**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 08:17

*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. Ana María Gaona Gómez |
| 2.Julián Palacio Roldán |
| 3.Natalia Carmona Restrepo |
| 4.Tomas Vergara Cardona |
| 5. |
| **REPO EN GITHUB: https://github.com/jpalacior/grandes-datos-unalmed** |

*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.

Olist es una empresa Brasilera del segmento del comercio electrónico, fundada en 2015 con el objetivo de acercar a comerciantes pequeños a los más grandes marketplaces nacionales e internacionales. Su modo de operación consiste en que los comercios, en calidad de clientes, se registren en Olist y creen un catálogo de sus productos para vender en línea, que luego Olist se encarga de publicar en marketplaces como Amazon, Mercado Libre, Americanas.com, entre otros. Además de brindarle al cliente exposición a sus productos, Olist ofrece:

1. Plataforma para la gestión financiera, de catálogos y ventas: permite a los comercios registrar sus productos y controlar toda la operación en una sola plataforma, sin recurrir a otros sistemas.
2. Tienda oficial en los marketplaces: los comercios cuentan con una tienda oficial en cada marketplace que les brinda una mejor credibilidad.
3. Servicios logísticos: se brinda un flete 15% más barato respecto del mercado, gracias a asociaciones de Olist con oficinas de correo.
4. Equipos dedicados: se le ofrece a los clientes un servicio con expertos para la revisión de catálogos y un equipo de servicio postventa.

A través de su plataforma web, Olist obtiene los datos de los comercios a través del registro que realizan inicialmente, y los datos de todos los productos que agreguen a su catálogo. Los datos de transacciones, órdenes, clientes y pagos son extraídos de cada uno de los marketplaces aliados por procesos ejecutados por los empleados de Olist e ingresados a la base de datos.

* 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.

Olist requiere diseñar e implementar un modelo de datos NoSQL que le permita almacenar los datos de su operación de forma eficiente para implementar modelos de analítica, teniendo en cuenta el alto volumen de transacciones generado por todos los productos vendidos por más de 90 negocios en todos los marketplaces asociados, y su reciente estrategia de expansión, financiada por softbank con 58 millones de dólares. (aetecno, 2021)

Tener un modelo de datos adecuado le facilitaría a Olist realizar exploración de los datos y modelos de analítica sobre los productos, clientes, opiniones, envíos y alianzas que actualmente maneja; de forma que pueda explotar el potencial de la gran masa de datos para clasificar o predecir, y estar un paso delante de sus competidores a la hora de tomar decisiones estratégicas.

* 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).

* Caracterizar las fuentes de datos en la operación de Olist y elaborar un diagrama del modelo de dominio del negocio, para tener una visión holística del flujo de operación.
* Diseñar un modelo entidad-relación y elaborar un diagrama del mismo, que sirva para mapear con claridad el modelo de datos que se elaborará.
* Elegir un sistema de bases de datos relacional y uno no relacional; y generar los scripts para la creación de tablas, siguiendo lo planteado en el modelo entidad-relación.
* Insertar los datos existentes en las tablas creadas, asegurándose de que los tipos de datos sean leídos adecuadamente.
* Realizar pruebas de exploración y consulta en ambos sistemas de bases de datos, para luego realizar un paralelo y evaluar la conveniencia de los sistemas NoSQL.
  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.

Olist cuenta con un sistema de bases de datos relacional que actualmente soporta la operación del negocio en producción, sin embargo, se plantea generar una base de datos no relacional que facilite los análisis estratégicos a través de soluciones de analítica, teniendo en cuenta que la empresa se encuentra en un periodo agresivo de expansión y que se espera un crecimiento en las ventas.

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 elegidos pertenecen a Olist Store, la cual es una plataforma brasileña de conexión entre pequeños negocios/emprendedores y marketplaces como Amazon y Mercado Libre, los datos consisten en información anonimizada relacionada a ventas realizadas entre el año 2016 y 2018. Olist funciona como intermediario (con cobro de comisión) entre el cliente final y el vendedor, por lo que los datos fueron recolectados directamente de las órdenes de los clientes en el website de Olist Store y en los marketplaces. Los datos se presentan en nueve archivos de formato csv, contienen en total 52 campos, pesan 120 MB y fueron obtenidos a través Kaggle (Olist et al., 2018).

