PLATAFORMAS DE VISUALIZACIÓN DE DATOS TOLERANTES A FALLOS POR MEDIO DE MONGODB

FAULT TOLERANT DATA VISUALIZATION PLATFORMS THROUGH MONGODB

Edy Gómez Coaboy, MSc.
https://orcid.org/0000-0001-7923-2311
Magister en Data Science (Ecuador).
Consultor en Data Science, Quito, Ecuador.
elgomez.mdat@uisek.edu.ec
Joe Carrión Jumbo, Ph.D.
https://orcid.org/0000-0003-3632-5352
Doctor en Informática (España).
Docente titular de la Universidad Internacional SEK, Quito, Ecuador.
joe.carrion@uisek.edu.ec

ARTÍCULO DE INVESTIGACIÓN
Recibido: 24 de noviembre de 2020
Aceptado: 2 de marzo de 2021

RESUMEN

El siguiente trabajo de investigación analiza las funcionalidades de un sistema distribuido para visualización de datos, por medio de un clúster de almacenamiento tolerante a fallos que permita el análisis estadístico, mediante la integración de herramientas. Se utiliza para experimentación con datos abiertos sobre matrimonios correspondientes al 2018. Se ha integrado MongoDB Atlas y Compass, que permiten la visualización de datos en un esquema distribuido en la nube. Se realiza la carga de datos que serán mostrados mediante gráficos y análisis detallado de cada uno. El uso de MongoDB Atlas para el modelado de documentos resulta simple de desplegar y proporciona funcionalidades para una mayor adaptabilidad escalable a los desarrolladores y analistas de datos. Posee interfaces gráficas útiles al momento de trabajar con bases de datos, también interpreta diferentes lenguajes, siendo una de las herramientas que ayudará en la actualidad y futuras generaciones al tratar con modelos de consultas. Enfocado para fines educativos con el objetivo de realizar la visualización de los datos y hacer réplicas de la información mediante servidores en la nube. Los resultados muestran las funciones de clúster tolerante a fallos con herramientas de visualización ágiles y simples de implementar, por lo que

https://doi.org/10.21855/ecociencia.82.386

es una alternativa a evaluar para las organizaciones con necesidades de plataformas de visualización.

Palabras clave: clúster, sistemas distribuidos, bases de datos, visualización de datos.

ABSTRACT

The following research work analyzes the functionalities of a distributed system for data visualization, by means of a fault-tolerant storage cluster that allows statistical analysis, through the integration of tools. It is used for experimentation with open data on marriages for 2018. MongoDB Atlas and Compass have been integrated, which allow the visualization of data in a distributed schema in the cloud. The data is loaded that will be shown through graphs and detailed analysis of each one. Using MongoDB Atlas for document modeling is simple to deploy and provides functionality for greater scalable adaptability to developers and data analysts. It has useful graphical interfaces when working with databases, it also interprets different languages, being one of the tools that will help today and future generations when dealing with query models. Focused for educational purposes in order to visualize the data and make replicas of the information through servers in the cloud. The results show fault-tolerant cluster functions with agile visualization tools that are simple to implement, making it an alternative to evaluate for organizations with visualization platform needs.

Keywords: cluster, distributed systems, databases, data visualization.

INTRODUCCIÓN

Se realizará la integración con MongoDB Compass y Atlas, herramientas tecnológicas que permiten una adaptabilidad muy eficiente al tratar con modelos de documentos y visualización de los datos. MongoDB soporta replicación de base de datos distribuidas y partición horizontal, el modelo de réplicas consiste en un nodo principal que acepta operaciones de escritura y las propaga hacia los nodos restantes (réplicas), (Tyulenev , y otros, 2019).

El rendimiento de este esquema ha sido analizado y comparado con herramientas similares como Cassandra y MongoDB (Haughian, Osman, & Knottenbelt, 2016), (Hammood & Murat, 2016), (Baruffa, Femminella, Pergolesi, & Reali, 2019), Postgresql y MongoDB (Makris, Tserpes, Spiliopoulos, & Anagnostopoulos, 2019), HBASE y MongoDB (Matallah, Belalen, & Bouamrane, 2017), acerca de comparación de varias herramientas y entre ellas MongoDB (Hammood & Murat, 2016). Para este trabajo se utilizará el modelo de réplicas, para la implementación se ha utilizado la guía oficial de MongoDB y los lineamientos de (Giamas, 2017).

La creación de clúster de datos, configuración de comunicación, conexión de internet, creación de nodos, creación usuarios y muestra de datos mediante gráficos, se realizará en MongoDB Atlas. Para la conexión con MongoDB Atlas y la carga de la base de datos se realizará mediante la interfaz gráfica de MongoDB Compass. Estas herramientas ayudan a los desarrolladores, analistas de datos e investigadores a tener una manera más ágil de tratar los datos y realizar clúster de servidores. Es por tal razón que se han utilizado estas tecnologías de información.

Se requiere analizar las funciones de clústeres tolerantes a fallos para desplegar una plataforma de visualización de datos en la nube. Es necesario diseñar, construir y poner en marcha un clúster de MongoDB para almacenamiento de datos distribuidos e integrarlo a herramientas de visualización de datos, para luego realizar un análisis estadístico del mismo. Se utilizará como caso de aplicación fuentes de datos abiertas publicadas por el INEC sobre matrimonios del año 2018 de todas las provincias y cantones del Ecuador.

Se busca analizar las funciones de MongoDB por medio del diseño de un clúster para el almacenamiento de datos distribuidos para la visualización de datos y análisis estadístico con replicación de la información en nodos redundantes a fin de disponer de un sistema tolerante a fallos.

Mediante la creación del clúster de almacenamiento de datos distribuidos se podrá realizar la replicación de la base de datos en varios nodos utilizando MongoDB Atlas por medio de 2 nodos. Con fines de analizar las funciones, se utilizará un ambiente controlado con dos nodos en la nube para validar la replicación y tolerancia a fallos. Los datos de un caso de aplicación corresponden al año 2018. El estudio no está enfocado al rendimiento.

METODOLOGÍA

Para el presente trabajo se ha utilizado para la evaluación de las características de software, el Estándar ISO/IEC 25040 (Organización Internacional de Normalización, 2011) y para el análisis de datos un caso de aplicación, por medio del uso de datos abiertos del INEC.

La descripción detallada de los elementos evaluados de acuerdo al estándar ISO/IEC 25040 y los criterios elegidos se muestra detallados en el Anexo 1. La Tabla 1 un resumen de las etapas establecidas en el estándar.

Tabla 1

Etapas de estándar ISO-IEC-25040.

N	Descripción
1	Establecer los requerimientos de evaluación
2	Especificar la Evaluación
3	Diseñar la Evaluación
4	Ejecutar la Evaluación

Fuente: Elaboración propia.

La Etapa 1, se fundamenta en el objetivo general del presente trabajo acerca de analizar las funciones de MongoDB por medio del diseño de un clúster. En la Etapa 2, se define los elementos del software que se detallan en la Tabla 2. La Etapa 3, se resume en el Anexo 1 y la Etapa 4 se resume también en el Anexo 1 y se detalla y analiza en el presente documento.

Tabla 2

Componentes a analizar

N	Descripción
1	Gestión de Clúster
2	Carga de Datos
3	Gestión de usuarios
4	Visualización de datos

Fuente: Elaboración propia.

Para probar las funciones con el caso de aplicación y analizar los datos se utilizó la metodología KDD (Fayyad, Shapiro, & Smyth, 1996), una de las principales características es que permite descubrimiento de conocimientos de pequeñas y grandes bases de datos, identificando patrones de gran significación estadística, esto implica mucho en la interpretación de patrones y modelos, ayudando a la toma de decisiones, ya sean personales u organizacionales (Rodríguez & García , 2016). La metodología KDD ha sido ampliamente documentada en (Liu & Hiroshi, 2012) y (Maimon & Rokach, 2009).

El proceso de KDD, se encuentra establecido en 5 etapas:

- La selección de fuentes de datos.
- El procesamiento y limpieza de los datos.
- La transformación y reducción de los datos.
- La minería de datos.
- La interpretación de los datos obtenidos.

Fuente de datos

Se determinó la base de datos que se va a utilizar para la exploración y el análisis respectivo. Se escogió la base de datos de interés social de la realidad del Ecuador. Para el presente reporte técnico se utilizan datos sobre los matrimonios correspondientes al año 2018, publicados por el Instituto Nacional de Estadísticas y Censos (INEC, 2018). Ver Figura 1.

