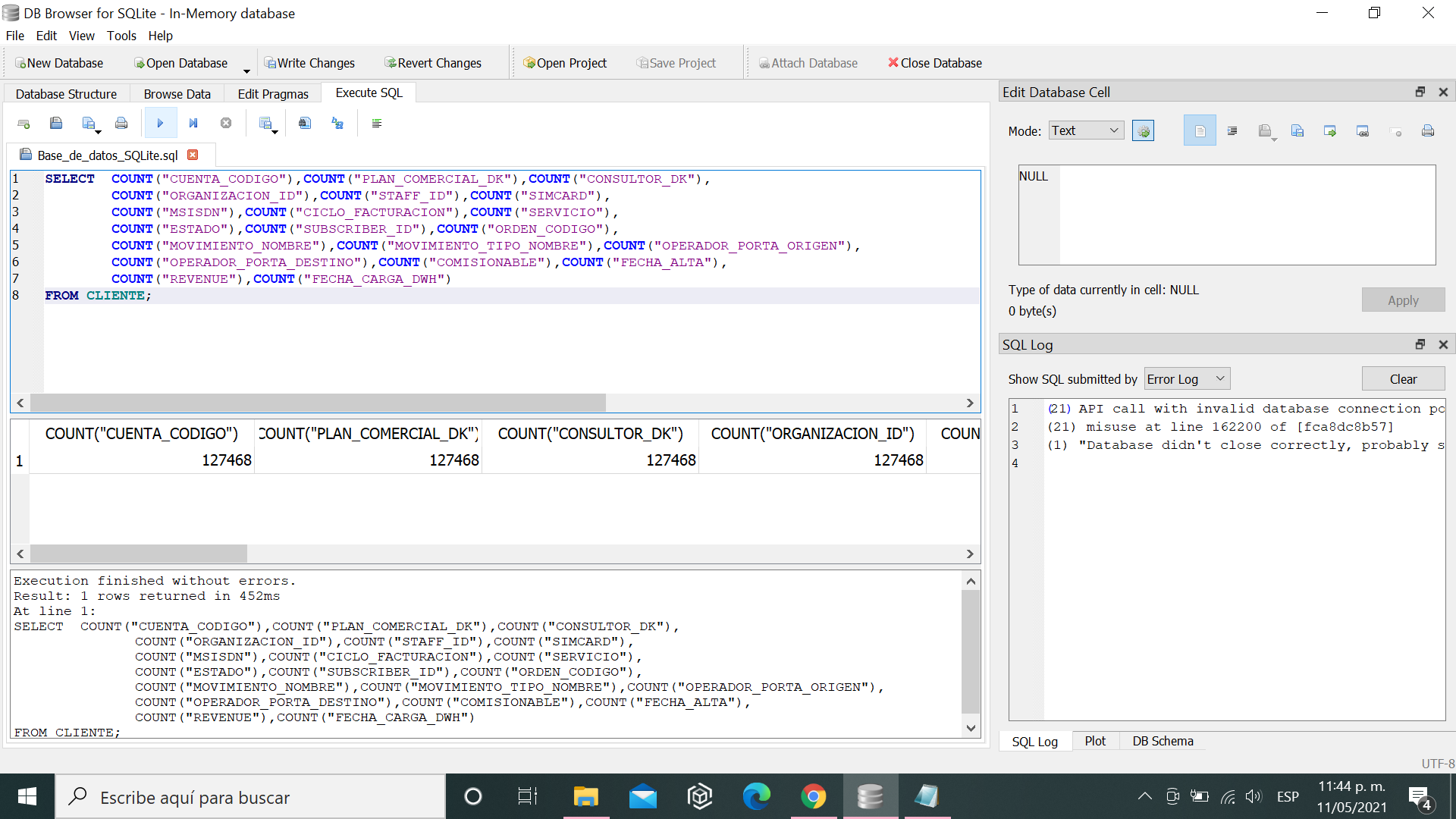
***Observación****: Escriba el código para insertar los datos en cada una de las tablas creadas. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.3.Insertar\_Datos.sql*

* 1. **Sentencia de consulta**

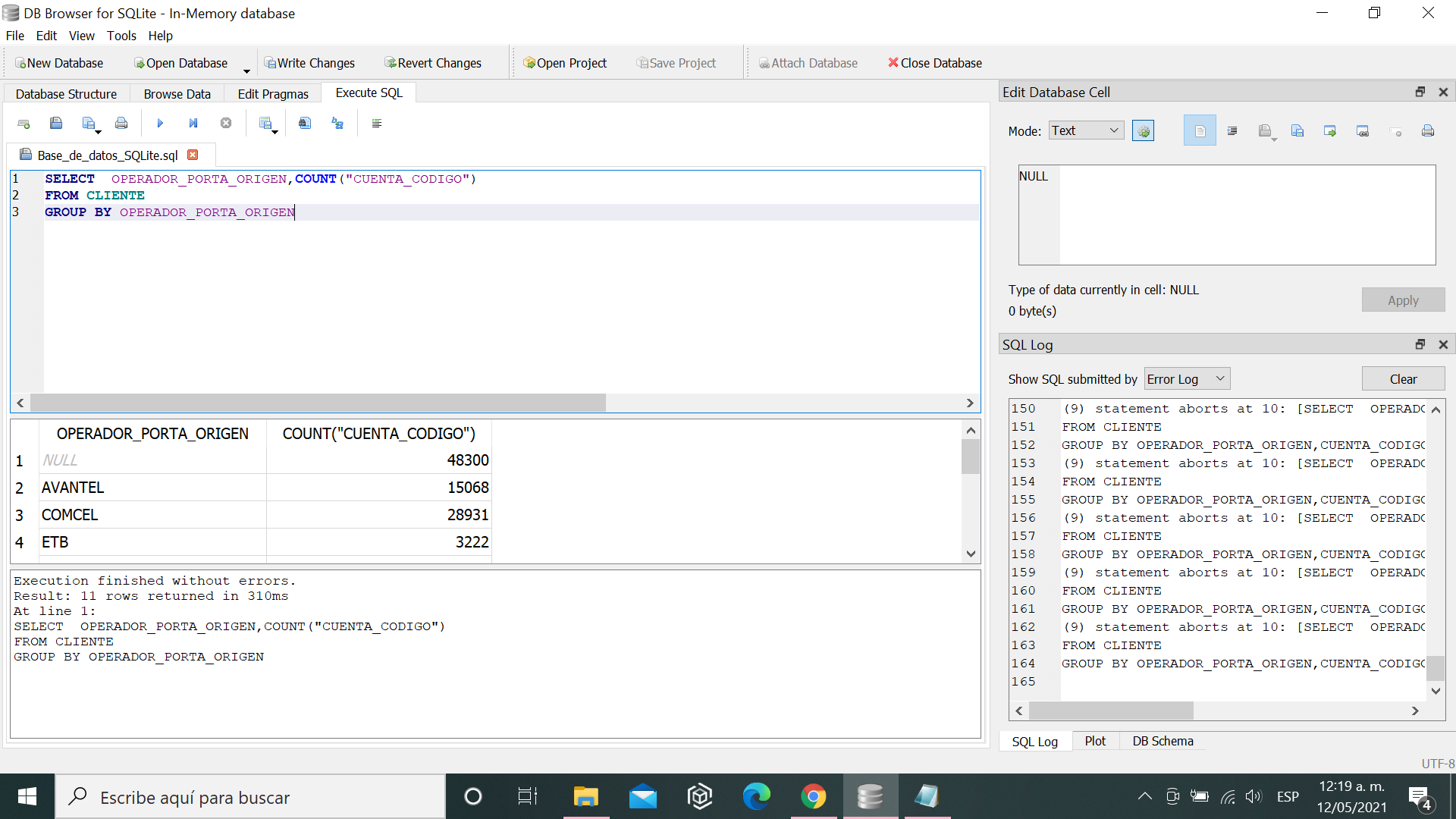
***Observación****: realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de 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*