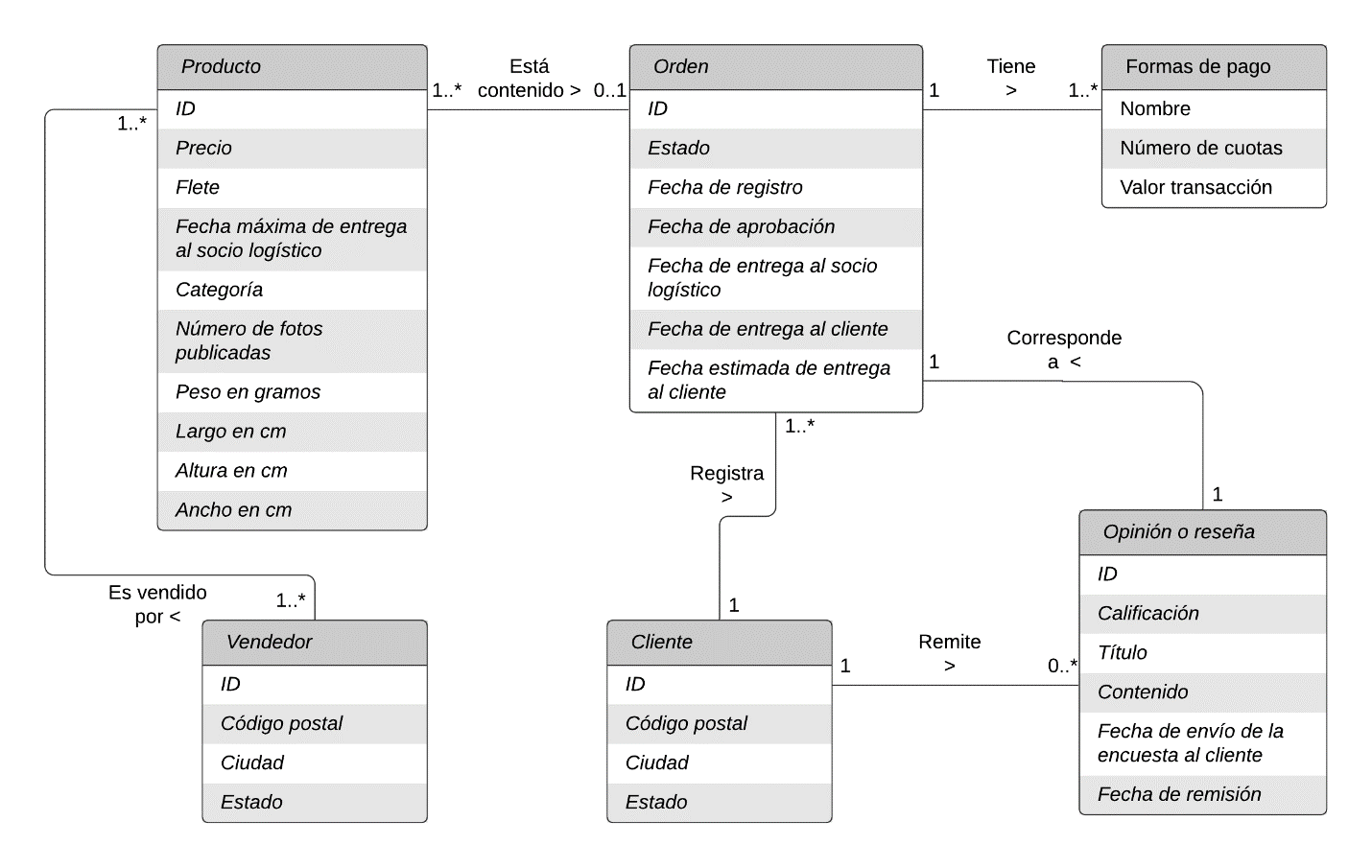
* 1. **Descripción de datos (diccionario):**