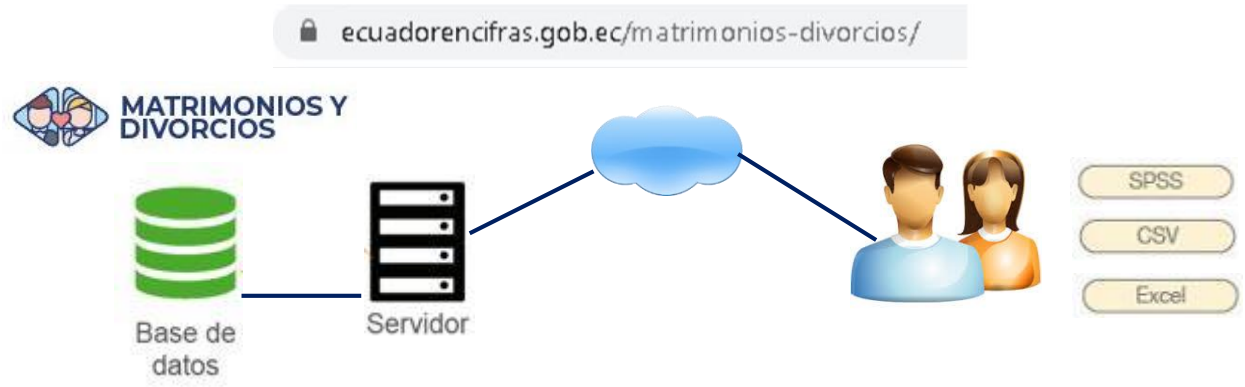

Figura 1. Enlace de descarga de los datos publicados por el INEC. Fuente: Elaboración propia.

Limpieza de datos

En primer lugar, se realizó un proceso de ETL (Extracción, Transformación y Carga), para obtener resultados confiables; se utilizó software de Hojas de Cálculo y Edición de Texto, con el fin identificar datos faltantes, a continuación, se remplazó separadores de datos y eliminación de espacios. El proceso ETL permite obtener una mejor interpretación de los datos y utilizarlos sin problemas de procesamiento. Este proceso está basado en lineamientos de (Kalman & Rendón, 2016). Ver Figura 2.

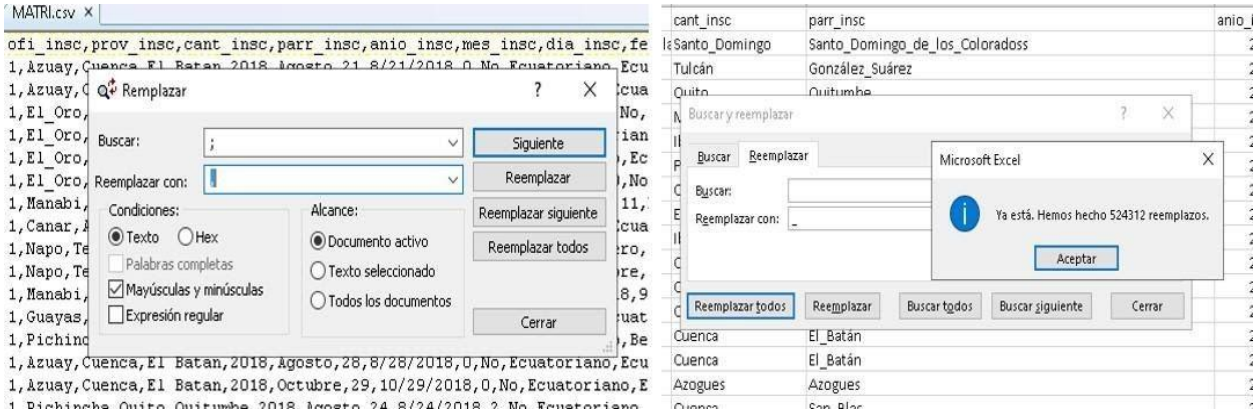

Figura 2. Procesamiento y limpieza de la base de datos. Fuente: Elaboración propia.

Transformación/Reducción

El objetivo del proceso de transformación se basa en extraer datos de cualquier fuente y luego convertirlos a un formato explotable y hacerlos llegar a un destino (Merlino, 2014).

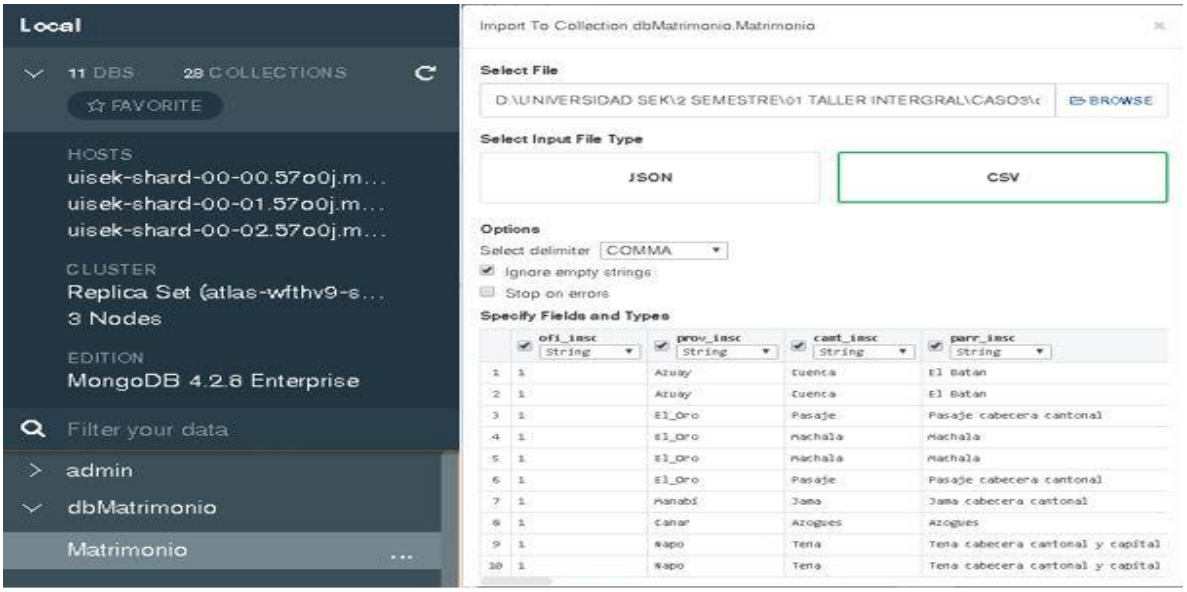

Figura 3. Transformación de los datos. Fuente: Elaboración propia.

Para este proceso se hizo la carga de datos a MongoDB Compass, en donde se importó de manera local la base de datos en formato CSV, para luego visualizarla en MongoDB Atlas en formato JSON. Ver Figura. 3.

Minería de Datos

Conceptualmente se refiere al proceso de extraer conocimiento útil y comprensible de los datos, previamente desconocido, a partir de grandes volúmenes de datos, con el objetivo de determinar el comportamiento de estos datos (Vallejo, Guevara, & Medina, 2018).

Durante el proceso de minería de datos se aplicaron técnicas que permitieron interpretar la base de datos. Una de las técnicas de minería de datos que se aplicó fue la exploración de los datos. Ver Figura 4.

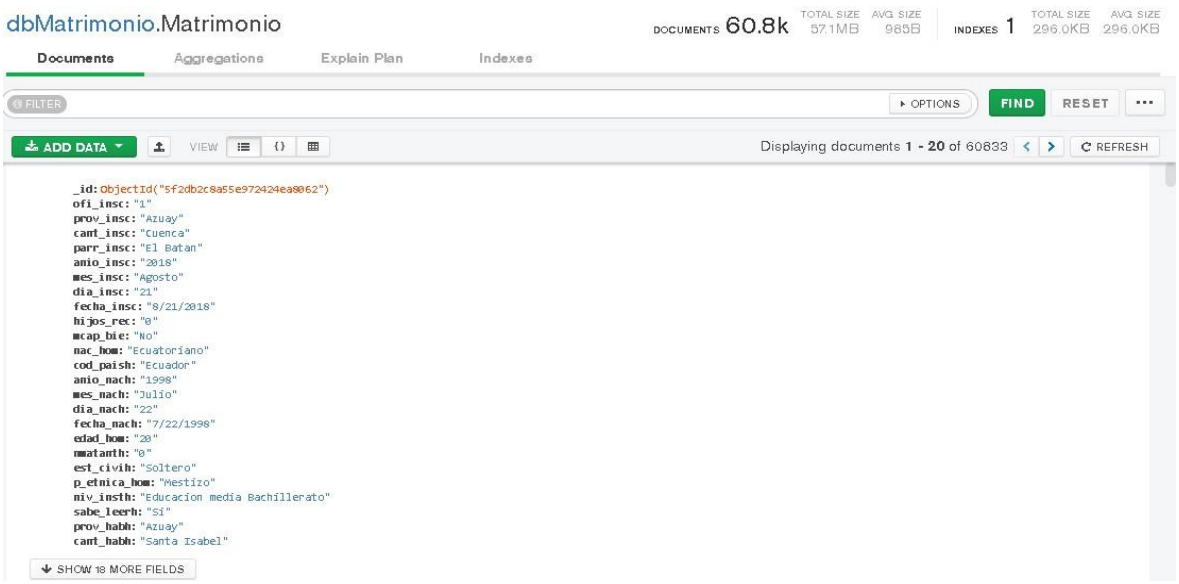

Figura 4. Exploración de datos. Fuente: Elaboración propia.

Representación de Datos

Para esta etapa se realizó la visualización de la base de datos en MongoDB Atlas luego de ser importada en MongoDB Compass. Se puede Visualizar en formato JSON (MongoDB, 2020). En la siguiente sección de Diseño Experimental se mostrarán cada uno de los gráficos con su respectivo análisis. Ver Figura 5.