**TABLA CLIENTE**: Estadística descriptiva

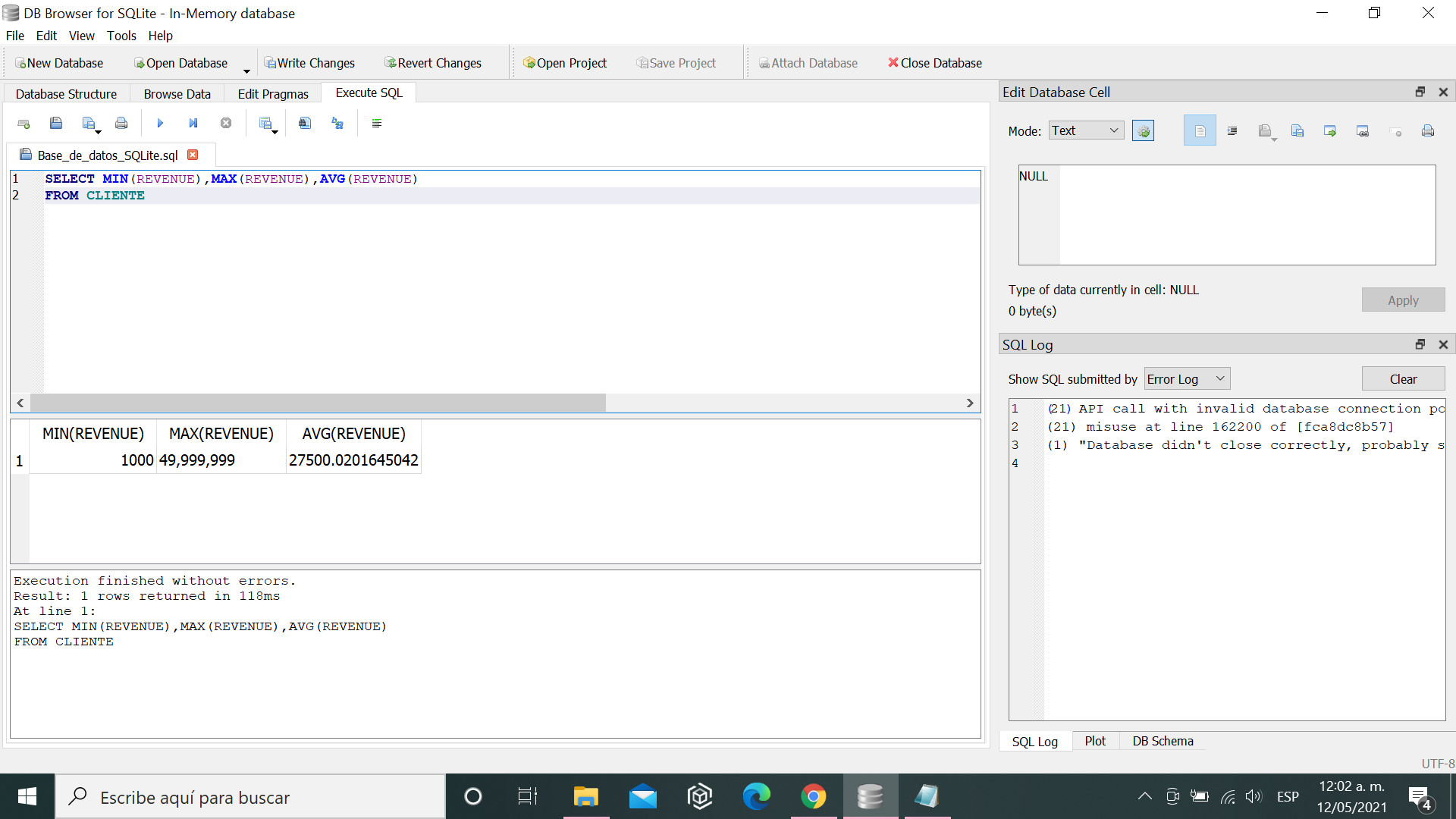
* Conteo total, por columnas



* Conteo por categoría, cantidad de cuentas que tienen registradas en los distintos operadores.