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



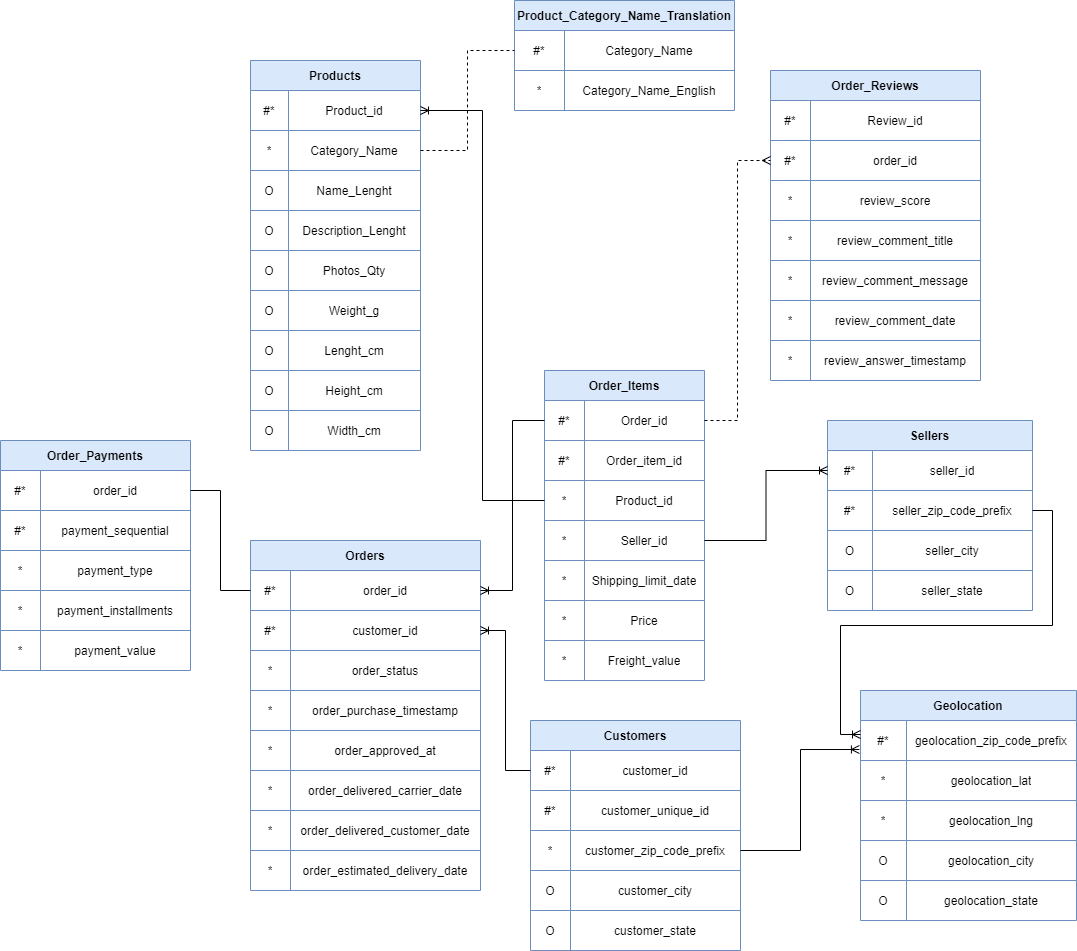
* 1. **Modelo del domino**

***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**

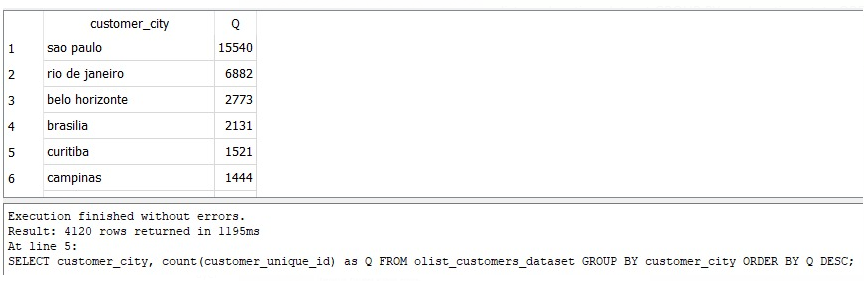
***Observación****: Escriba el código para insertar los datos en cada una de las tablas creadas. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.3.Insertar\_Datos.sql*