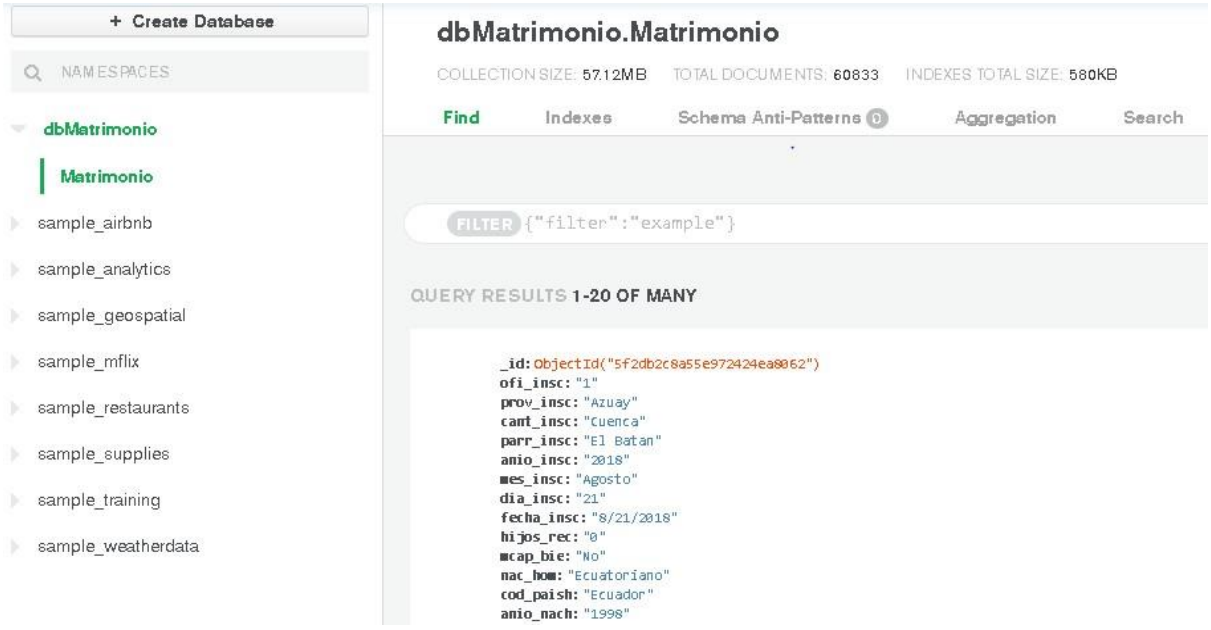

Figura 5. Carga de Datos en MongoDB Compass. Fuente: Elaboración propia.

MODELOS DE DATOS

Se presentarán dos tipos de modelos que fueron diseñados para tener una mejor interpretación de la funcionalidad de la base de datos. A continuación, se define la estructura física y lógica (Piedrabuena, 2007).

Modelo Físico

Se cuenta con una sola tabla en el cual se encuentra registrada toda la información referente a la base de datos de Matrimonios, para realizar este proceso de ETL (Extract, Transform, Load) se utilizó Pentaho Server versión 8.1, PDI (Process Data integration) y la herramienta Spoon (Ioana , Diaconita, & Bologa, 2015) .La finalidad de esta transformación de datos se la efectuó para llevar de un archivo CSV a una base de datos. Se crearon dos procesos, uno de entrada y otro de salida. Ver Figura 6.

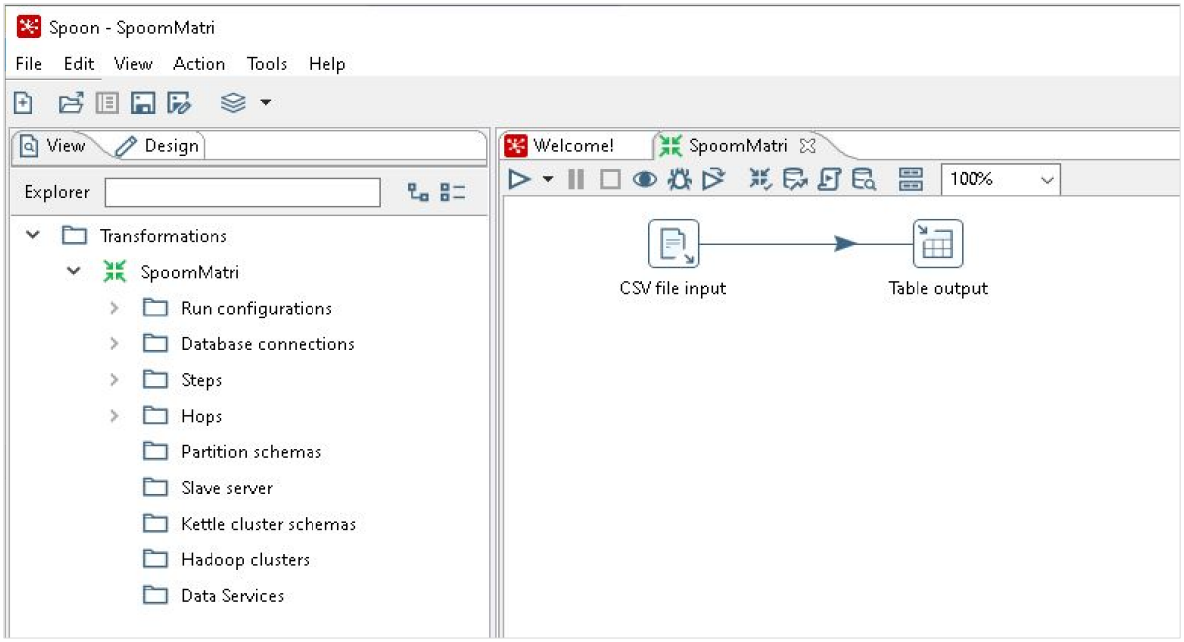

Figura 6. Creación para el archivo de entrada y salida en Spoon. Fuente: Elaboración propia.

Archivo de entrada: La transformación, aplicando el proceso de PDI, inicia con la creación de una Nueva Transformación y seleccionado dos elementos, el primero es un CSV Input y el segundo es una Tabla Output. En CSV Input, se elige el archivo de origen y se selecciona el formato de origen como UTF-8. Se realiza la carga de datos, se eligió la base de datos de los matrimonios del 2018, se efectúa una previsualización de 5.000 datos. Luego se selecciona *Get Fields (Obtener campos)*, se elige también la cantidad de 5.000 registros (es recomendable ya que se va a realizar una exploración de datos extensa). Ver Figura 7.

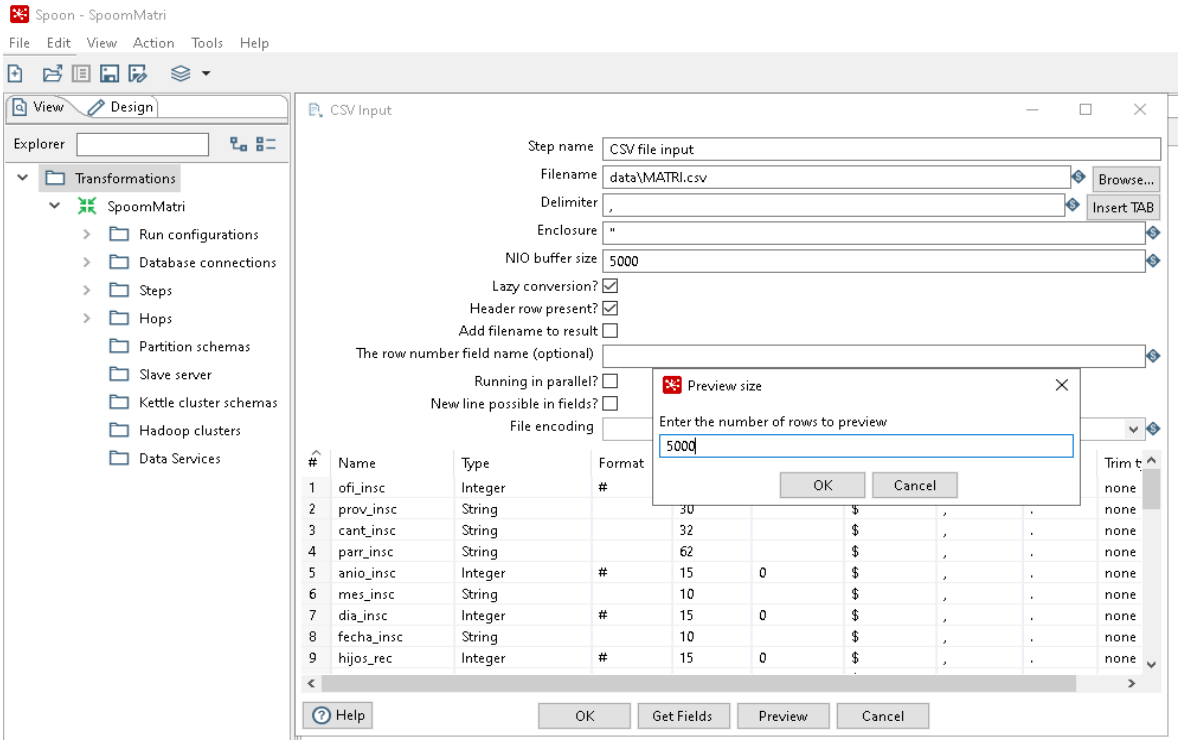

Figura 7. Carga del archivo en formato CSV de matrimonios 2018 utilizando Pentaho Data Integration. Fuente: Elaboración propia.