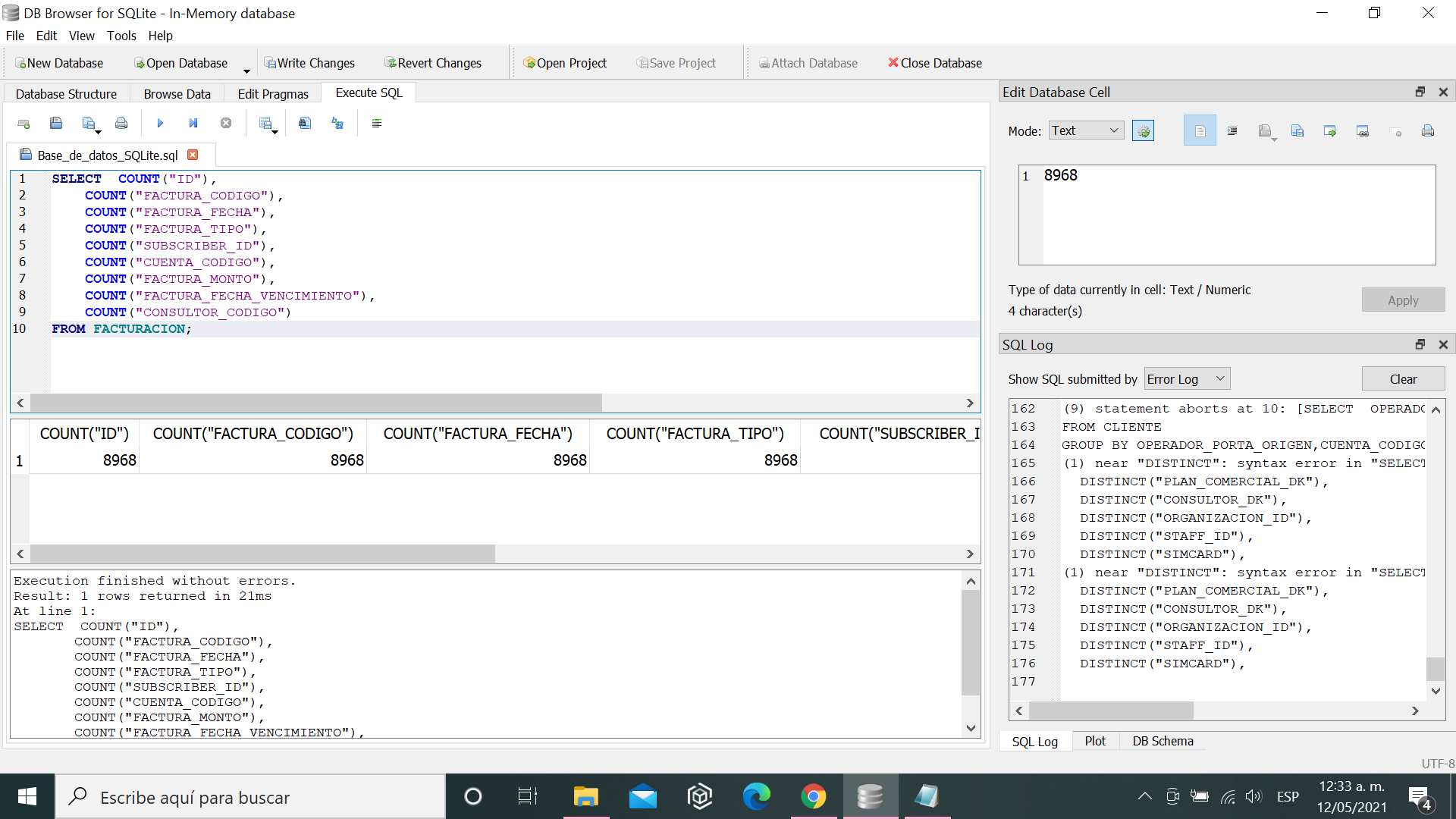


* Máximo, mínimo y promedio de la columna REVENUE

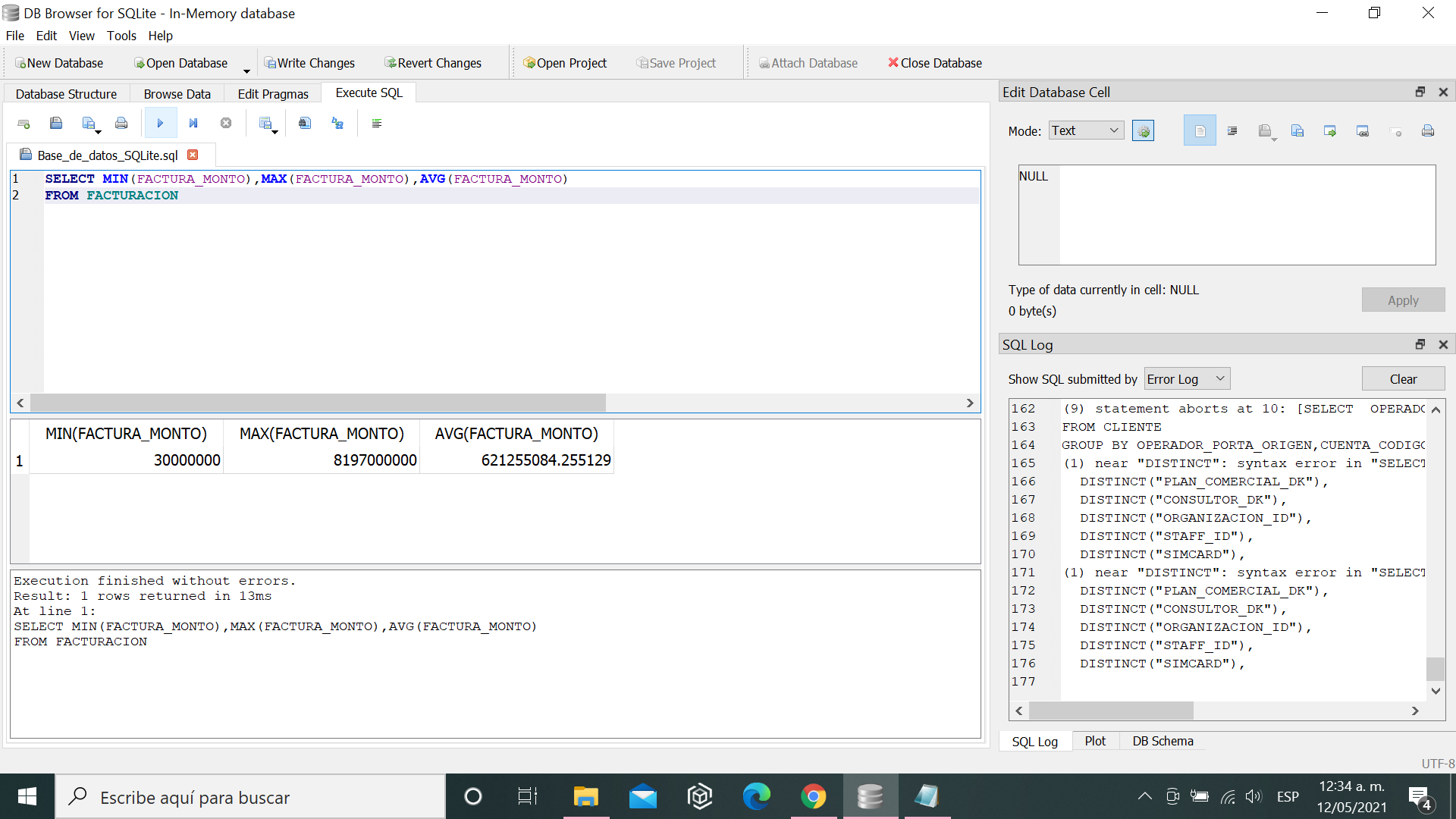


**TABLA FACTURACIÓN**: Estadística descriptiva

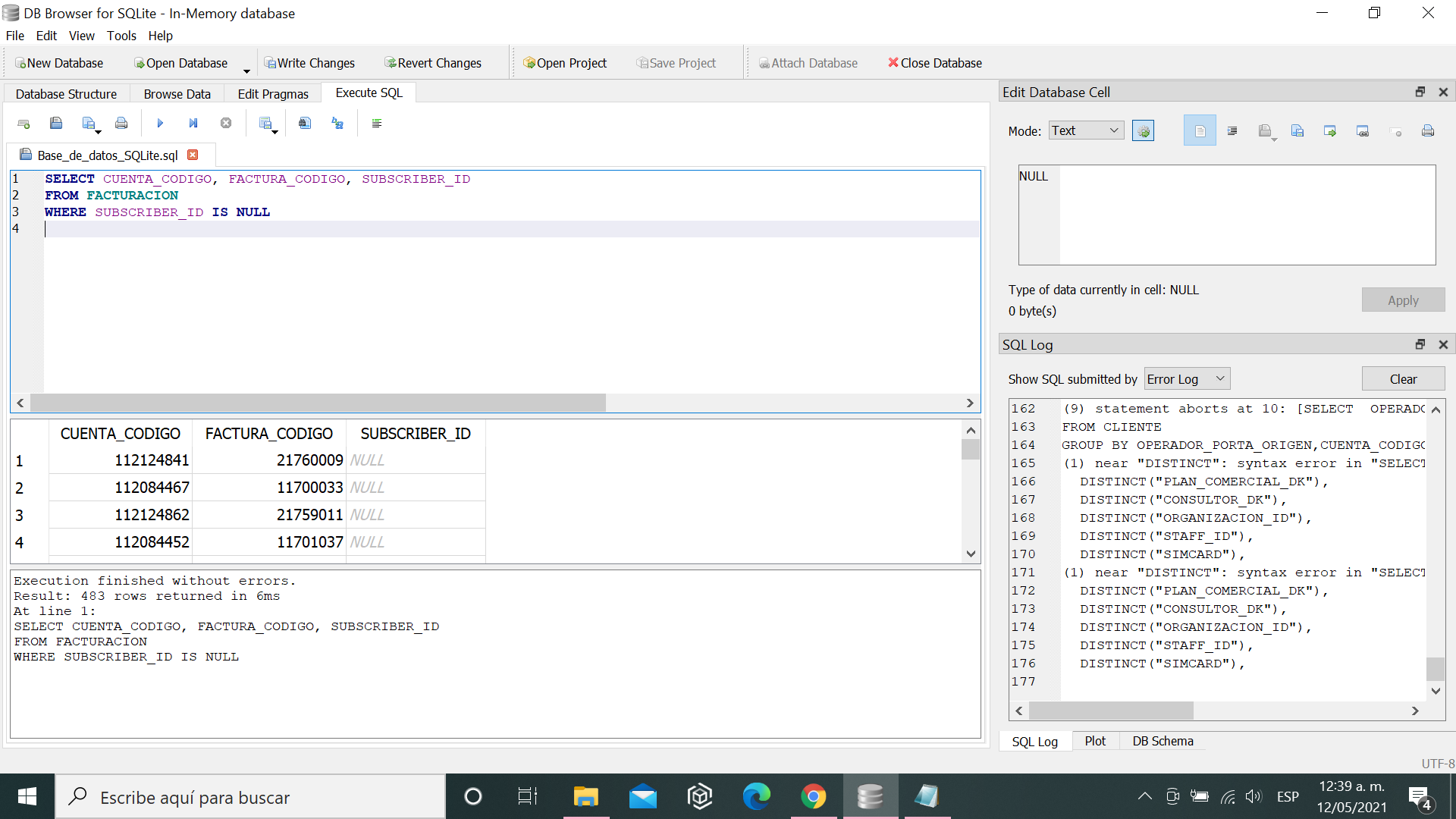
* Conteo total, por columnas



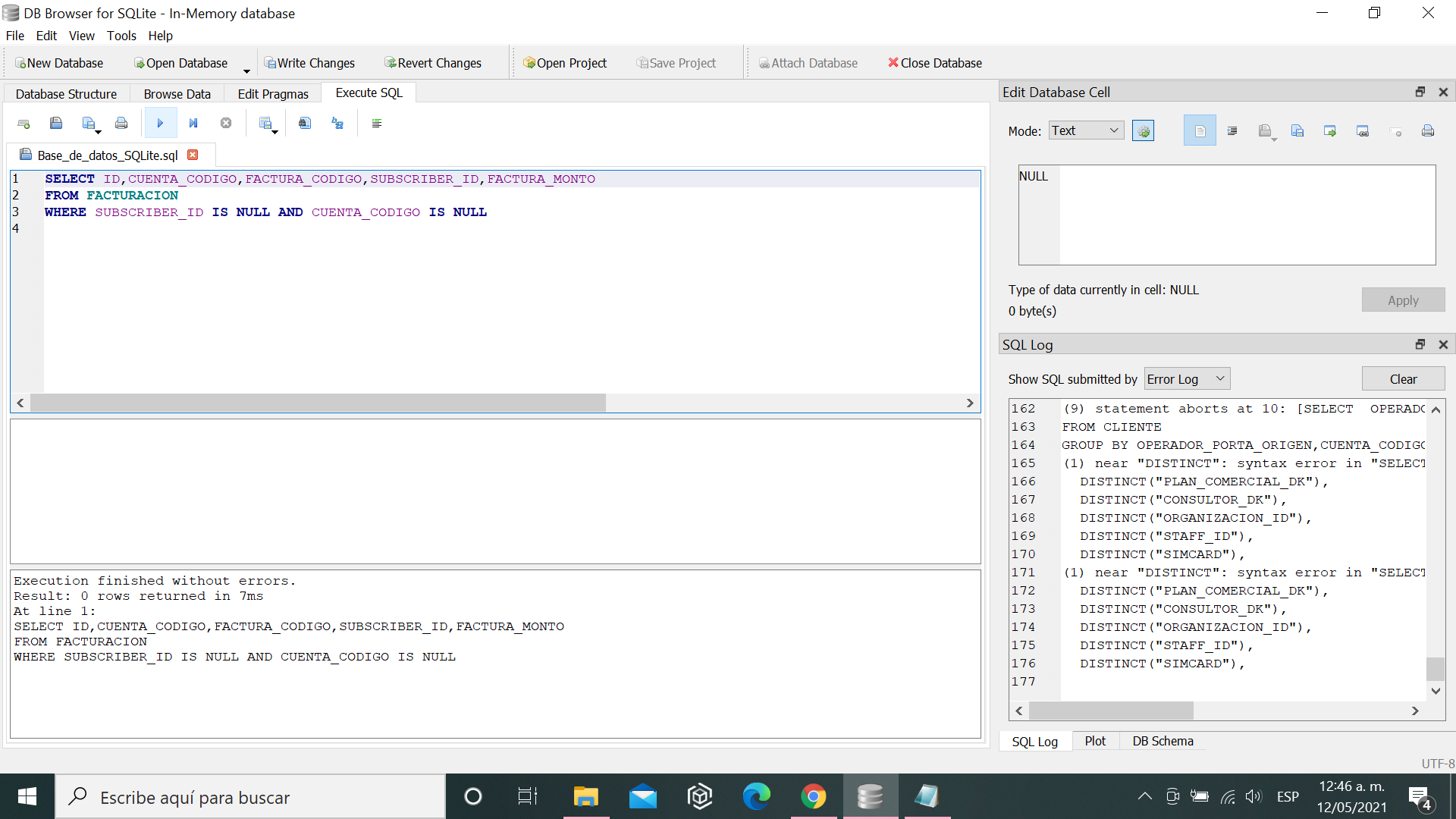
* Máximo, mínimo y promedio de la columna FACTURA\_MONTO



