**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: 2021.05.24

*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. Maria del Pilar Hernández Moreno - 1152458223 |
| 2. Rubén Esneyder Morales Barbosa - 1072429901 |
| 3. Andrés Felipe Hernández Velandia - 1143390458 |
| 4. Gabriel Jaime Sánchez Montoya - 1017229936 |
| **REPO EN GITHUB: https://github.com/pilarhernandez96/AGD-2021.git** |

*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.

**R/:**

Los datos son generados a partir del estudio del mercado laboral en tiempo de pandemia COVID-19, realizado por la compañía Greenwich.HR en EE. UU. El conjunto de datos se deriva del volumen de ofertas de trabajo que se rastrean diariamente a 3,3 millones de organizaciones, agrupados por estado geográfico, industria, grupos de cargos y tipos de cargos, comparados con el comportamiento de ofertas de trabajo en tiempo pre-Covid. La tasa se representa como el promedio de ofertas de trabajo durante 7 días hasta una fecha específica en proporción al valor fijo de las ofertas de trabajo ocurridas entre el 24 de febrero y el 1 de marzo del 2020, correspondiente al tiempo pre-Covid.

Greenwich.HR estima que, en promedio, se capturan datos del 70% de todos los nuevos puestos de trabajo en EE. UU. Los datos son recolectados en línea, se calculan diariamente y se publican en la noche. Greenwich.HR es una compañía dedicada entre otras funciones, a la recolección de datos del mercado laboral como lo son clasificaciones de trabajo, niveles de pago, tiempos de contratación, entre otros; manejan limpieza, normalización de datos y otras funciones de higiene para obtener datos limpios y listos para usar.

Para efectos del trabajo y por el volumen de la información original (más de 308 millones de registros a abril/2021), se tomó la decisión de trabajar solo con el estado de Alaska (AK) con registros hasta el 30 de junio 2020 y en 6 industrias.

* 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.

**R/:**

1. Determinar cuáles fueron las cinco industrias más afectadas por los efectos de la pandemia del COVID-19 al 30 de junio de 2020 en Alaska.
2. Identificar: 2.1 cuáles fueron los 20 tipos de cargos con las tasas más bajas y 2.2 cuáles fueron los 10 tipos de cargos con las tasas más altas durante la pandemia al 30 de junio de 2020 en Alaska.
3. Determinar el grupo de cargos (job family) más afectado por la pandemia al 30 de junio de 2020 en Alaska.

Las respuestas a estas preguntas ayudarán a la identificación de los sectores de industrias que requieren menos y más atención haciéndolas visibles ante el Estado para ser beneficiados o no por las políticas y ayudas financieras que se tengan destinadas a la atención del Covid-19.

* 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).

**R/:**

* Establecer el comportamiento de afectación y recuperación del mercado laboral en Alaska mediante el uso de estadística descriptiva, para focalizar y priorizar los apoyos económicos en los estados y grupos familiares que más lo necesitan.
* Identificar cuáles industrias necesitan re activarse más rápidamente para poder disminuir la tasa de desempleo y enfocar en estas el establecimiento de políticas gubernamentales para fortalecer la economía de esos sectores.
  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.

**R/:**

Los impactos laborales de la pandemia se están moviendo de manera rápida, dependiendo del sector industrial, el territorio y las profesiones. El ritmo del desempleo y su recuperación no logra ser interpretado de manera adecuada por las fuentes tradicionales que otorgan datos del mercado laboral en Estados Unidos. Greenwich y OneModel concentran sus esfuerzos en tener los datos más actualizados y precisos posibles combinando la mayor fuente de datos de contratación en EE. UU. con estadísticas gubernamentales semanales. El objetivo es mostrar las decisiones dentro de las empresas, indicando su confianza y preocupaciones sobre la forma y la velocidad de recuperación que se avecina en el futuro en la economía local e internacional pese a la incertidumbre que se tiene por el COVID-19.

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

**R/:**

El conjunto de datos que se utilizarán para la comprensión del problema y la situación actual que ha hoy enfrenta el mercado laboral de los Estados Unidos proporciona actualizaciones diarias sobre el volumen de ofertas de trabajo en Alaska Filtradas por categoría de trabajo, industria, rol desempeñado y geográfica. Estos datos fueron suministrados a través de la empresa Greenwich.HR recolectados diariamente por medio de su propia plataforma de datos mediante la compilación de millones de ofertas de empleo en línea. Para la obtención de este DataSets fue necesario crear nuevas tablas que hicieran conexión con los registros únicos que logramos identificar en cada tabla, debido a que la información que se registraba en dichas tablas no correspondía a un modelo entidad relación y se duplica en cada una de las 10 tablas. Por lo que se recurrió a la herramienta pyspark para consolidar la información en 6 tablas relacionales.