Tabla de salida: Se realiza la conexión a MS SQL Server, primero se requiere que tenga una base de datos de destino previamente creada en el servidor de SQL Server. Configurar el servidor, puede ser la dirección IP o el nombre del host y la instancia de SQL. Ver Figura 8.

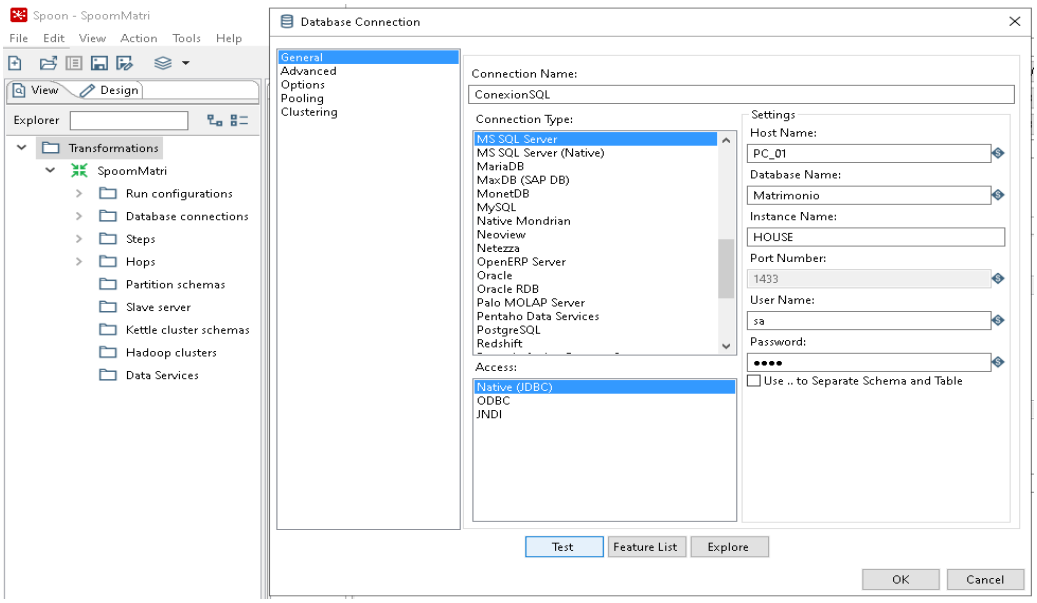

Figura 8. Conexión de Spoon a MS SQL Server. Elaboración propia.

Seleccionar la opción “Specify database fields” para indicar los nombres de los campos, “Get fields” para leer los campos que se definieron en el archivo, seguidamente en SQL seleccionar “Execute”. Este proceso crea la base de datos en MS SQL Server. Ver Figura 9.

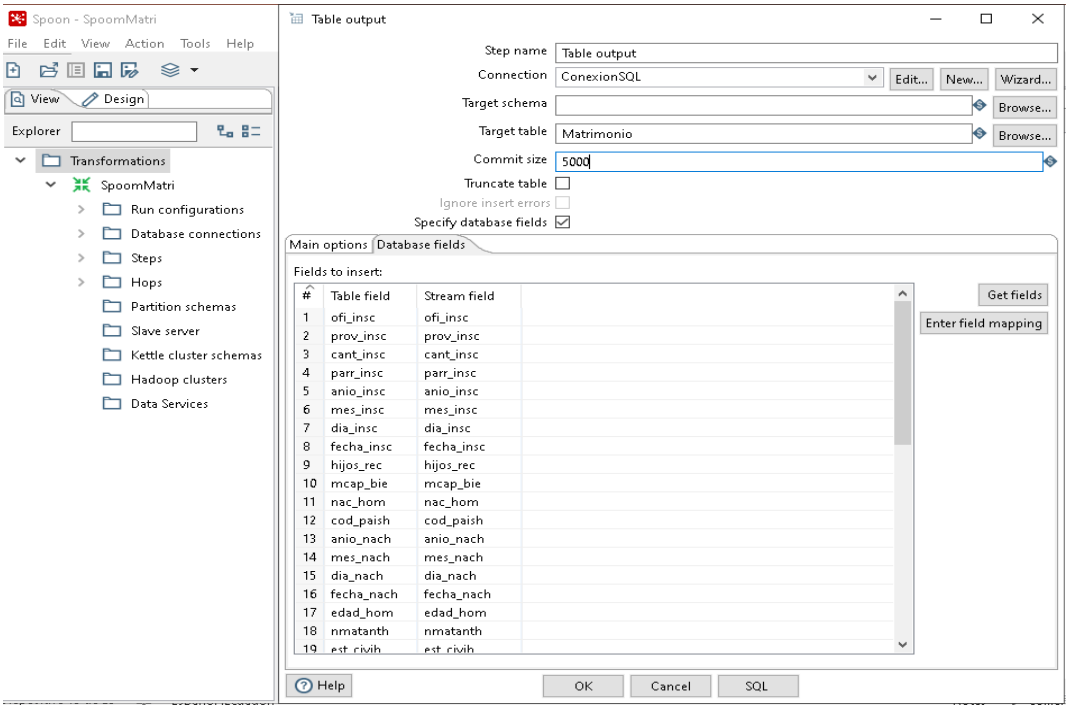

Figura 9. Crear la base de datos en MS SQL Server. Elaboración propia.

Para enviar los datos y realizar el proceso de transformación se inicia la ejecución. Se recomienda verificar en MS SQL Server que se haya creado la tabla e insertado los datos, ver Figura 10. Esta información será utilizada para crear el clúster de almacenamiento.

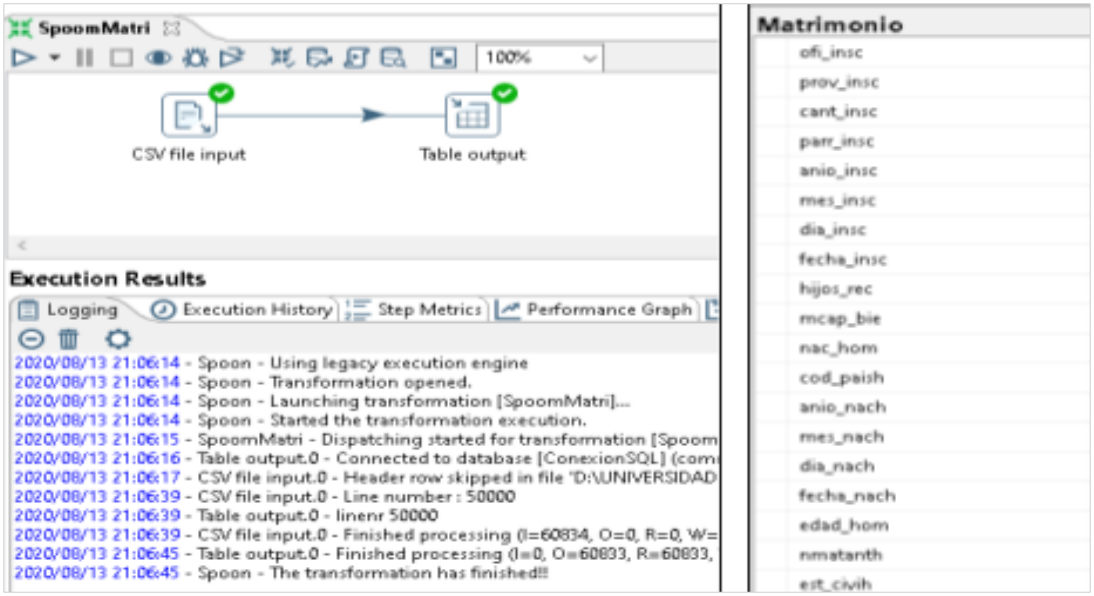

Figura 10. Ejecución y creación de la base de datos en SQL Server. Fuente: Elaboración propia.

Modelo Lógico

La estructura del clúster en un ambiente real podría ser por medio de equipos físicos interconectados a la misma red. Otro esquema puede ser en ambiente virtual, en el que los nodos pueden estar desplegados sobre la misma plataforma de virtualización. Un escenario recomendado es con sistema operativo (únicamente en modo línea de comandos ya sea Linux o Windows) dedicado para cada servicio de MongoDB.

En este modelo se define con la estructura lógica de una base de datos y cómo será la interfaz de acceso para los usuarios (Tello, 2015) .

1. *Usuario Admin2:* Tiene los permisos de administrador, realiza peticiones al servidor.
2. *Usuario Invitado :* Tiene los permisos de lectura y escritura para la base de datos, realiza las peticiones al servidor.
3. *Usuario Admin :* Contiene todos permisos de administrador del servidor principal y realiza peticiones a los servidores que contienen la replicación de los datos como también a los servidores administrados de MongoDB Atlas. Ver Figura 11.