* Cuentas que no tienen código interno de servicio al cliente (SUSCRIBER\_ID)

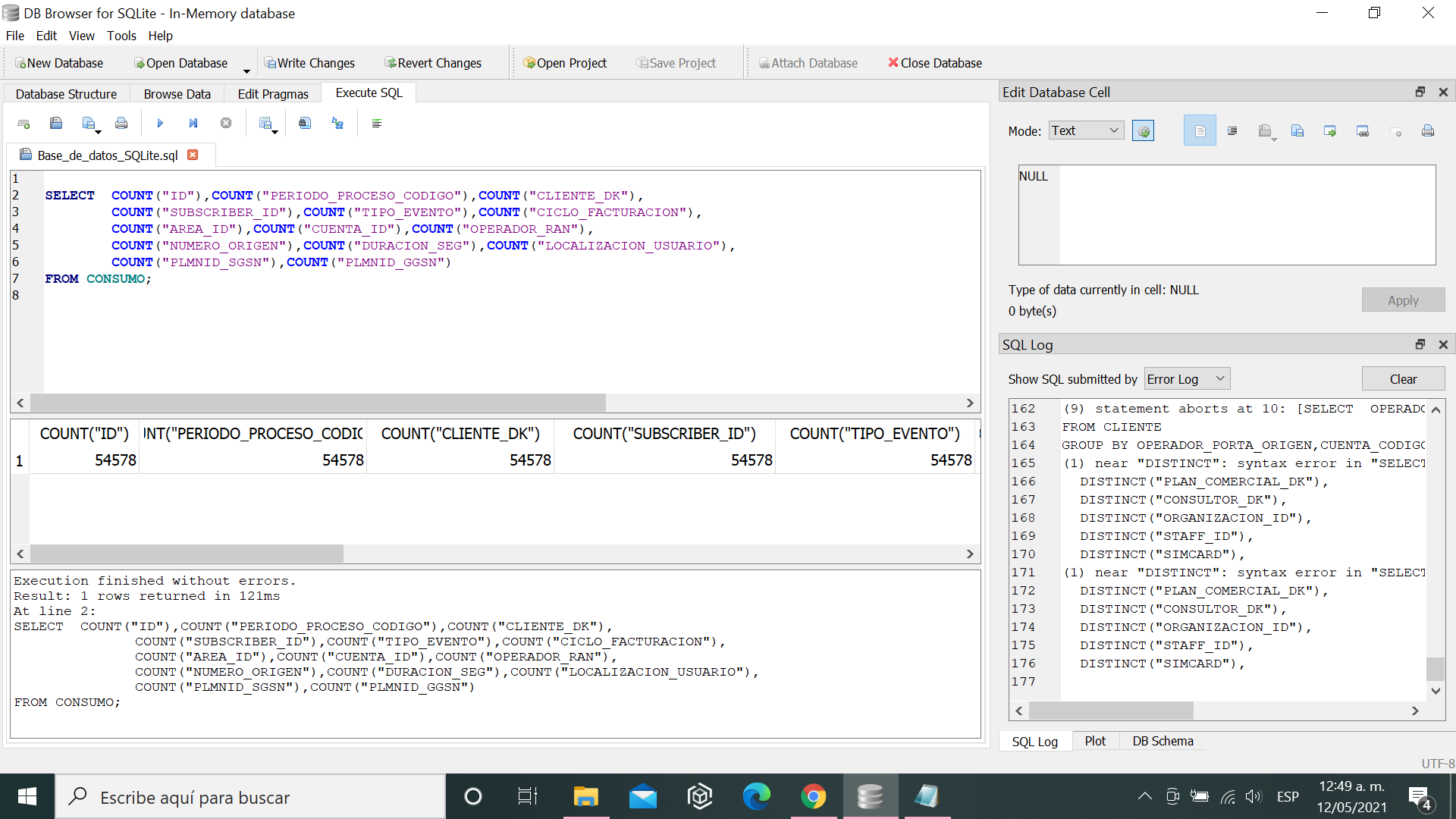


* Códigos de facturas que tienen cuentas Nulas sin código interno de servicio pero sí monto: Debido a la limpieza de los datos realizada previamente se borraron los registros de estas cuentas. Por lo que el resultado es ninguno.