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

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

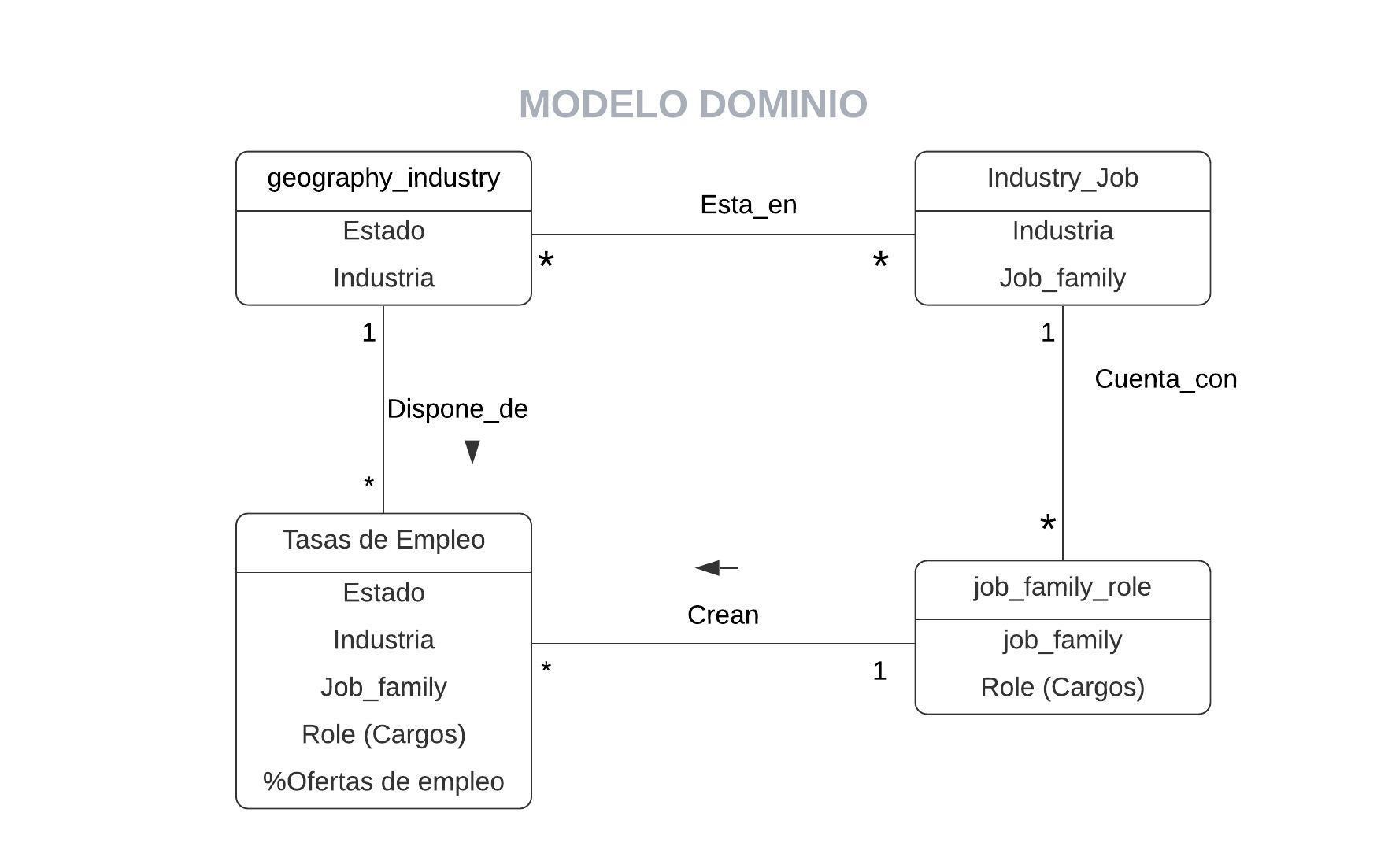
**R/:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tabla** | **Nombre del atributo / variable** | **Formato o Tipo de Dato** | **Descripción** |
| Geography\_industry | Id\_geography\_industry | Int | Llave única de la conexión |
| Id\_State | Int | Llave conexión única para representar cada estado |
| State | Varchar(2) | Abreviatura del estado en EE.UU. |
| Id\_Industry | Int | Llave conexión con Industry |
| Job\_family\_role | ID\_Job\_family\_role | Int | Llave única de Job\_family\_role |
| Id\_job\_family | Int | Llave conexión única para representar cada Job\_family |
| Job\_family | Varchar(60) | Nombre de familia de empleo |
| Id\_role | Int | Llave única de Role |
| Tasas\_Empleo | ID | Int | Llave única de Tasas\_Empleo |
| post\_date | DATE | Fecha análisis |
| id\_State | Int | Llave conexión con Geography\_industry |
| id\_Indt | Int | Llave conexión con Industry |
| id\_job\_family | Int | Llave conexión con Job\_family\_role |
| id\_role | Int | Llave conexión con Role |
| count\_id\_indexed | Float | Recuento del número total de anuncios únicos de empleos en EE.UU. Por estado, industria, familia de empleos y cada role(cargo) vinculado durante un período de 7 días que finaliza en post\_date, dividido por el recuento del número total de ofertas únicas para empleos en EE.UU. dentro de la familia de empleos para el período 24 febrero 2020-1 marzo 2020. |
| Industry\_job\_family | ID\_Industry\_Job | Int | Llave única de la conexión |
| id\_Indt | Int | Llave conexión con Industry |
| Id\_job\_family | Int | Llave conexión con Job\_family\_role |
| Industry | Id\_Indt | Int | Llave única de Industry |
| Industry | Varchar(20) | Nombre de la industria |
| Role | Id\_role | Int | Llave única de role |
| Role | Varchar(60) | Nombre del rol o cargo (nota: se excluyen los roles de conteo bajo). |