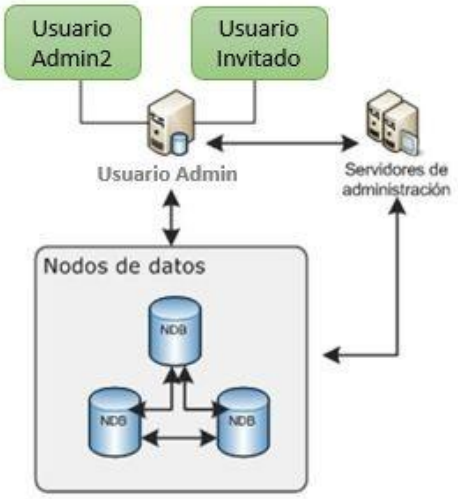

Figura 11. Modelo lógico de la base de datos. Fuente: Elaboración propia.

DISEÑO EXPERIMENTAL

En este apartado se realizará la explotación y uso de los recursos configurados. Se demostrará mediante gráficos y análisis el uso de MongoDB Compass y las funciones de visualización disponibles.

Iniciar Base de datos de MongoDB

Se inicia el proceso de gestión de la base de datos mediante la línea de comandos (para el experimento se utilizó la línea de comandos de Windows). La ejecución debe hacerse con perfil de Administrador. El programa para invocar es “mongod.exe”. Ver Figura 12.

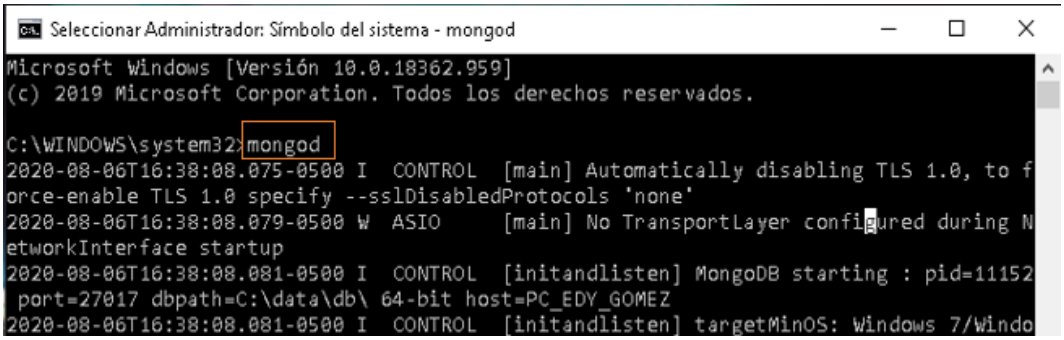

Figura 12. Inicio del núcleo de gestión de la Base de datos con MongoDB. Fuente: Elaboración propia.

En este apartado se demostrará la explotación y uso de los recursos configurados. Se demostrará mediante gráficos y análisis el uso de MongoDB Compass y las funciones de visualización disponibles.

Control de la Base de Datos

Una vez iniciada la base de datos, se procede a ejecutar otra consola de Windows (Línea de comandos), para el cual se ejecuta el comando “mongo”. Este proceso no es obligatorio, si recomendable e importante para determinar si existe conexión al servidor de MongoDB. Ver Figura 13.

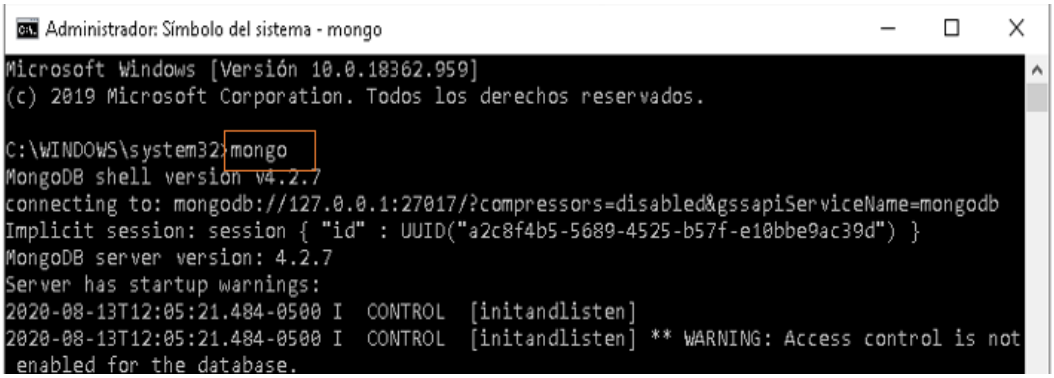

Figura 13. Conexión a la base de datos con MongoDB. Fuente: Elaboración propia.

Uso de la plataforma MongoDB Atlas

MongoDB Atlas es un servidor de base de datos con modelo Software as a Service (SaaS). Este servicio lo provee MongoDB Inc. y el servicio proporciona tres servidores a elección del usuario *Amazon Web Service* , *Azure* y *GCP* (Google Cloud Platform) (Ashraff , 2018). En los siguientes pasos del proceso se utilizará con la interfaz gráfica de MongoDB Atlas. La interfaz de MongoDB Atlas no cuenta con todas las características que la administración mediante línea de comandos, lo cual es una desventaja. Sin embargo, para las funcionalidades que se van a realizar no es inconveniente.

Creación de Usuarios

Es muy importante realizar la creación de los usuarios y asignarles los respectivos roles de interacción con la base de datos (Gómez, 2013). Se crearon tres usuarios:

- *Admin*: Contiene los permisos de administrador.
- *Admin2*: Tiene el respaldo de copia de seguridad del usuario
- *Admin Invitado*: Tiene los permisos de lectura y escritura para la base de datos.

Donde el usuario administrador *Admin* cuenta con todos los privilegios (editar, eliminar y consultar). Ver Figura. 14.

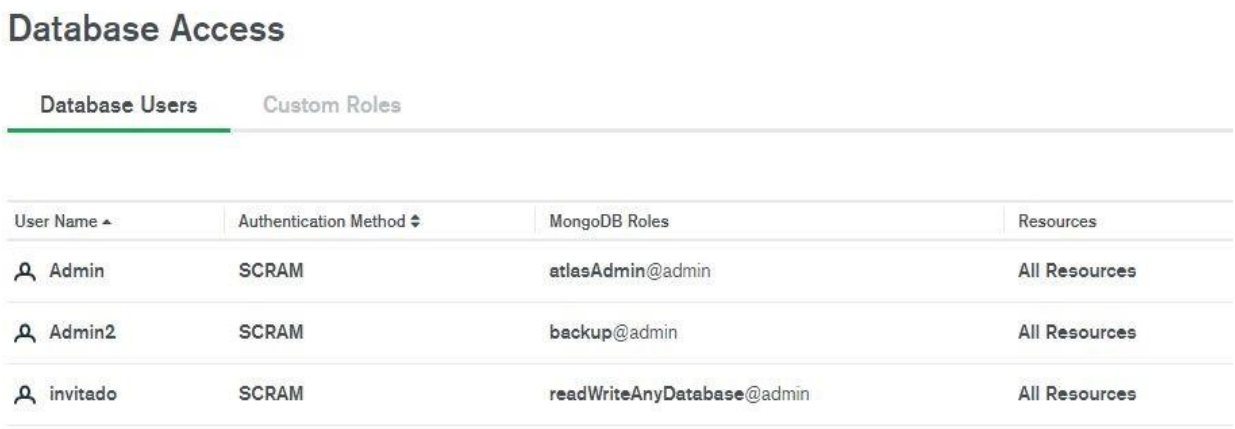

Figura 14. Creación de usuarios en MongoDB Atlas. Fuente: Elaboración propia.

Conexión al Servidor

Es de gran utilidad realizar esta configuración que permite la salida a Internet del servidor Web, ya que permite la conexión con la red y con aplicaciones, es decir cualquier dispositivo con dirección IP se puede conectar (Vázquez, 2015). Para este trabajo se dejó establecido para que cualquier equipo o dispositivo con dirección IP se pueda conectar al servidor. Ver Figura 15.

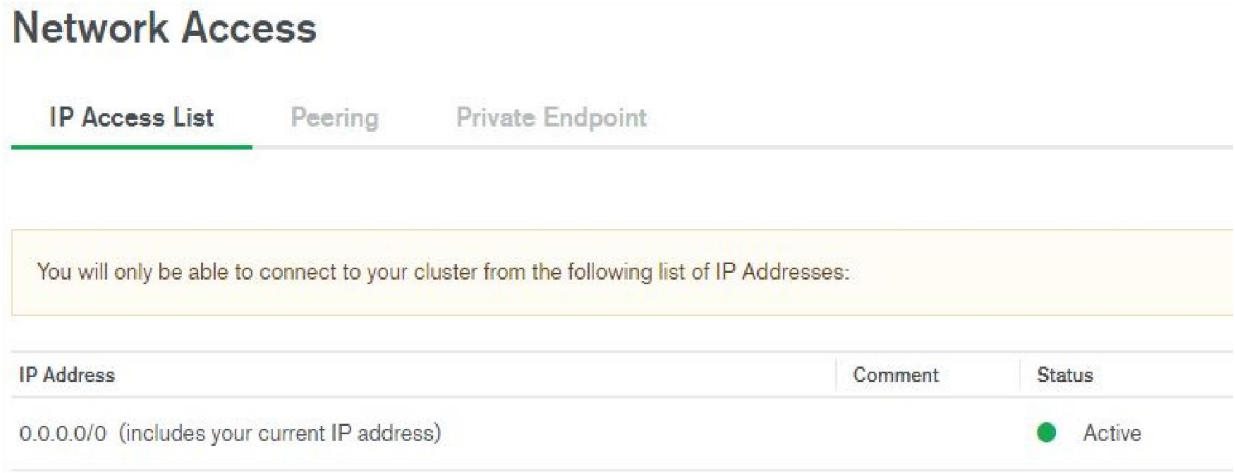

Figura 15. Configuración de conexión del servidor. Fuente: Elaboración propia

Alojamiento del Servidor

El alojamiento de los datos se realizó en la nube con el servicio de MongoDB Atlas. Se utilizó una cuenta experimental sin costo que permite elegir en qué lugar se desea alojar la base de datos. Para el presente informe técnico se utilizó la región Norte Virginia. Ver Figura 16.