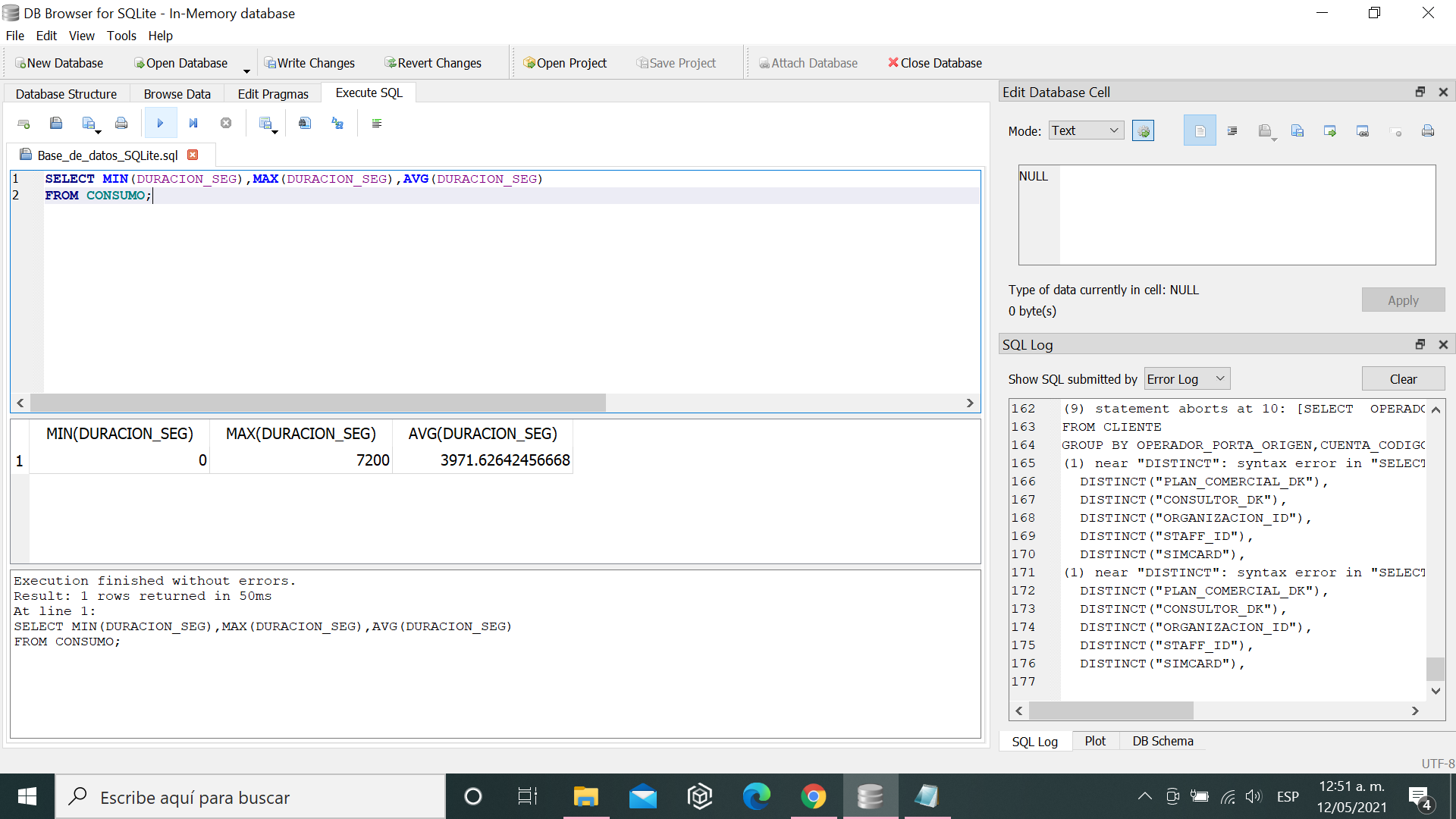


**TABLA CONSUMO**: Estadística descriptiva

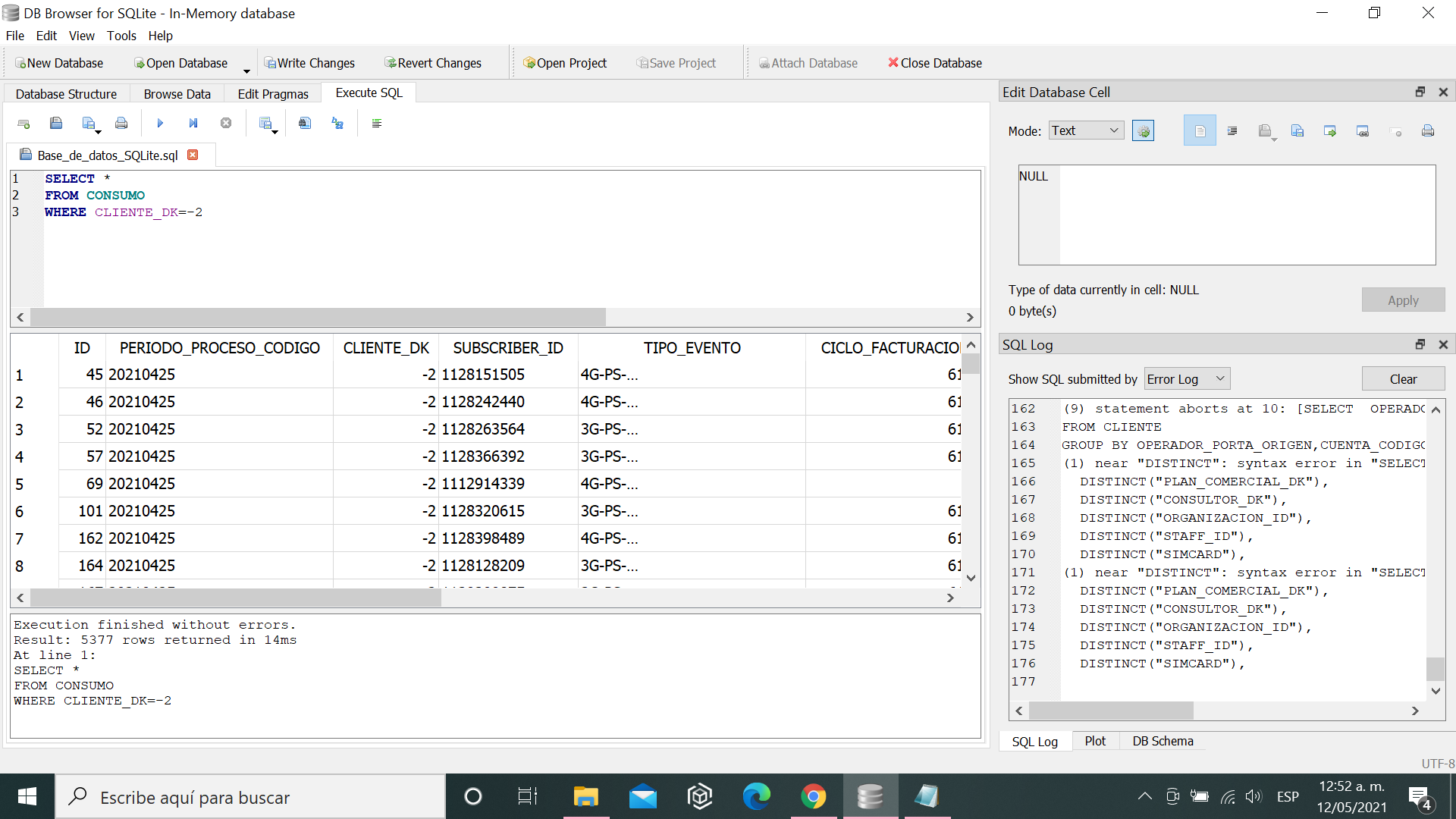
* Conteo total, por columnas



* Máximo, mínimo y promedio de la columna DURACION\_EN\_SEG



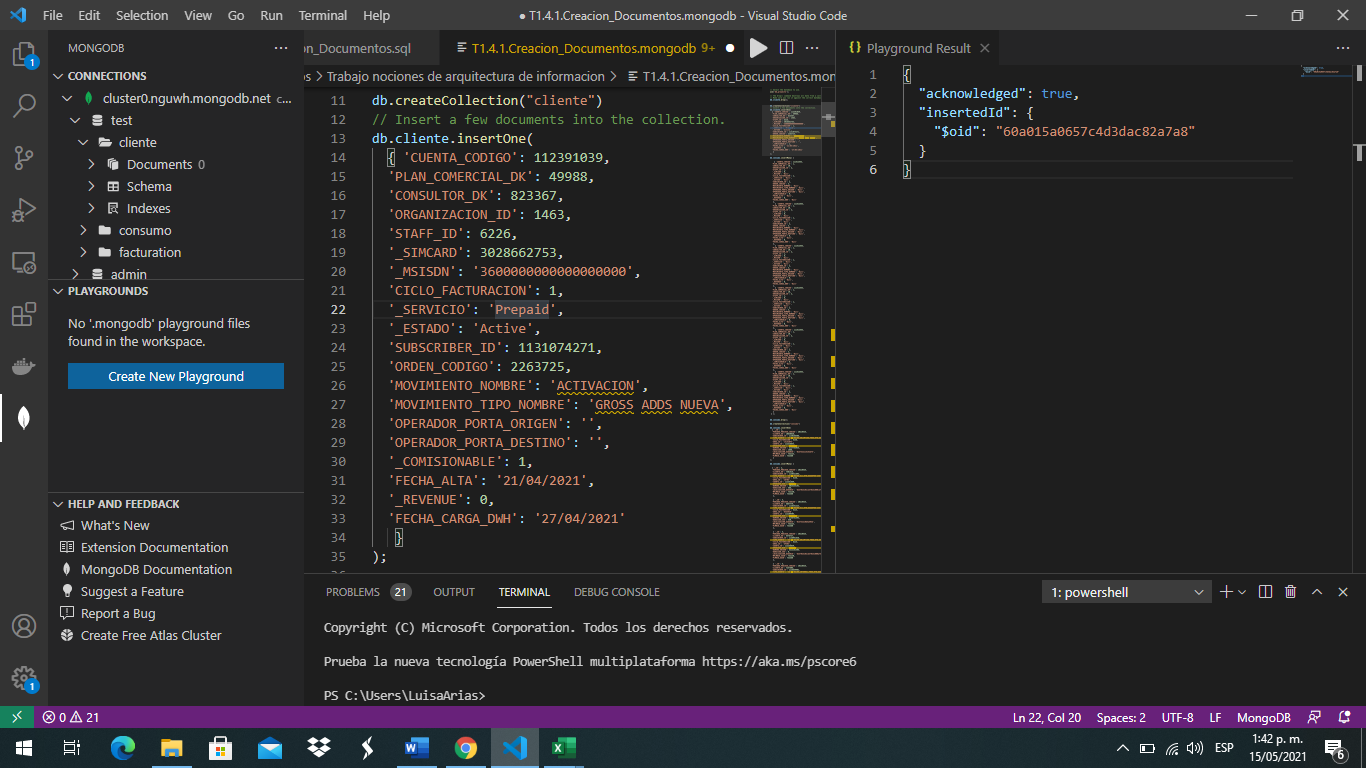
* Descripción de las columnas donde CLIENTE\_DK=-2