* 1. **Modelo del dominio**

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

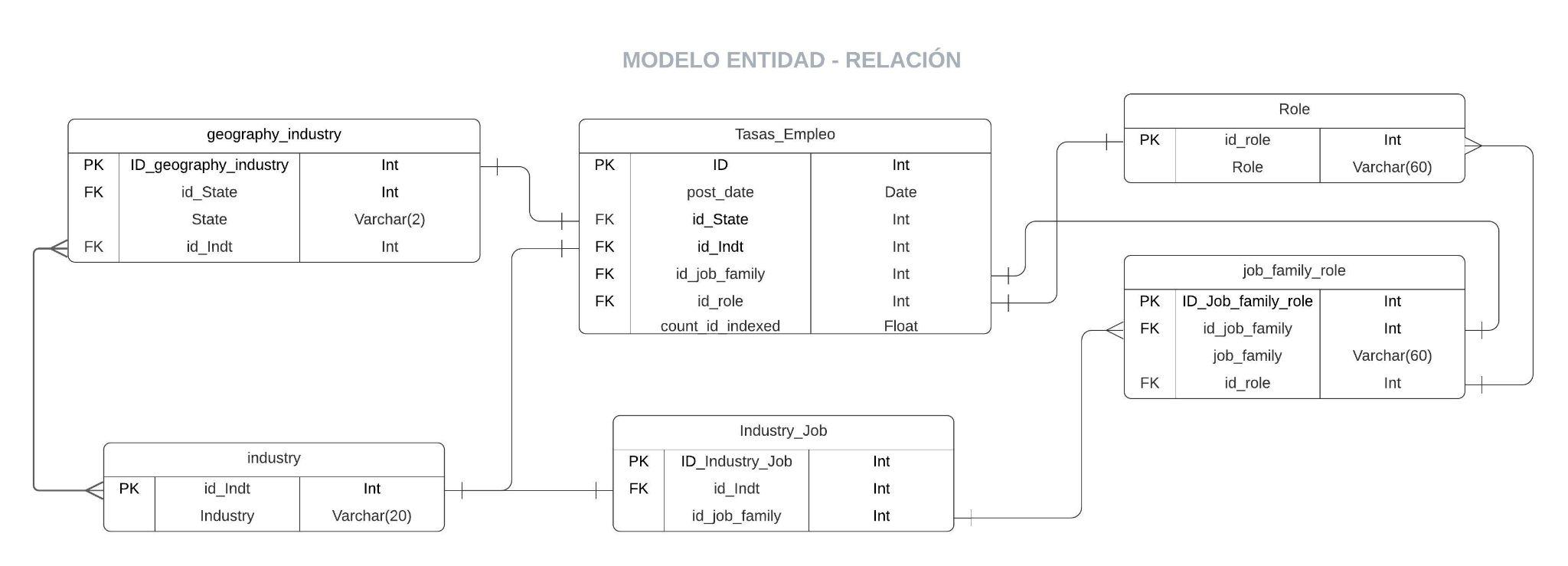
**R/:**



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.*

**R/:**



* 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)

**R/:**

DROP TABLE IF EXISTS "Role";

CREATE TABLE [Role] (

[id\_role] Int primary key,

[Role] Varchar(100),

);

DROP TABLE IF EXISTS job\_family\_role;

CREATE TABLE [job\_family\_role] (

[ID\_Job\_family\_role] Int primary key,

[id\_job\_family] Int,

[job\_family] Varchar(100),

[id\_role] Int,

);

DROP TABLE IF EXISTS Industry\_Job;

CREATE TABLE [Industry\_Job] (

[ID\_Industry\_Job] Int primary key,

[id\_Indt] Int,

[id\_job\_family] Int,

);

DROP TABLE IF EXISTS industry;

CREATE TABLE [industry] (

[id\_Indt] Int primary key,

[Industry] Varchar(50),

);

DROP TABLE IF EXISTS geography\_industry;

CREATE TABLE [geography\_industry] (

[ID\_geography\_industry] Int primary key,

[id\_State] Int,

[State] Varchar(2),

[id\_Indt] Int

);

DROP TABLE IF EXISTS Tasas\_Empleo;

CREATE TABLE [Tasas\_Empleo] (

[ID] Int primary key,

[post\_date] Date,

[id\_State] Int,

[id\_Indt] Int,

[id\_job\_family] Int,

[id\_role] Int,

[count\_id\_indexed] float,

);

* 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*

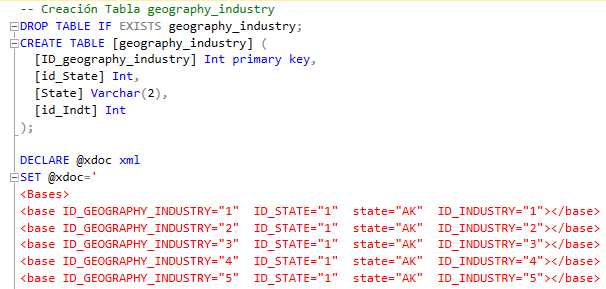
**R/:**

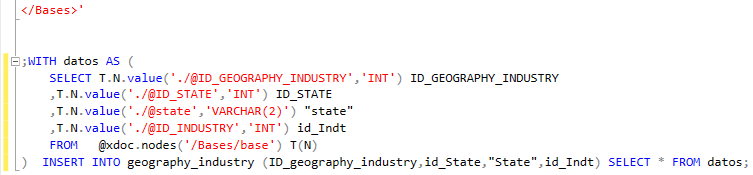
Para realizar la inserción de los datos se usó en primera instancia la creación de una tabla virtual en cada caso y luego se realizó la inserción utilizando un Select From. A Continuación, se detalla el código usado en cada caso.

Nota: por la cantidad de registros que se deben insertar en algunas tablas, se ilustran solo las cinco primeras líneas en los casos que aplique. Para ver el código en su totalidad, por favor remitirse a la carpeta donde reposan los script de cada tabla.

* **Tabla “geography\_industry”:**

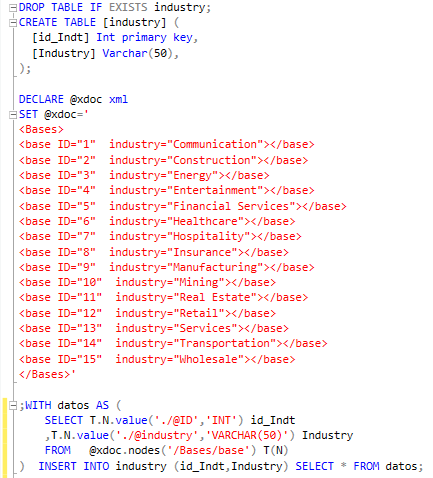
Código para insertar datos (Sólo 5 líneas. En total son 1.080 registros):