Figura 16. Alojamiento en servidor MongoDB con el modelo SaaS. Fuente: Elaboración propia.

Replicación de Datos

Mantener los datos distribuidos en diferentes servidores mejora el rendimiento en el tiempo de respuesta, este esquema es tolerante a fallos y mejora la disponibilidad del sistema. Se puede tener miles de peticiones al servidor simultáneamente y no se ve afectado (Power, 2016). Se implementó una replicación con tres nodos, uno primario y dos secundarios como se muestra en la Figura 17.

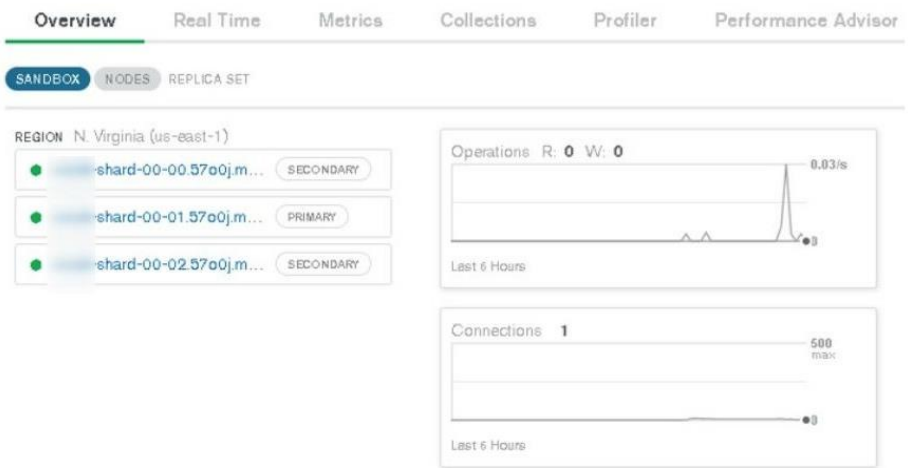

Figura 17. Monitoreo de los nodos del clúster de MongoDB. Fuente: Elaboración propia.

Nodo Primario

Es el nodo principal donde se encuentra almacenada la información de la base de datos de Matrimonios del año 2018. Es el nodo principal del Clúster en responder a las peticiones. Luego se realizan las réplicas a los nodos secundarios. Ver Figura 18.

Figura 18. Monitoreo del Nodo primario del Clúster de MongoDB. Elaboración propia.

Nodos Secundarios

Los dos nodos adicionales mantienen las réplicas de la información, es decir si el nodo principal deja de funcionar en cualquier momento, ellos se encuentran activados para seguir respondiendo a las peticiones. Es muy importante la replicación de los datos ya que se evita el corte del servicio y la ventaja que se tiene, es que mejora el rendimiento de solicitudes. Se puede observar que los dos nodos secundarios se encuentran activos. MongoDB Atlas por defecto los mantiene activados. Ver Figura 19.

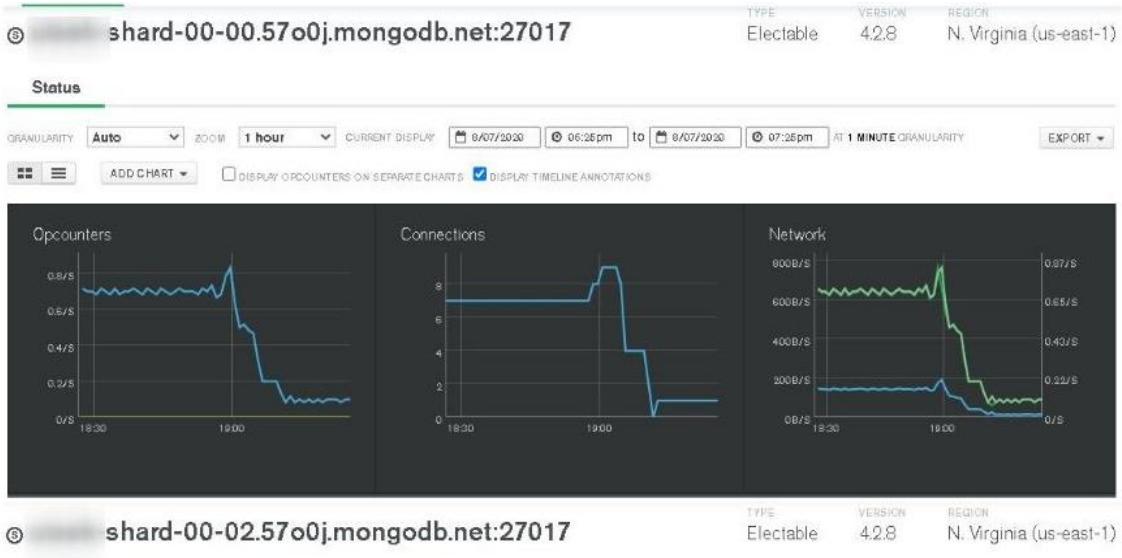

Figura 19. Monitoreo del Nodo secundario del Clúster de MongoDB. Elaboración propia.

RESULTADOS

En esta sección se presentan los resultados obtenidos de la base de datos del INEC sobre Matrimonios del año 2018, mismos que son visualizados mediante gráficos, utilizando la misma plataforma de MongoDB Atlas, que permite la creación de un *Dashboard*. Cabe recalcar que se puede realizar en otra herramienta de visualización de datos (Power BI, Tableau, aplicaciones de R Studio Shiny, etc.), pero en el presente trabajo se describe el proceso con la misma herramienta. Esta información estará ubicada en el servidor web de MongoDB Atlas.

Caso de Aplicación: Cantidad de Matrimonios por Provincias

Fue uno de los primeros análisis que se realizó, donde se deseaba conocer las cantidades de matrimonios por provincias. Como se puede ver Guayas y Pichincha son las dos provincias con mayor cantidad de habitantes, por tal razón existe mayor cantidad de matrimonios. Se observa que Guayas ocupa el primer lugar con un total de 16.256 matrimonios en el año 2018. Ver Figura 20.

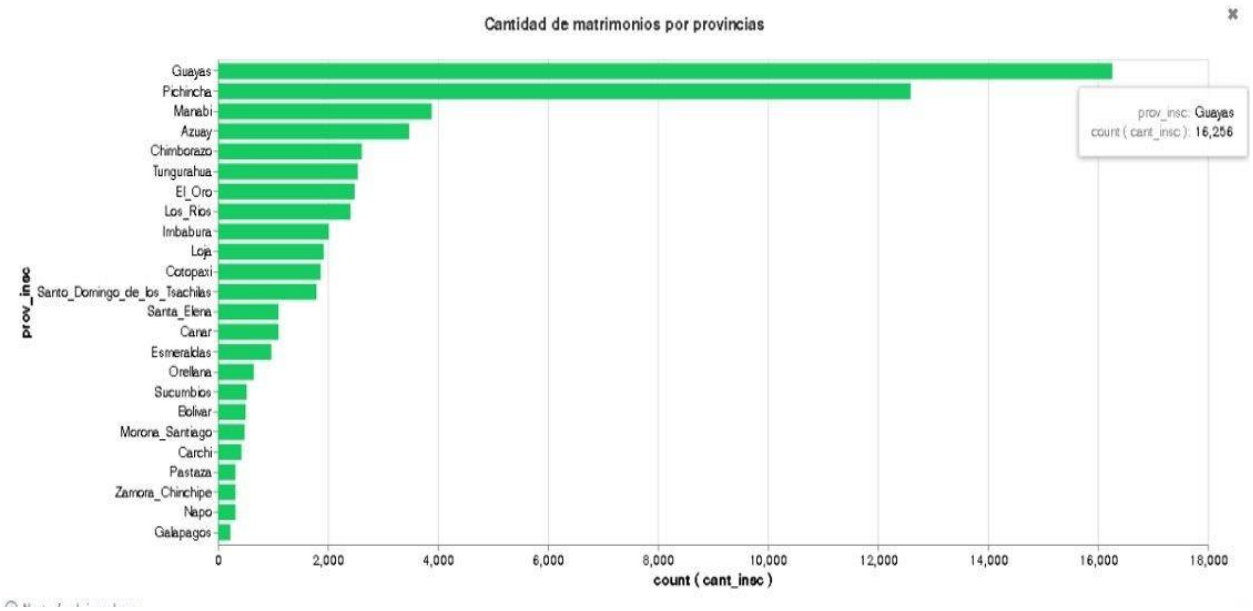

Figura 20. Gráfico sobre la cantidad de matrimonios por provincia. Elaboración propia.