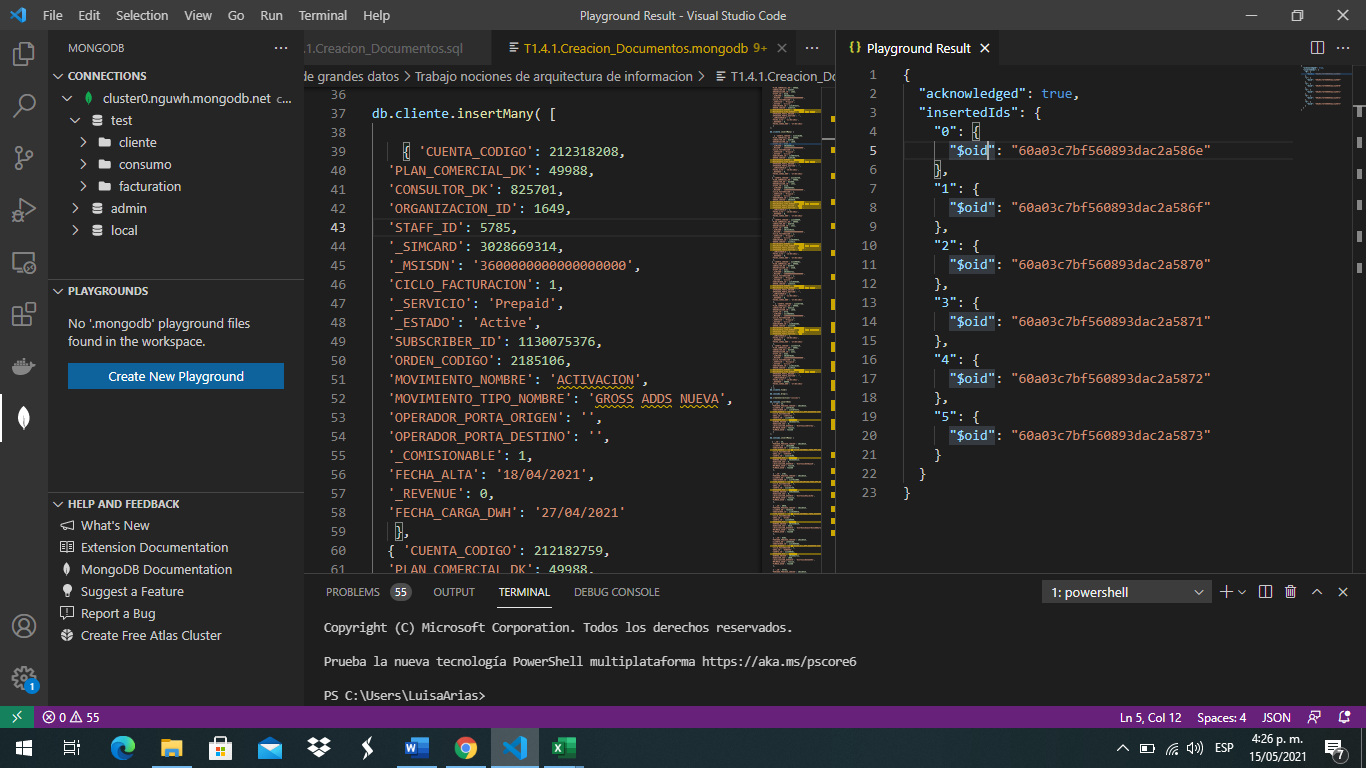


* **MongoDB**
  1. **Sentencia o consulta de creación del documento(s)**

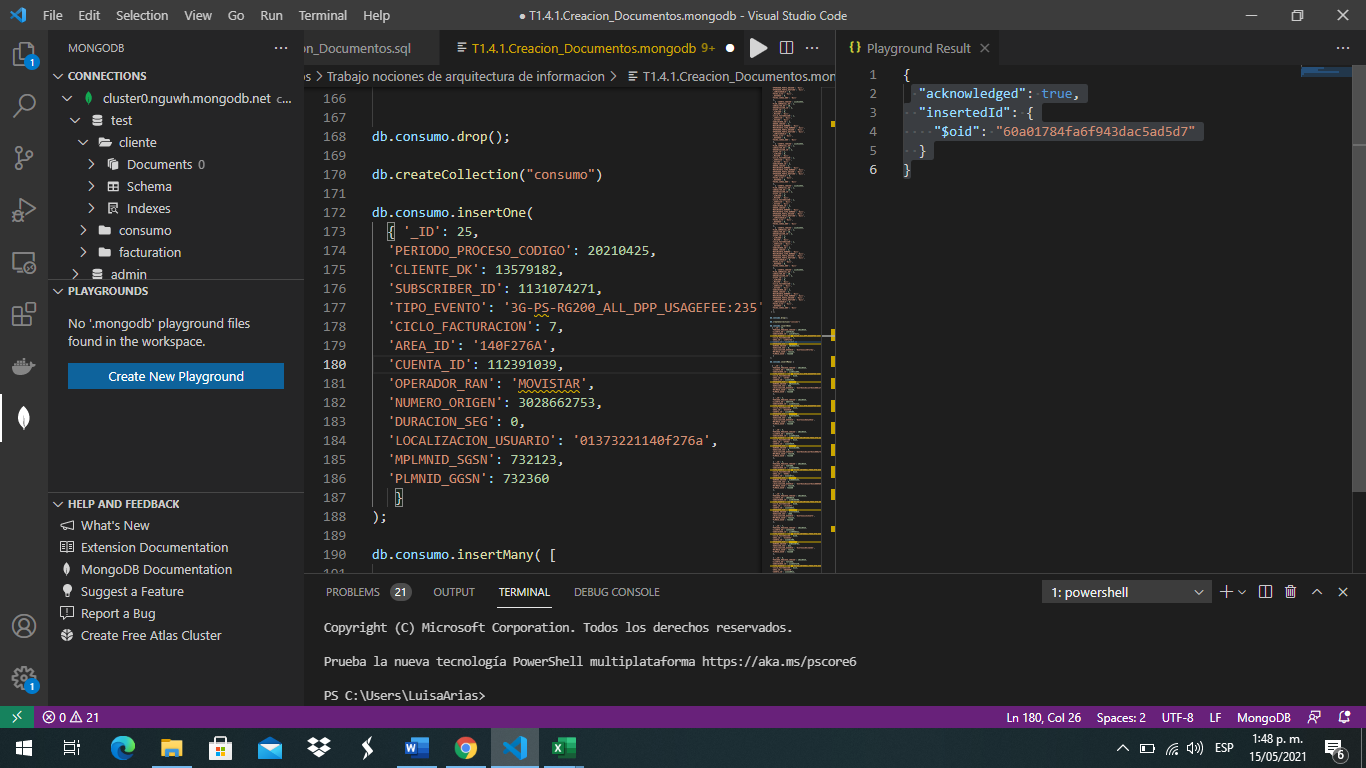
***Observación****: Escriba el código en MongoDB para crear al menos 20 documentos que correspondan a su conjunto de datos específico. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.4.1.Creacion\_Documentos.sql*

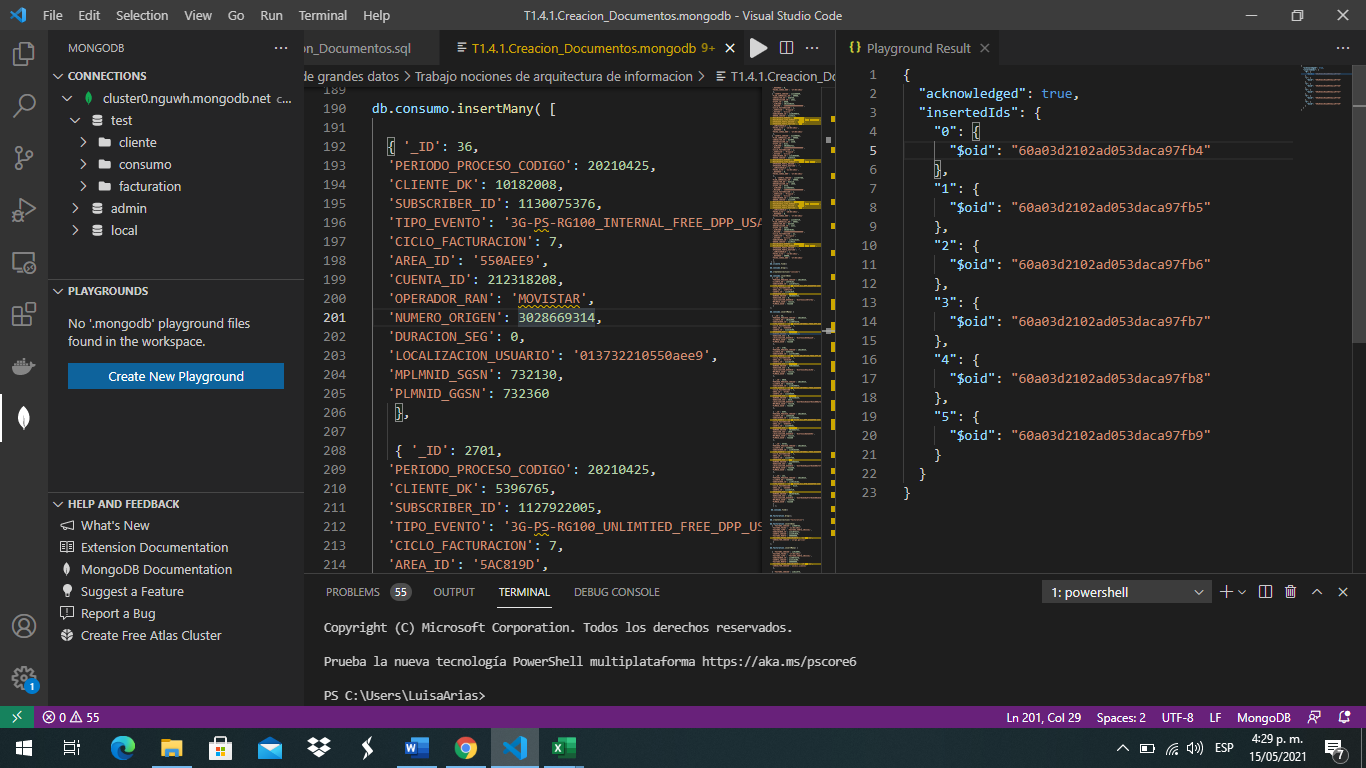
**COLECCIÓN CLIENTE**: creación de 7 documentos por colección cliente.