* 1. **Sentencia de consulta**

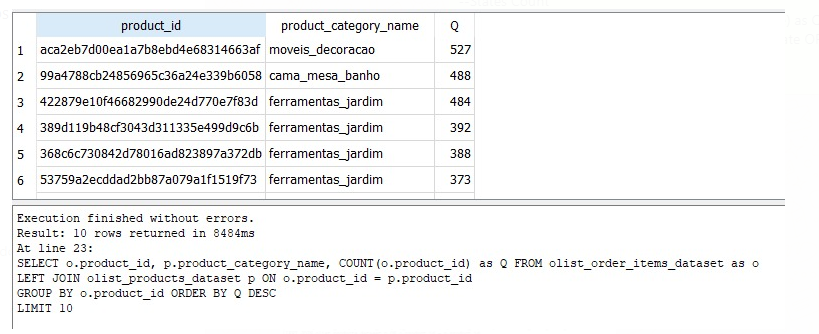
***Observación****: realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de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

IMPORTANTE: A continuación, **algunas** capturas de pantalla de SQLite*,* para ver la totalidad de las consultas remítase al archivo***T1.3.4.Consultar\_Datos.sql***.

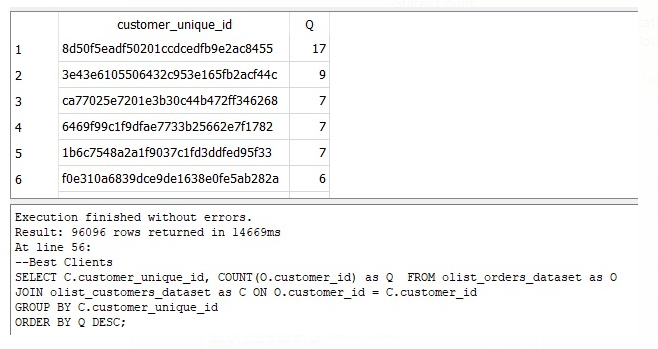
--Customers by City



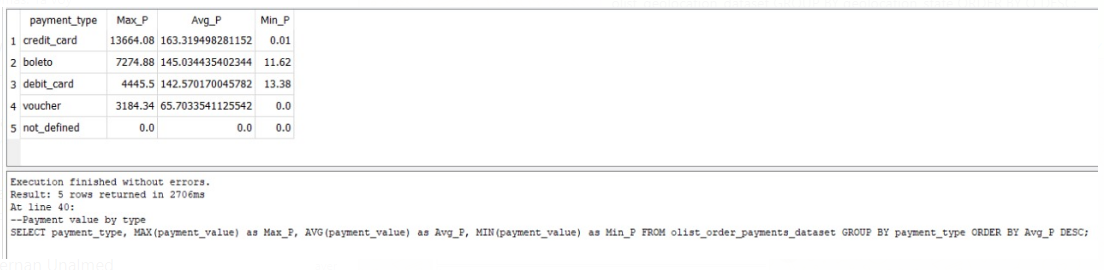
--Most Sold Products (Top 10)