Caso de Aplicación: Mes de Inscripción

Para este análisis se eligió determinar la cantidad de inscripción de matrimonios por mes, para lo cual se tomó en consideración a la provincia de Esmeraldas. Se puede visualizar que durante el mes de agosto fue donde tuvieron más inscripciones de matrimonios.

En los experimentos de evaluación de la herramienta se plantea conocer en qué cantón existen mayores índices y se observa que el cantón Esmeraldas se encuentra mayor cantidad con 59 inscripciones seguidamente se encuentra el cantón Quinindé. Ver Figura 21.

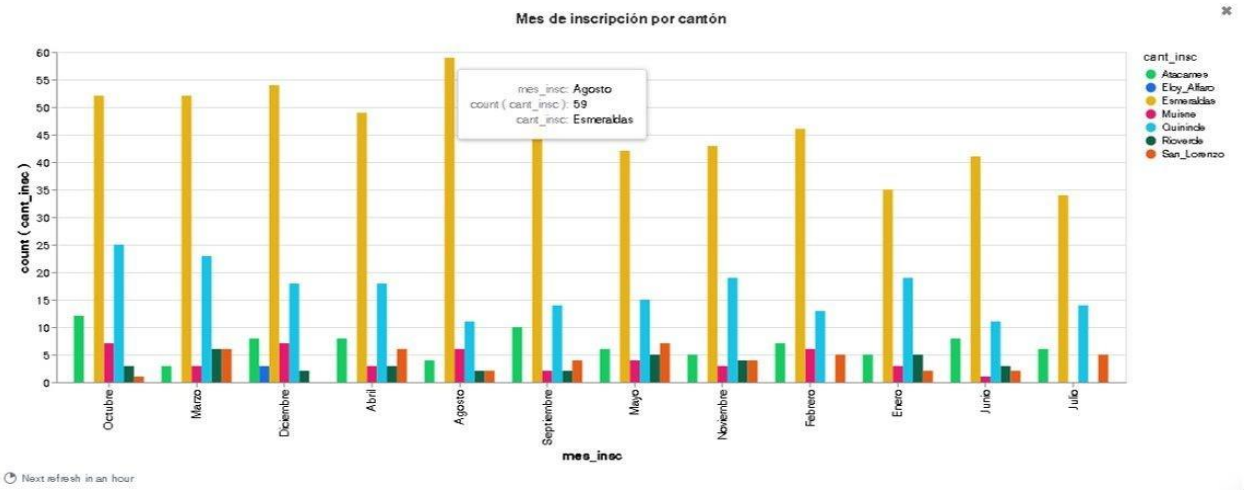

Figura 21. Gráfico mes de inscripción en la provincia de Esmeraldas. Elaboración propia.

Caso de Aplicación: Nivel de Instrucción

Para este análisis se tomaron en cuenta ambos sexos para determinar el nivel de instrucción que tienen las personas que contrajeron matrimonio en el año 2018 de la provincia de Esmeraldas. Se puede observar para ambos géneros las cantidades son similares, es decir que el nivel de instrucción que más sobresale es la Educación Media (Educación Bachillerato), para el sexo Masculino con un total de 456 personas y para el sexo Femenino con 457 personas. Ver Figura 22.

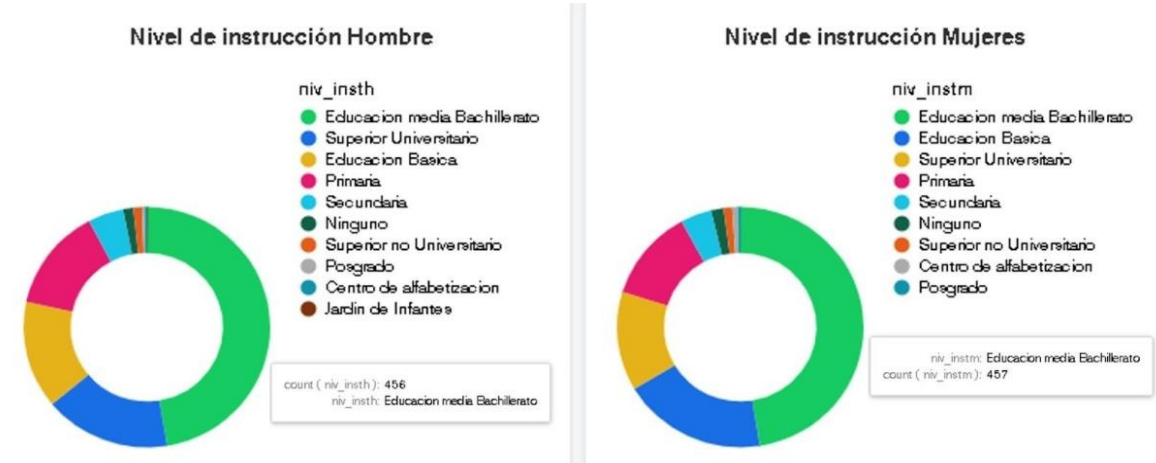

Figura 22. Gráfico nivel de instrucción provincia de Esmeraldas. Elaboración propia.

DISCUSIÓN

MongoDB Atlas contiene muchas características importantes que no se encuentran en la versión libre, es decir, se puede contratar servidores dedicados con una contratación mensual de 57 dólares y si se desea tener servidores en varias regiones se requiere una contratación mensual de 95 dólares (Julio del 2020). Con la operatividad de todas las funcionalidades de MongoDB Atlas.

- Se analizaron las funciones de MongoDB.
- Se implementó el esquema de Tolerancia Fallos.
- Se evaluaron las funcionalidades de Visualización.
- Se probó la simplicidad del modelo de replicación.

La Tabla 3, resume las funciones analizadas de acuerdo a requerimientos integrales para una plataforma de visualización.

Tabla 3

Resultados de funciones específicas analizadas.

N	Descripción	Resultado
1	Descarga de Aplicación	Local simple. Remota, no aplica
2	Documentación sobre instalación	Para modo local y en la nube disponible
3	Documentación sobre configuración	Para modo local y en la nube disponible
4	Proceso de instalación	Local, se recomienda virtualización. Sistema Operativo Mínimo
5	Configuración para Replicación	Local, complejidad media.
6	Despliegue de nodos	Local, complejidad media.
7	Verificación de replicación	Remota, configuración con complejidad simple
8	Despliegue en la Nube	Remota, configuración con complejidad simple
9	Monitoreo de servicios	Remota, configuración con complejidad simple
10	Configuración de Usuarios	Local y remota simple
11	Carga de datos en la Nube	Local y remota simple
12	Herramientas de Visualización	Local: complejidad alta, en la Nube complejidad simple
13	Publicación de Visualización	Local, complejidad alta, en la Nube complejidad simple

Fuente: Elaboración propia.

La Tabla 4 resume valoración general para cada componente analizado.

Tabla 4

Resultado general

N	Descripción	Resultado
1. Gestión de Clúster	1) Operativo, replica los datos, operativo en fallos de un nodo.	Complejidad media
2. Carga de Datos	Se cargan conjunto de datos de al menos 100k registros. Carga de acuerdo al caso de aplicación	Simple
3. Gestión de usuarios	Se crean los usuarios con los perfiles aceptados. Se verifica las funciones de acceso de acuerdo a cada perfil	Simple
4. Visualización de datos	Se evalúa gráficos de visualización estadística.	Alta

Fuente: Elaboración propia.

Mediante la elaboración del clúster de almacenamiento de datos, se puede concluir que al mantener los datos distribuidos en diferentes servidores es posible mejorar el rendimiento en el tiempo de respuesta (esta medición está fuera del alcance del presente trabajo) y es de gran importancia ya que se pueden tener miles de peticiones al servidor y no se ve afectado al momento de dar respuestas. Se espera como trabajo futuro evaluar por nivel de carga, volumen de usuarios y número de solicitudes.

Al momento que el servidor principal deja de funcionar hay dos servidores secundarios activos que están dando respuestas a las peticiones, es de gran ventaja contar con la réplica de los datos, porque la información debe estar disponible.

Es de gran importancia realizar la exploración y limpieza de los datos antes de comenzar a realizar cualquier análisis. Las herramientas que se utilizaron como: Editor de Archivos Planos, Hojas de Cálculo, Pentaho Data Integration, MongoDB Compass y Atlas se integran de forma ágil.

MongoDB Atlas es una herramienta muy importante para el almacenamiento de datos en la nube y muy útil para los desarrolladores, analistas o quienes deseen interactuar con datos distribuidos en la nube. Cuenta con la ventaja de ser completamente administrada, tiene un sinnúmero de características que permiten utilizar clúster de almacenamiento de datos, mismos que estarán disponibles en varios servidores en la web. Este tipo de análisis es de gran importancia para la formación académica y sirve de referencia a futuros estudios e investigaciones sobre herramientas para desplegar y administrar clústeres empresariales, para servicios de visualización de datos para la toma de decisiones.