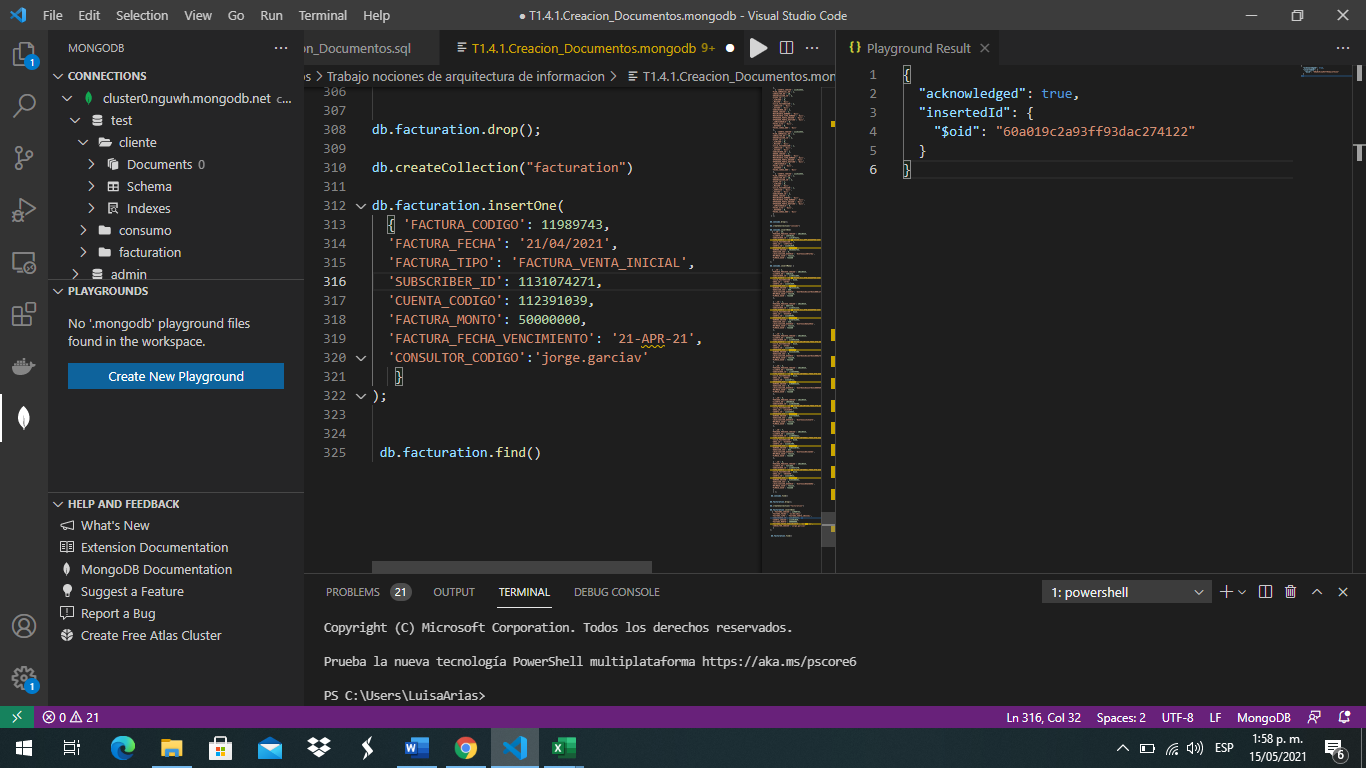


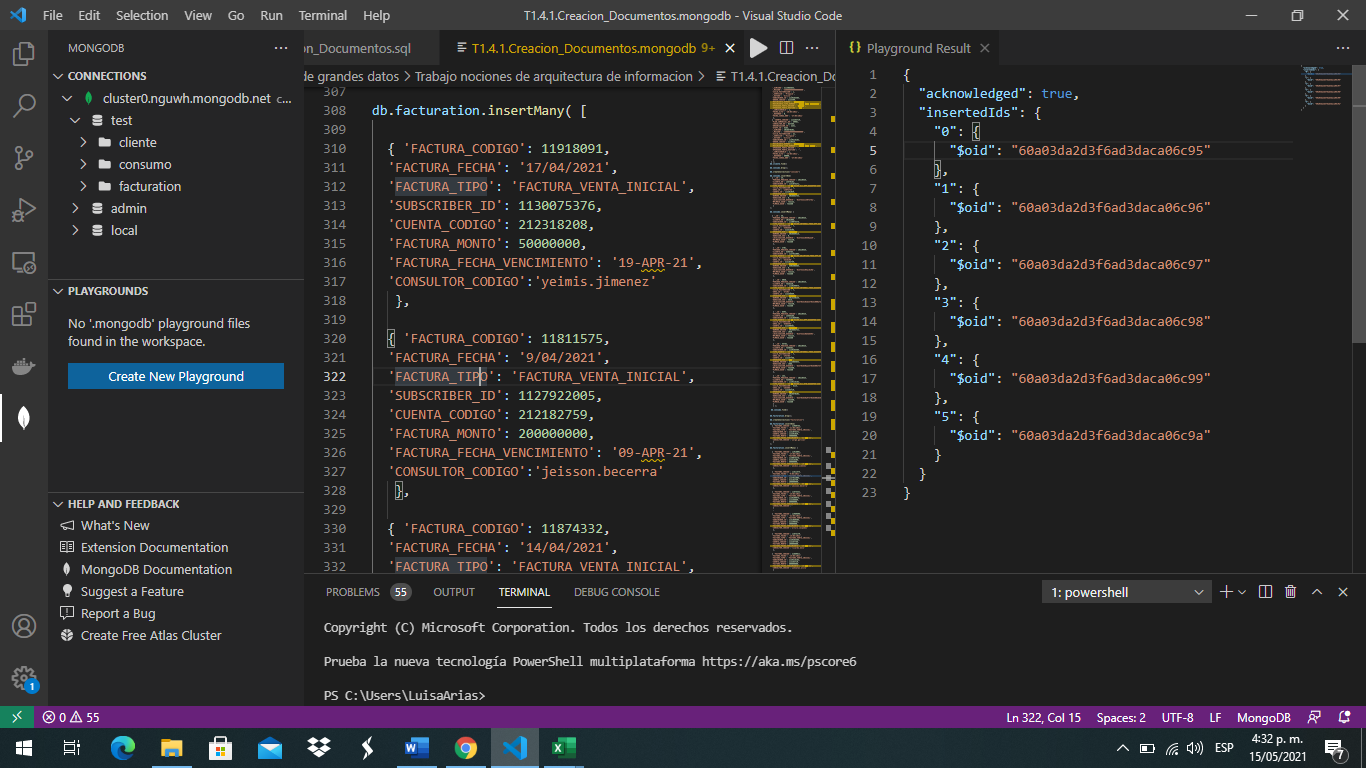
**COLECCIÓN CONSUMO**: creación de 7 documentos por colección consumo.





**COLECCIÓN FACTURACIÓN**: creación de 7 documentos por colección facturación.





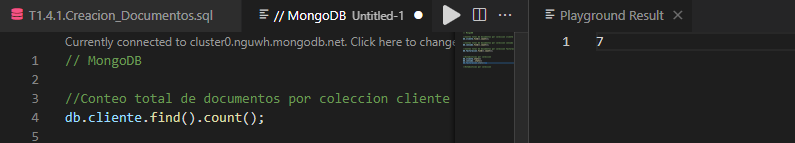
**Total de documentos creados:** 21

* 1. **Sentencia de consulta**

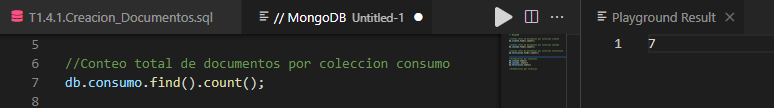
***Observación****: Realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de 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*

**EXPLORACION DE DATOS**

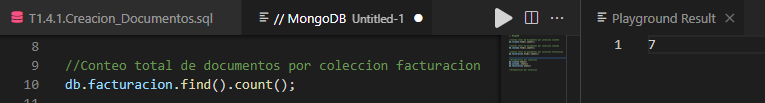
**COLECCION CLIENTE**: Estadística descriptiva, conteo total de documentos



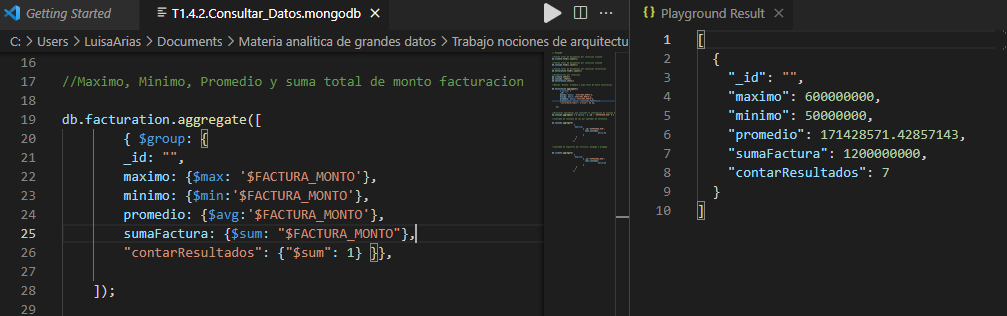
**COLECCION CONSUMO**: Estadística descriptiva, conteo total de documentos