****

****

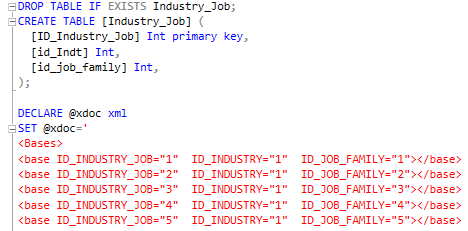
* **Tabla “industry”:**

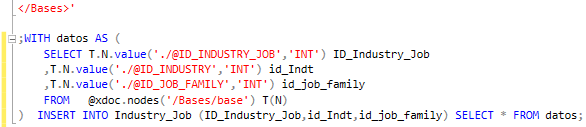
Código para insertar datos:

****

* **Tabla “Industry\_Job”:**

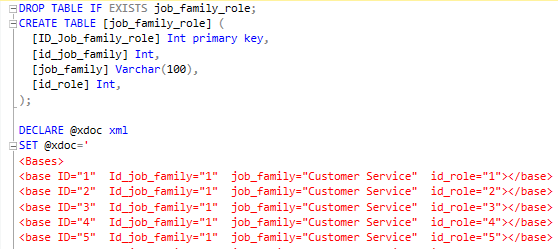
Código para insertar datos (Sólo 5 líneas. En total son 270 registros):

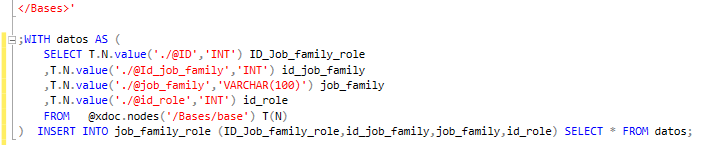
****

****

* **Tabla “job\_family\_role”:**

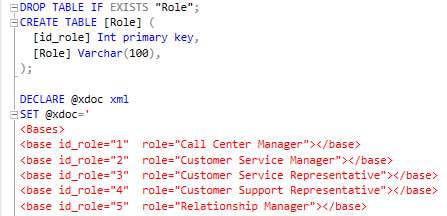
Código para insertar datos (Sólo 5 líneas. En total son 1.700 registros):

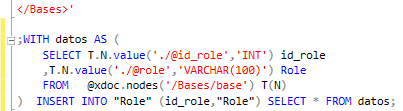
****

****

* **Tabla “Role”:**

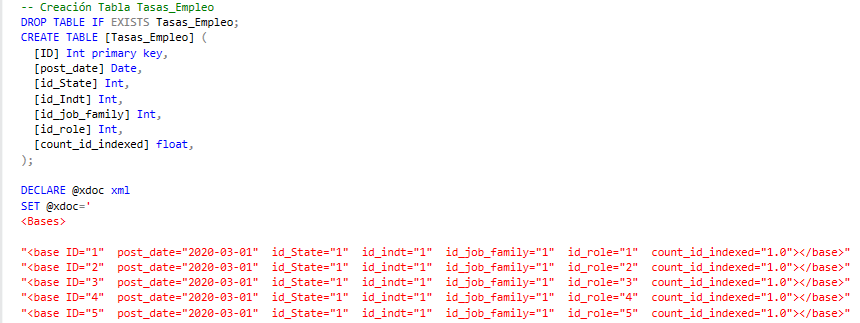
Código para insertar datos (Sólo 5 líneas. En total son 1.700 registros):

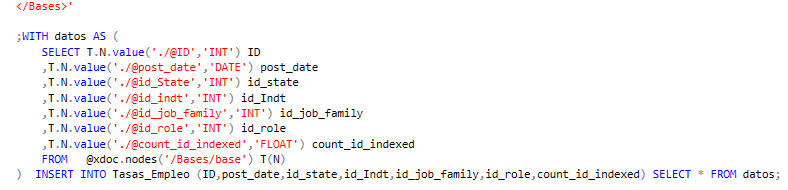
****

****

* **Tabla “Tasas\_Empleo”:**

Código para insertar datos (Sólo 5 líneas. En total son 821.357 registros):

****

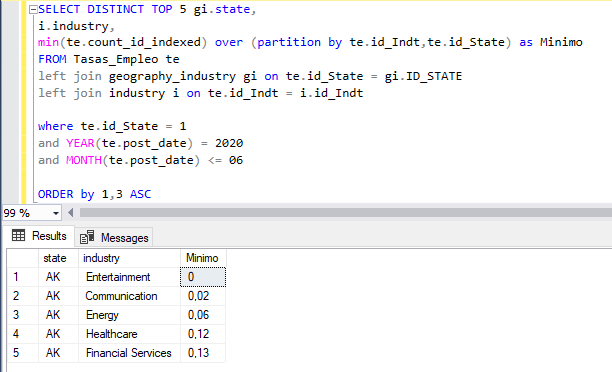
****

* 1. **Sentencia de consulta**

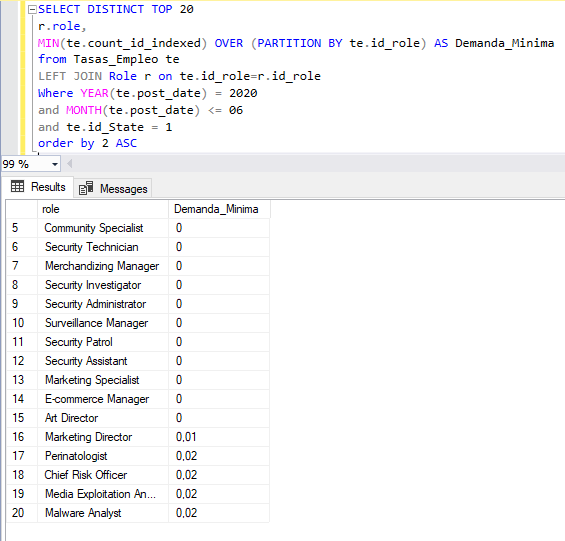
***Observación****: realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, apcce estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de pantalla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