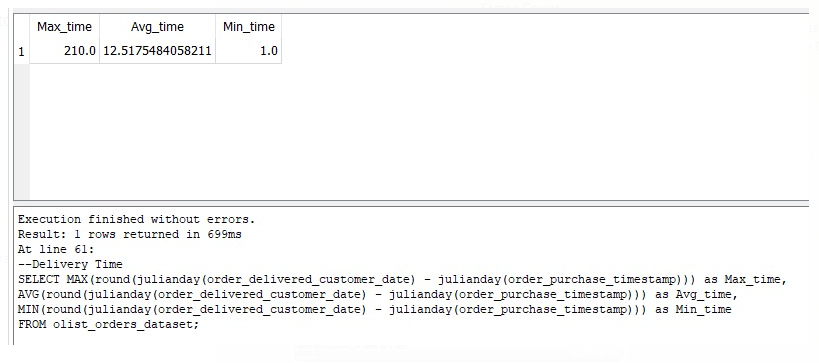
--Best Clients



--Payment value by type



--Delivery time



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

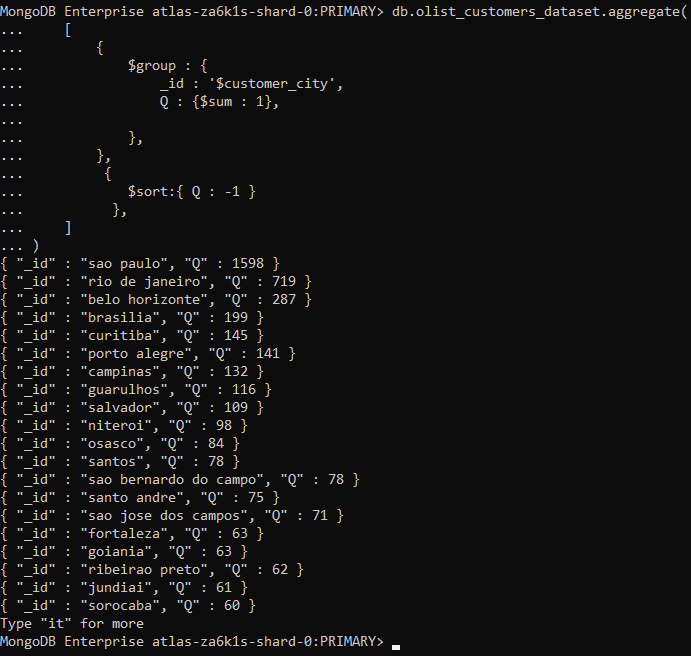
***Observación****: Escriba el código en MongoDB para crear al menos 20 documentos que correspondan a su conjunto de datos específico. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.4.1.Creacion\_Documentos.sql*

* 1. **Sentencia de consulta**

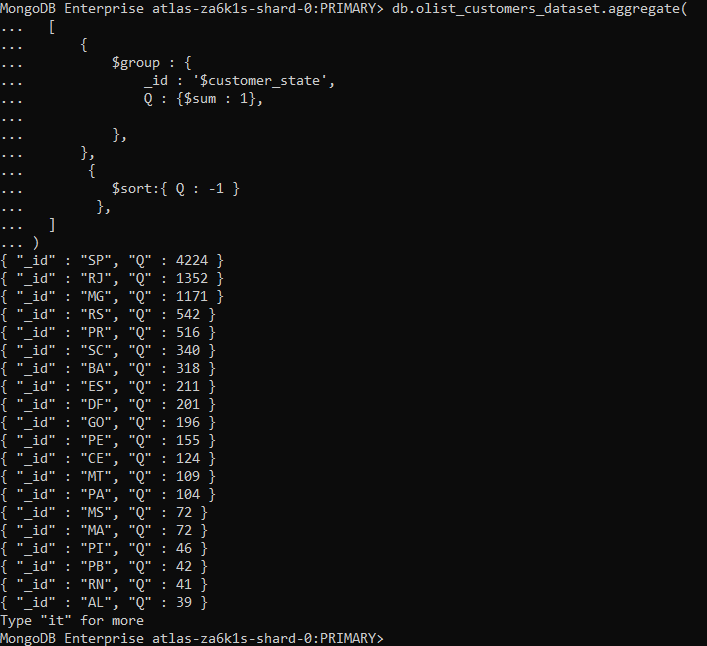
***Observación****: Realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de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

IMPORTANTE: A continuación, **algunas** capturas de pantalla de mongodb*,* para ver la totalidad de las consultas remítase al archivo***T1.4.2.Consultar\_Datos.sql***.