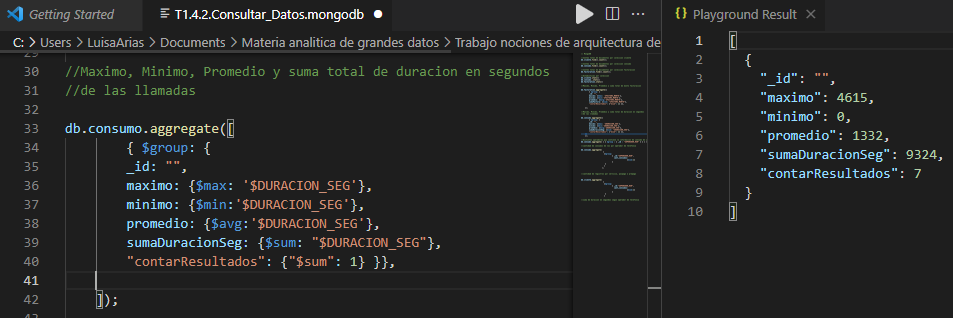
**COLECCION FACTURACIÓN**: Estadística descriptiva, conteo total de documentos



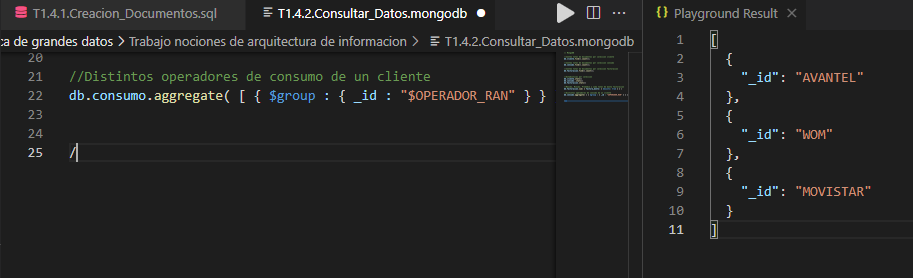
* Maximo, Mínimo, Promedio, Conteo y Suma total de montos de facturación



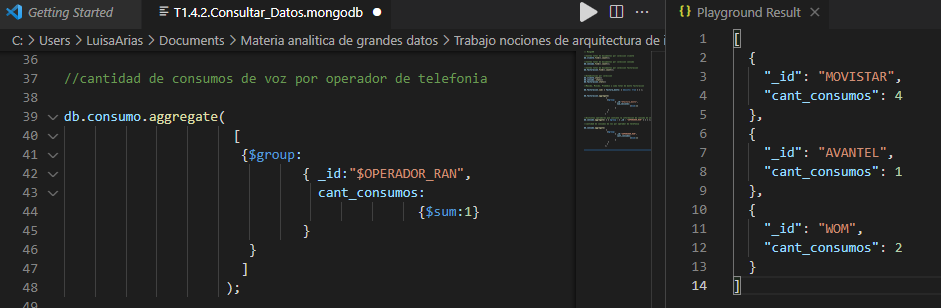
* Maximo, Mínimo, Promedio, Conteo y Suma total de duración de las llamadas en segundos



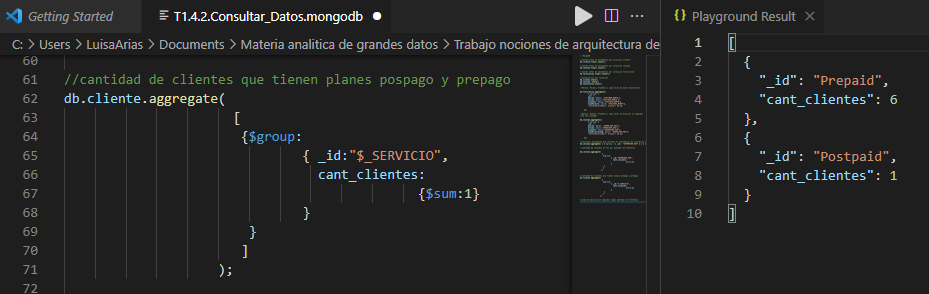
* Distintos Operadores de telefonía que registran en la colección de consumos de voz de los clientes



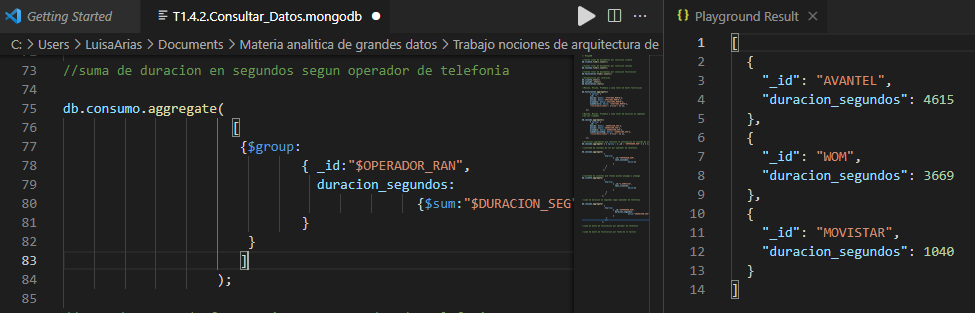
* Cantidad de Consumos de voz que se registraron por cada operador de telefonía



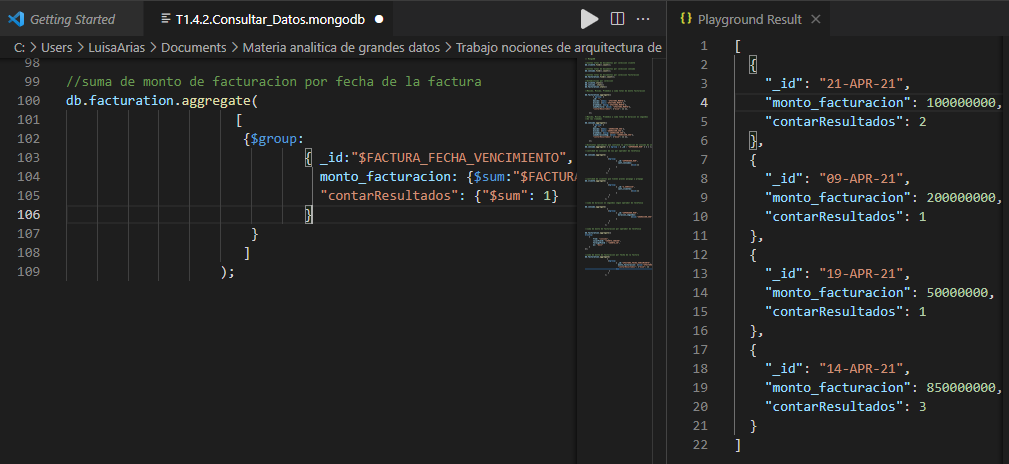
* Cantidad de clientes por planes pospago y prepago



* Suma total de duración en segundos según operador de telefonía



* Suma de monto de facturación y conteo por fecha de vencimiento de la factura



* **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?

La primera limitación es el reto de transformar la idea de datos en filas y columnas a datos en forma de objetos relacionados entre sí. En el mundo real, filas y columnas no es una manera que exista de almacenar información, sin embargo, objetos y relaciones entre ellos sí. Ese sería el primer problema: las bases de datos relacionales no están pensadas para contener la información de manera tan real o precisa. Otra limitación a mencionar, ésta más técnica, es que las bases de datos relacionales no se adaptan bien a los cambios frecuentes en los datos. Funcionan bien con problemas con información bien definida al principio del todo y que no vaya variando demasiado con el tiempo. Para algunos tipos de problemas, esta particularidad de las bases relacionales puede suponer una importante limitación.

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

Con las bases de datos no relacionales se pierde un poco el esquema ACID (atomic, consistent, isolated, durable). Esas garantías se pierden un poco en favor a mayor rendimiento de la base de datos. Por ejemplo, esto se ve en las bases de datos no relacionales en el hecho de que no tienen la posibilidad de hacer operaciones de manera atómica, por tema de performance, cosa que las bases de datos relacionales sí pueden. Un equipo de desarrollo que esté haciendo una migración a NoSQL deberá considerar esto para cambiar las consultas que hagan a la base de datos adaptándolas al sistema no relacional. Otra limitación de las bases de datos no relacionales es en relación a los huecos de información que pueda haber. En las bases relacionales, la clave foránea garantiza que la base de datos siempre esté en un formato válido y consistente mientras que las no relacionales no suelen tener tal garantía. El orden y validez de los datos corre mucho más de cuenta del desarrollador, que deberá prestar más atención a validar sus datos antes de escribirlos en su base de datos.

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

Cuando tenemos un conjunto de datos muy volátil y creciente, es posible que las bases de datos relacionales empiecen a ser insuficientes pues es complicado escalar este tipo de bases de datos. Por otro lado, si queremos minimizar los recursos computacionales de nuestras máquinas, lo mejor será ir a un modelo de bases de datos no relacionales, pues son menos demandantes a este respecto. Si tenemos un conjunto de datos que presentan mucha volatilidad de los mismos o que su forma de representación es algo que pueda ir cambiando con el tiempo, entonces merece la pena usar NoSQL. De todos modos, y en general, dependerá mucho esta decisión del stack tecnológico que se maneje en el caso concreto. También adaptar el tipo de sistema de almacenamiento a nuestro stack es un punto muy a tener en cuenta y algo importante en el proceso de toma de decisiones.