**R/:**

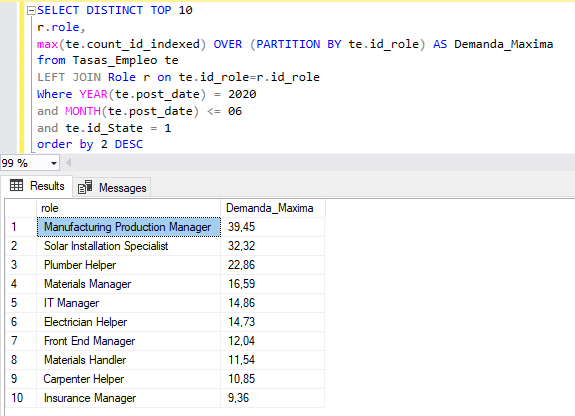
* Sentencia para problematica 1



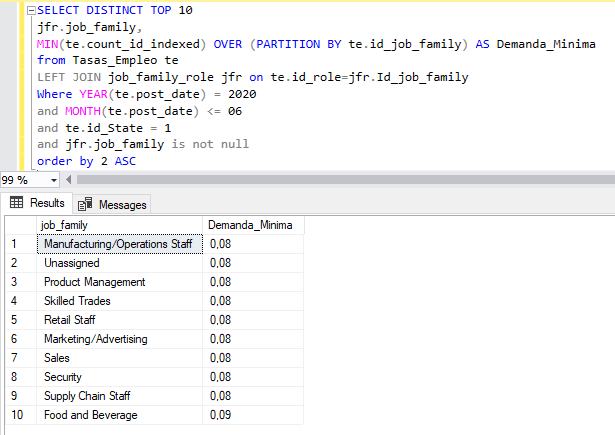
* Sentencia para problematica 2.1



* Sentencia para problematica 2.2



* Sentencia para problematica 3



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*

**R/:**

* Se crea la colección (tabla) con el nombre “tasa\_empleo”

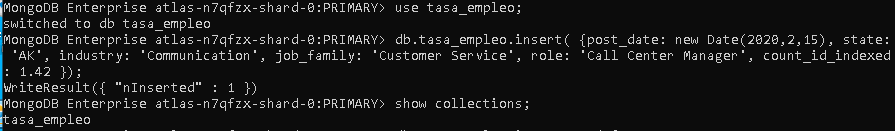
*use tasa\_empleo;*

* Se ingresa un documento (registro) para crear colección tasa\_empleo

*db.tasa\_empleo.insert( {post\_date: new Date(2020,2,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Call Center Manager', count\_id\_indexed: 1.42 });*

* Luego, se verifica la creación de la colección tasa\_empleo

*show collections*



* Se insertan los demás documentos correspondientes al mes de marzo, recordando que JavaScript maneja los meses del año con la numerología de de 0 a 11.

*db.tasa\_empleo.insertMany( [*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Customer Service Manager', count\_id\_indexed: 1.41 },*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Customer Service Representative', count\_id\_indexed: 1.38 },*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Customer Support Representative', count\_id\_indexed: 1.46 },*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Relationship Manager', count\_id\_indexed: 1.39 },*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Service Desk Manager', count\_id\_indexed: 1.44 },*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Service Manager', count\_id\_indexed: 1.4 },*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Call Center Manager', count\_id\_indexed: 1.03 },*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Customer Service Manager', count\_id\_indexed: 1.02 },*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Customer Service Representative', count\_id\_indexed: 0.99 },*

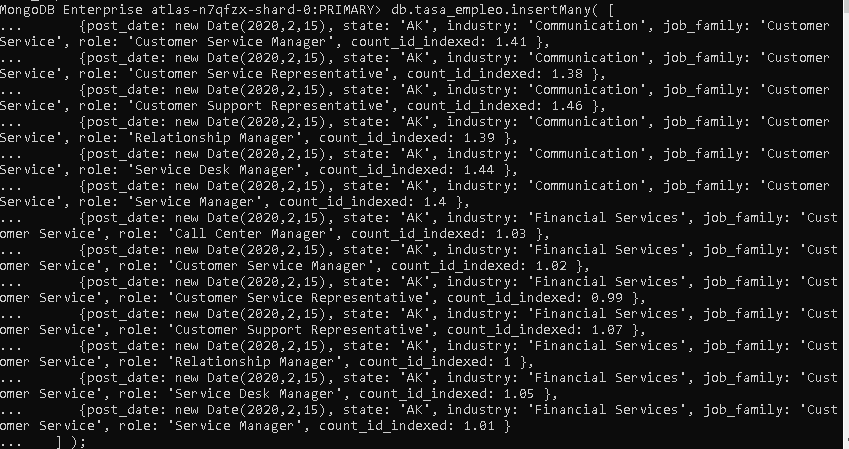
*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Customer Support Representative', count\_id\_indexed: 1.07 },*

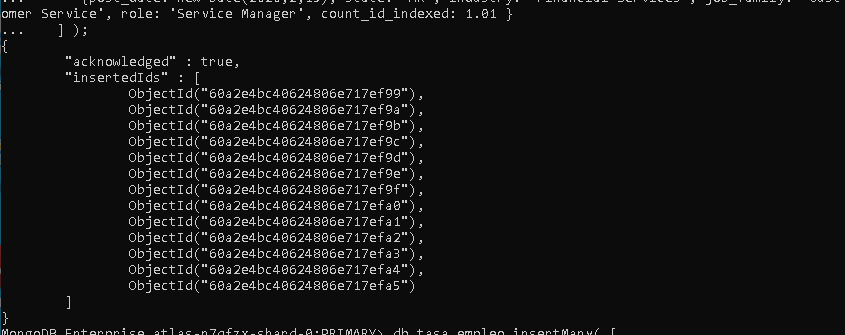
*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Relationship Manager', count\_id\_indexed: 1 },*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Service Desk Manager', count\_id\_indexed: 1.05 },*

*{post\_date: new Date(2020,2,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Service Manager', count\_id\_indexed: 1.01 }*