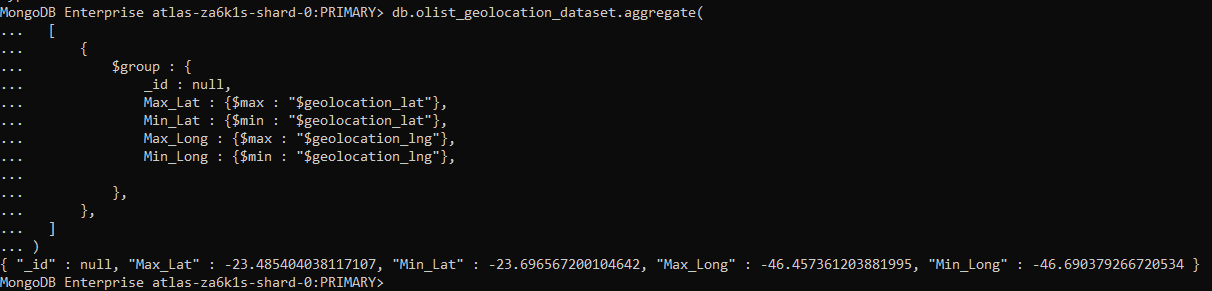
// Customers by City



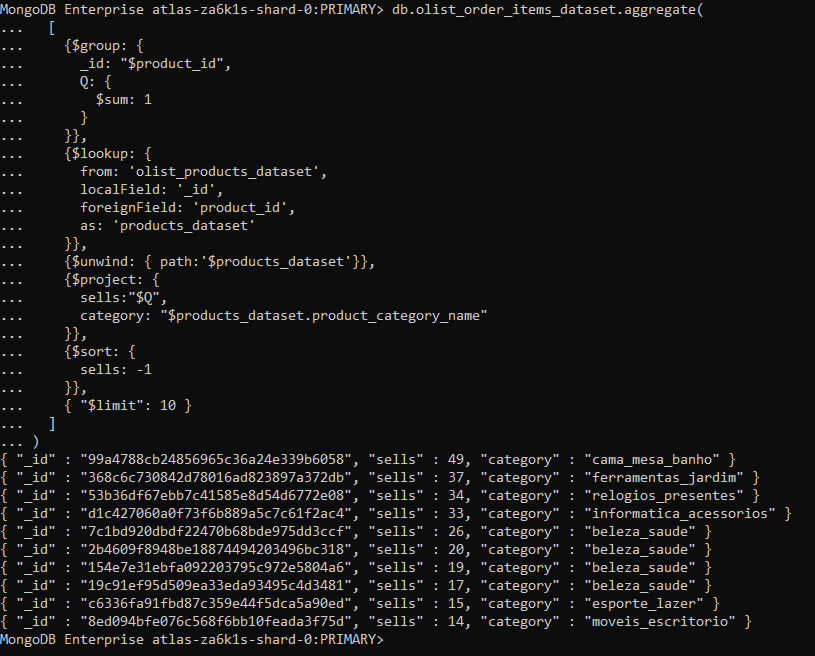
// Customers by State



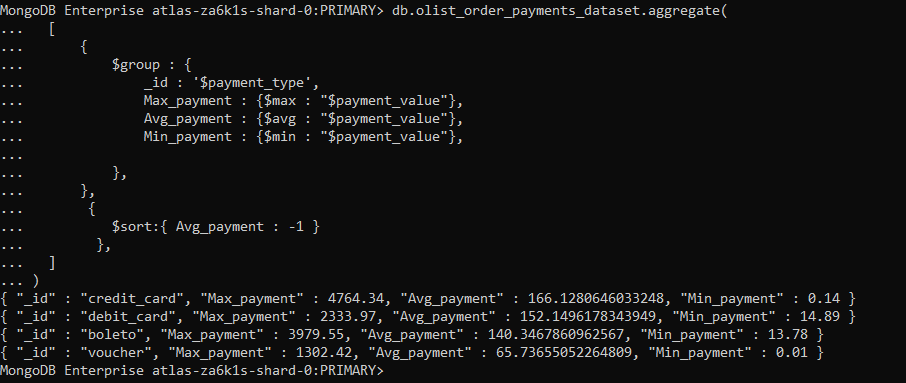
// Operation Area



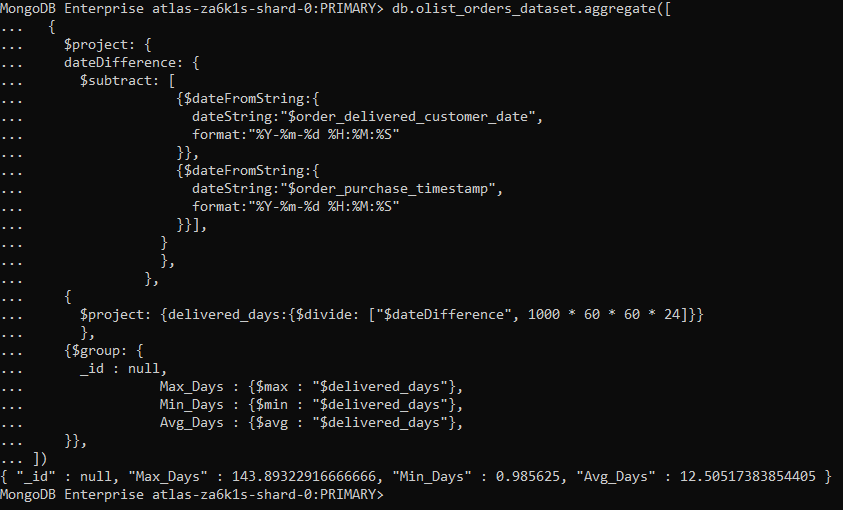
// Most Sold Products (Top 10)



// Payment value by type



// Delivery Time



/\*Category Name Translation\*/



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?**

Los autores contextualizan e identifican las ventajas del uso de las bases de datos gráficos sobre los conjuntos de datos relacionales. A pesar de lo anterior, destacan como una limitación el proceso de migración pues este proceso requiere una adaptación desde el modelo de datos relacional al de NoSQL o de gráficos. Específicamente, los autores destacan **el requerimiento de tiempo** y **tecnología** como una limitación en este proceso. Argumentan que “las migraciones proporcionan un enfoque estructurado y por pasos para la refactorización de la base de datos para que pueda evolucionar para cumplir con los requisitos cambiantes. A diferencia de las refactorizaciones de código, que suelen tardar unos minutos o segundos, las refactorizaciones de la base de datos pueden tardar semanas o meses en completarse, con tiempo de inactividad para los cambios de esquema.”

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

En la práctica la migración de datos relacionados a NoSQL conlleva muchos cambios a nivel de desarrollo, a nivel de despliegue de la aplicación, de coherencia de los datos, conocimiento de bases de datos no relacionales o NoSQL del equipo, por lo tanto, la decisión debe ser tomada después de un análisis general de ventajas y desventajas de la migración. Una vez tomada la decisión, lo más difícil es desnormalizar la data y colocarla en el formato que se quiere la NoSQL; reconstruir los datos deseados y adaptarlos en la NoSQL toma tiempo. Pero en la práctica lo más difícil es entender el nuevo entorno, en la NoSQL de forma que sea eficiente el manejo, la consulta y la interpretación de los datos.