*] );*





* Se insertan los demás documentos correspondientes al mes de abril

*db.tasa\_empleo.insertMany( [*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Call Center Manager', count\_id\_indexed: 0.45 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Customer Service Manager', count\_id\_indexed: 0.41 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Customer Service Representative', count\_id\_indexed: 0.39 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Customer Support Representative', count\_id\_indexed: 0.42 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Relationship Manager', count\_id\_indexed: 0.34 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Senior Customer Service Representative', count\_id\_indexed: 0.42 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Service Desk Manager', count\_id\_indexed: 0.38 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Communication', job\_family: 'Customer Service', role: 'Service Manager', count\_id\_indexed: 0.38 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Call Center Manager', count\_id\_indexed: 0.3 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Customer Service Manager', count\_id\_indexed: 0.26 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Customer Service Representative', count\_id\_indexed: 0.24 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Customer Support Representative', count\_id\_indexed: 0.27 },*

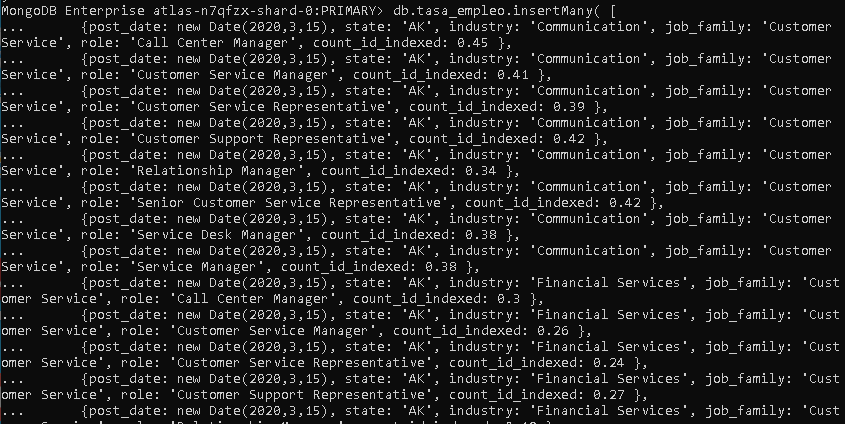
*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Relationship Manager', count\_id\_indexed: 0.19 },*

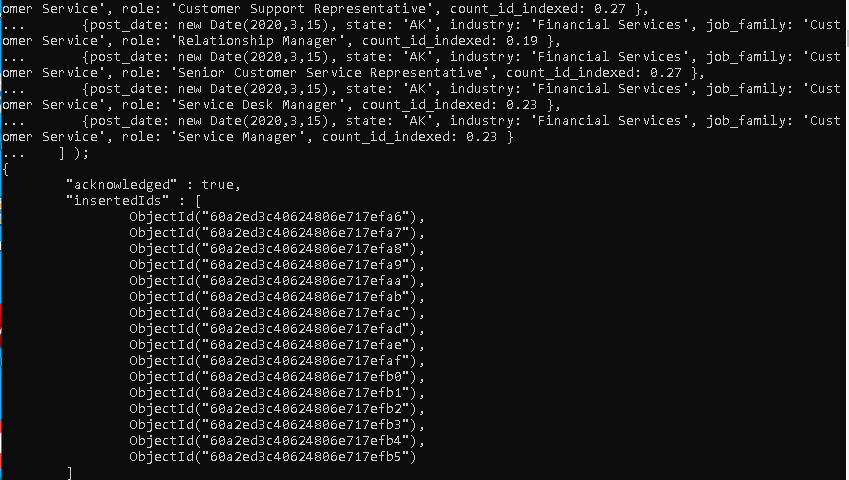
*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Senior Customer Service Representative', count\_id\_indexed: 0.27 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Service Desk Manager', count\_id\_indexed: 0.23 },*

*{post\_date: new Date(2020,3,15), state: 'AK', industry: 'Financial Services', job\_family: 'Customer Service', role: 'Service Manager', count\_id\_indexed: 0.23 }*

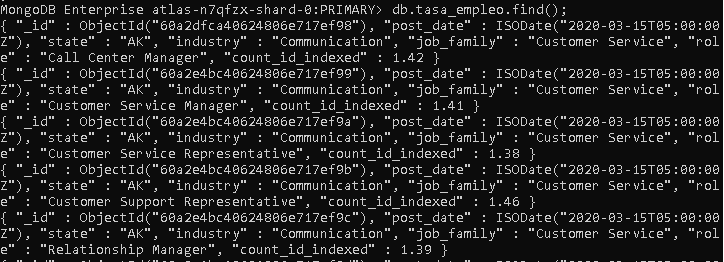
*] );*





* Se verifica la creación de los documentos en la colección tasa\_empleo

*db.tasa\_empleo.find();*



* 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*

**R/:**

**CONTEOS**

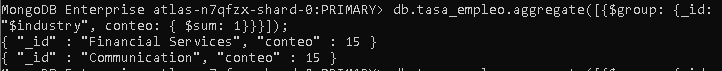
* ¿Cuántos registros tiene la base de datos “tasa\_empleo”? (conteo total)

*db.tasa\_empleo.count();*



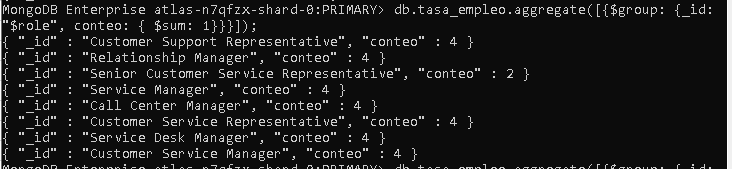
* ¿Cuántos registros existen por tipo de industria? (conteo por categoría industry)

*db.tasa\_empleo.aggregate([{$group: {\_id: "$industry", conteo: { $sum: 1}}}]);*



* ¿Cuántos registros existen por tipo de role? (conteo por categoría role)

*db.tasa\_empleo.aggregate([{$group: {\_id: "$role", conteo: { $sum: 1}}}]);*



* ¿Cuántos registros existen por tipo de job\_family? (conteo por categoría job\_family)