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

Los conjuntos de datos relacionados dificultan la manipulación e interpretación de los datos. Por lo anterior, los autores describen esta migración como una alternativa para el manejo del volumen, la velocidad y la variedad de los datos y afirman que, “el mayor valor que aportan los gráficos a la pila de desarrollo es su capacidad para almacenar relaciones y conexiones como entidades de primera clase”. Algunas otras ventajas que enumera son; el rendimiento de minutos a milisegundos debido a que una base de datos de gráficos convierte JOINs complejos en recorridos rápidos de gráficos, manteniendo así un rendimiento de milisegundos independientemente del tamaño general del conjunto de datos, ciclos de desarrollo altamente acelerados, capacidad de respuesta empresarial extrema. Estas bases de datos admiten un modelo de datos muy flexible y detallado que permiten modelar y administrar dominios enriquecidos de una manera fácil e intuitiva. Estos tipos de datos están diseñados para el entendimiento de la realidad.

**Referencias**:

Olist, Sionek, A., Dabague, Galeazzi, T., Oliveira, M. de, & Horning, W. (2018). Brazilian E-Commerce Public Dataset by Olist. Recuperado el 10 de mayo de 2021, de https://www.kaggle.com/olistbr/brazilian-ecommerce/version/2

aetecno (2021). SOFTBANK INVIERTE EN LA PLATAFORMA MARKETPLACE BRASILEÑA OLIST. Recuperado el 10 de mayo de 2021, de https://tecno.americaeconomia.com/articulos/softbank-invierte-en-la-plataforma-marketplace-brasilena-olist