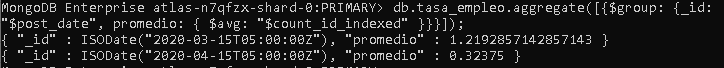
*db.tasa\_empleo.aggregate([{$group: {\_id: "$job\_family", conteo: { $sum: 1}}}]);*



**PROMEDIOS**

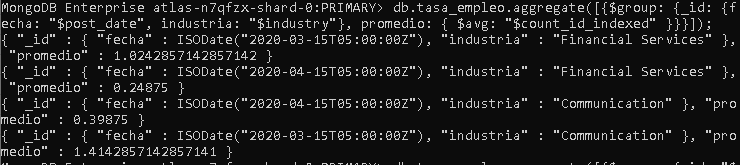
* ¿Cuál es el promedio de tasa de nuevos anuncios por mes? (promedio por mes)

*db.tasa\_empleo.aggregate([{$group: {\_id: "$post\_date", promedio: { $avg: "$count\_id\_indexed" }}}]);*



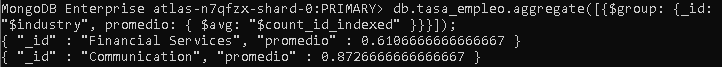
* ¿Cuál es el promedio de tasa de nuevos anuncios por mes por cada industria? (promedio por mes e industria)

*db.tasa\_empleo.aggregate([{$group: {\_id: {fecha: "$post\_date", industria: "$industry"}, promedio: { $avg: "$count\_id\_indexed" }}}]);*



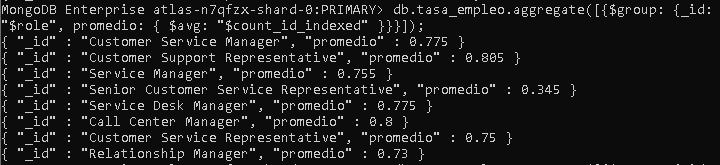
* ¿Cuál es el promedio de tasa de nuevos anuncios por industria? (promedio por industria)

*db.tasa\_empleo.aggregate([{$group: {\_id: "$industry", promedio: { $avg: "$count\_id\_indexed" }}}]);*



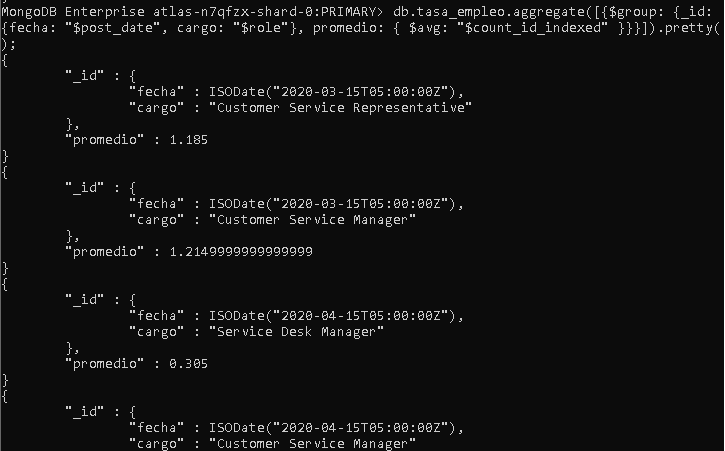
* ¿Cuál es el promedio de tasa de nuevos anuncios por role? (promedio por role)

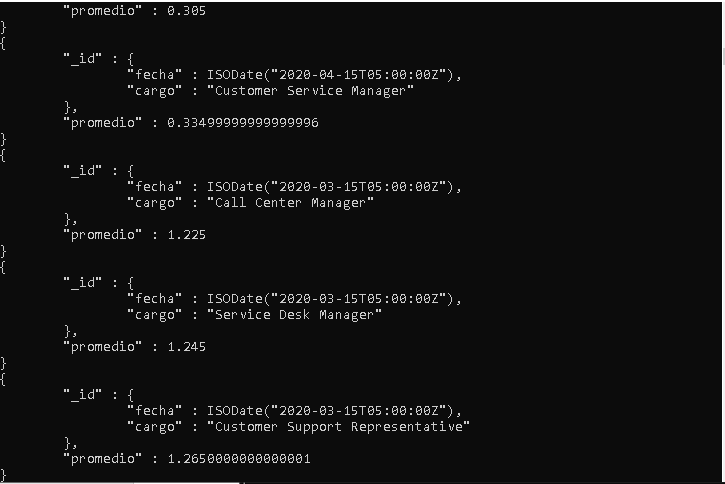
*db.tasa\_empleo.aggregate([{$group: {\_id: "$role", promedio: { $avg: "$count\_id\_indexed" }}}]);*

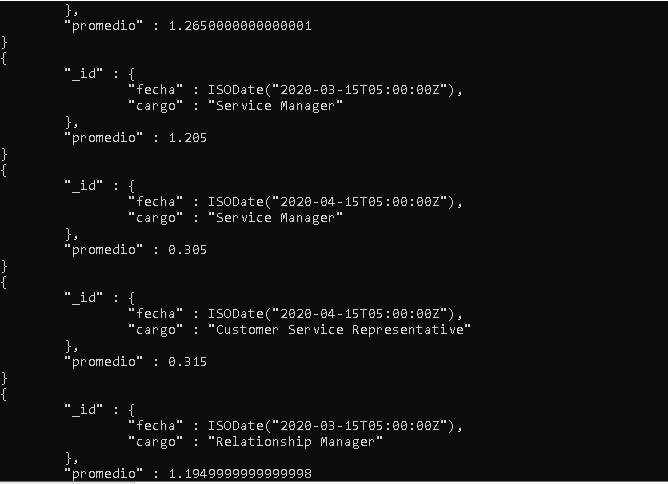


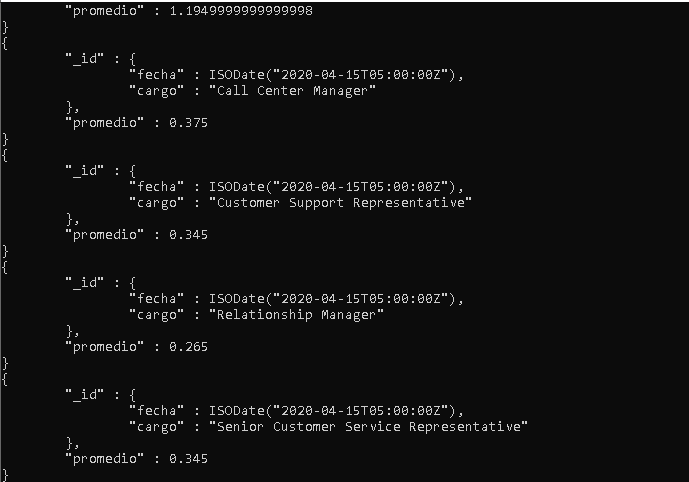
* ¿Cuál es el promedio de tasa de nuevos anuncios por mes por cada cargo? (promedio por mes e industry)

*db.tasa\_empleo.aggregate([{$group: {\_id: {fecha: "$post\_date", cargo: "$role"}, promedio: { $avg: "$count\_id\_indexed" }}}]).pretty();*





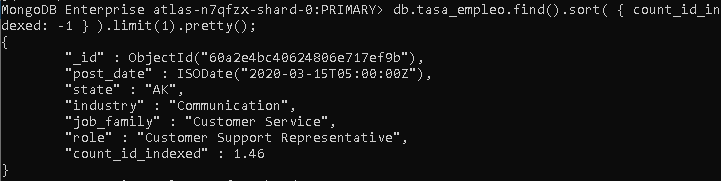




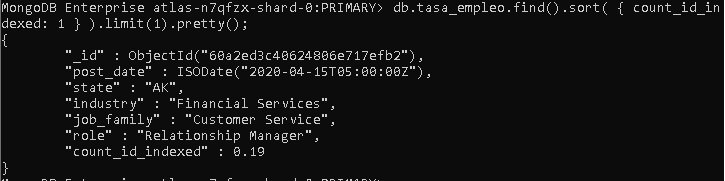
**MÁXIMOS Y MÍNIMOS**

* ¿Cuales fueron los registros con la mejor y peor tasa de nuevos anuncios? (máximo y mínimo por tasa)

*db.tasa\_empleo.find().sort( { count\_id\_indexed: -1 } ).limit(1).pretty();*

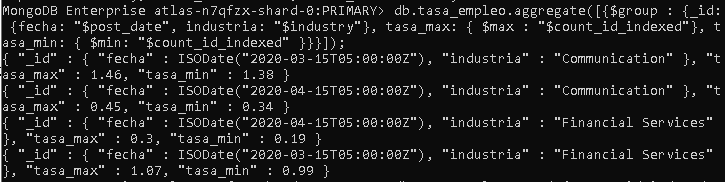


*db.tasa\_empleo.find().sort( { count\_id\_indexed: 1 } ).limit(1).pretty();*



* ¿Cuales fueron las tasas maximas y minimas registradas por industria por cada mes?

*db.tasa\_empleo.aggregate([{$group : {\_id: {fecha: "$post\_date", industria: "$industry"}, tasa\_max: { $max : "$count\_id\_indexed"}, tasa\_min: { $min: "$count\_id\_indexed" }}}]);*



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 bases de datos relacionales no están diseñadas para la velocidad de la agilidad empresarial porque no almacenan sólidamente las relaciones entre los elementos de datos, es decir, presentan falta de coincidencia entre las bases de datos relacionales y las relaciones de los datos. Las relaciones de los datos son a menudo más valiosas que los datos en sí. El problema es que dichas bases de datos no están diseñadas para capturar esa relación por esto las empresas que las crean están perdiendo conexiones esenciales para las decisiones actuales basadas en datos y que tienen alta complejidad y riesgo. Adicional, las bases de datos relacionales no se adaptan bien a los cambios porque su esquema fijo funciona mejor para los problemas que están bien definidos desde el principio ya que estos cambios de esquema son un verdadero problema para los desarrolladores y tienen un alto costo de mantenimiento.

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

**R/:**

Se deben considerar: el crecimiento ya que cuando las bases de datos tienden a crecer demasiado en el almacenamiento y el mantenimiento es sumamente difícil y costoso, suelen presentar fallas en tiempo de respuesta. La elección al más adecuado porque el mercado de RDBMS, se pueden encontrar servicios gratuitos o comprados, dependiendo de varios factores, como pueden ser, cantidad de información, tiempo de disponibilidad, fuerza de operación, etc., Cada solución de RDBMS que se presente en el mercado ofrece también ventajas y desventajas, no precisamente lo más caro vaya a ser la solución a todo. Finalmente, la complejidad en la instalación, pues algunos RDBMS dependen del sistema operativo donde se vayan a instalar, no garantizan el buen funcionamiento si no cumplen con los requerimientos mínimos de instalación.

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

**R/:**

Primero, el rendimiento pasa de minutos a milisegundos ya que una base de datos de gráficos convierte JOINs complejos en recorridos rápidos de gráficos, que son operaciones de tiempo constante manteniendo así el rendimiento en milisegundos independientemente del tamaño total del conjunto de datos. Segundo, los ciclos de desarrollo drásticamente acelerados porque reduce la sobrecarga de desarrollo de la traducción de ida y vuelta entre un modelo de objeto y un modelo relacional tabular. Tercero, la capacidad de respuesta empresarial extrema pues permite que su base de datos de gráficos evolucione al mismo ritmo que el resto de la aplicación y los requisitos comerciales cambiantes. Por último, está preparado para empresas porque emplea en una aplicación de misión crítica, una tecnología de datos debe ser robusta, escalable y, la mayoría de las veces, transaccional.