CONCLUSIONES

Con el trabajo realizado se evidencia las funciones de MongoDB en un ambiente de replicación de datos por medio de la creación de dos nodos integrados como un clúster. El esquema permite que los datos se repliquen entre los nodos por lo que es factible continuar con la operación aún en el fallo de uno de los nodos. Esta función es apropiada para los requerimientos organizacionales.

Se ha realizado el despliegue de MongoDB con las funciones de minería de datos de forma ágil y las funciones de visualización han sido probadas por su simplicidad y nivel de automatización para generar gráficos en base a la estructura de los datos.

La replicación se logró por medio de la modificación de datos en uno de los nodos y se verificó la actualización inmediata de los datos. Sin embargo, el objetivo no ha sido medir el rendimiento, por lo que es una línea de investigación que resulta muy interesante para el futuro con diferentes escenarios, de capacidad de nodos y volumen de datos.

El proceso de análisis requiere una previa revisión de la calidad de la base de datos, por lo que fue necesario aplicar técnicas de extracción, transformación previa por medio de otras herramientas.

Se utilizó las funciones de visualización para análisis descriptivo de datos y MongoDB permite creación de forma ágil de diferentes alternativas de visualización. Para el futuro se ha visto necesario profundizar la evaluación con nuevas cargas de datos para evaluar otras funciones.

Con al análisis de las funciones de MongoDB se concluye que es una alternativa para el uso en ambiente empresarial a considerar en los procesos de selección de plataformas de visualización de datos de acuerdo a las necesidades de cada organización

REFERENCIAS BIBLIOGRÁFICAS

Ashraff , H. (2018). *rear un Cluster de Base de Datos en la Nube con MongoDB Atlas.* Obtenido de https://code.tutsplus.com/es/tutorials/create-a-database-cluster-in-the-cloud-with-mongodb-atlas--cms-31840

Baruffa, G., Femminella, M., Pergolesi, M., & Reali, G. (2019). *Comparison of MongoDB and Cassandra Databases for Spectrum Monitoring As-a-Service.* Perugia: IEEE.

Fayyad, U., Shapiro, G. P., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework., (págs. 1-7).

Giamas, A. (2017). *Mastering MongoDB 3.x: An expert's guide to building fault-tolerant MongoDB applications.* Kindle.

Gómez, M. (2013). *Bases de datos, notas del curso.* Cuajimalpa: Prolongación Canal de Miramontes.

Hammood, A., & Murat, S. (2016). Comparison Of NoSQL Database Systems: A Study On MongoDB, Apache Hbase, And Apache Cassandra. *International Conference on Computer Science and Engineering*, (págs. 626-631). Tekirdağ,.

Haughian, G., Osman, R., & Knottenbelt, W. (2016). Benchmarking Replication in Cassandra and MongoDB NoSQL Datastore. *Springer*, 152-166.

INEC. (2018). *Instituto Nacional de Estadísticas y Censos.* Obtenido de https://www.ecuadorencifras.gob.ec/matrimonios-divorcios/

Ioana , A., Diaconita, V., & Bologa, R. (2015). Data integration approaches using ETL. 19-27.

Kalman, J., & Rendón, V. (2016). Uso de la hoja de cálculo para analizar datos cualitativos. *Redalyc*, 29-48.

Liu, H., & Hiroshi, M. (2012). Feature selection for knowledge discovery and data mining. New York: Springer Science.

Maimon , O., & Rokach, L. (2009). Introduction to knowledge discovery. (págs. 1-15). Boston: Springer.

Makris, A., Tserpes, K., Spiliopoulos, G., & Anagnostopoulos, D. (2019). *Performance Evaluation of MongoDB and PostgreSQL for.* Lisbon: CEUR.

Matallah, H., Belalen, G., & Bouamrane, K. (2017). Experimental comparative study of NoSQL databases: HBASE versus MongoDB by YCSB. *Comput. Syst. Sci. Eng*, *32(4)*, págs. 307-317.

Merlino, H. (2014). *Metodología de transformacion de datos.* Buenso Aires: Reportes Técnicos en Ingeniería del Software.

MongoDB. (2020). *Interact with Cluster Data. Load Sample Data into Your Atlas Cluster.* Obtenido de https://docs.atlas.mongodb.com/sample-data/

Organización Internacional de Normalización. (2011). Calidad de Software y Datos (ISO/IEC 25040). Obtenido de https://iso25000.com/index.php/normas-iso-25000/iso-25040

Piedrabuena, F. (2007). *Diseño lógico y físico de bases de datos.* Uruguay.

Power, D. (2016). *Beneficios de la Replicación de Base de Datos.* Obtenido de https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/beneficios-de-la-replicacion-de-base-de-datos

Rodríguez , C., & García , M. (2016). Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor. *Scielo*, 43-53. Obtenido de http://rus.ucf.edu.cu/

Tello, R. (2015). Base de datos en la ingeniería y los negocios. *Redalyc*, 1-5.

Tyulenev , M., Schwerin, A., Kamsky, A., Tan, R., Cabral, A., & Mulrow , J. (2019). Implementation of Cluster-wide Logical Clock. *SIGMOD* (págs. 1-15). Amsterdam: Storage & Indexing .

Vallejo, H., Guevara, E., & Medina, S. (2018). Minería de Datos. *Recimundo*, 339-349.

Vázquez, S. (2015). Tecnologías de almacenamiento de información en el ambiente digital. *Redalyc*, 3-16.

ANEXO 1.

Aplicación de estándar ISO/IEC 25040.

N	Etapa	Tarea	Tarea / módulo	Criterios de evaluación
1	**Establecer Los requerimientos de Evaluación**			
	1.1	Establece el propósito	Conocer las funciones de MongoDB para clústeres tolerantes a fallos para análisis de datos	N/A
	1.2	Definir los requerimientos de calidad	Nodos del clúster. Funciones de carga de datos Herramientas de visualización	Creados y operativos Carga satisfactoria o no Funciones de gráficos: Lista
	1.3	Identificar partes a ser evaluadas	1. Gestión de Clúster 2. Carga de Datos 3. Gestión de usuarios 4. Visualización de datos Documentación de cada ítem: 1,2,3,4	N/A
	1.4	Definir el rigor de la evaluación	El rigor establecidas se enfoca en la funcionalidad operacional básica, sin criterios de rendimiento. Verificar que el módulo funciones, acepte datos de entrada y genere respuestas	N/A
2	**Especificar la Evaluación**			
	2.1	Seleccionar medidas de calidad	1. Gestión de Clúster 2. Carga de Datos 3. Gestión de usuarios 4. Visualización de datos	1) Operativo, replica los datos, operativo en fallos de un nodo. 2) Se cargan conjunto de datos de al menos 100k registros. Carga de acuerdo al caso de aplicación 3) Se crean los usuarios con los perfiles aceptados. Se verifica las funciones de acceso de acuerdo a cada perfil 4) Se evalúa gráficos de visualización estadística.

	2.2	Definir criterios de decisión para medidas de calidad	En enfoque será de usabilidad de cada módulo: 1. Gestión de Clúster 2. Carga de Datos 3. Gestión de usuarios 4. Visualización de datos	1) Simple, Complejidad media, Complejo 2) Simple, Complejidad media, Complejo 3) Simple, Complejidad media, Complejo 4) Simple, Complejidad media, Complejo
	2.3	Definir criterios de decisión para la evaluación	En enfoque será de usabilidad de cada módulo: 1. Gestión de Clúster 2. Carga de Datos 3. Gestión de usuarios 4. Visualización de datos	1. Realizar el procedimiento de instalación. Monitorear la operatividad de cada nodo Probar la replicación Probar la operatividad en falla de un nodo. 2. Cargar el archivo de datos. Definir si el proceso es Simple, medio o complejo 3, Crear usuarios de cada perfil. Definir si el proceso es Simple, medio o complejo 4. Crear un gráfico de al menos 3 tipos de gráficos. Para cada uno definir si el proceso es Simple, medio o complejo
3	**Diseñar la Evaluación**			
	3.1	Planificar actividades de evaluación	A) Recursos de Hardware: B) Recursos de Software: C) Personal (horas necesarias de cada uno) D) Módulos: 1. Gestión de Clúster 2. Carga de Datos 3. Gestión de usuarios 4. Visualización de datos	A. Computador: Intel XX, Memoria, Procesador, almacenamiento. B. Sistema Operativo, Excel, Editor de textos, Herramienta ETL C.1 Ing. Sistemas. 40 horas. Evaluación 1. Ing. Sistemas 10 horas. Revisión D. 10 horas para cada módulo. 1) Revisión de documentación (Manuales técnicos) y requerimientos. 2) Descarga del software, instalación y configuración 3) Operación: Depuración de datos, carga, configuración, revisar operación.

4	**Ejecutar la Evaluación**			
	4.1	Realizar mediciones	Módulos: 1. Gestión de Clúster 2. Carga de Datos 3. Gestión de usuarios 4. Visualización de datos	Calificación y justificación de calificación, documentación del proceso. Elaboración de informe,
	4.2	Aplicar criterios de decisión para medidas de calidad	Módulos: 1. Gestión de Clúster 2. Carga de Datos 3. Gestión de usuarios 4. Visualización de datos	Criterios de Decisión: 1) Simple: Documentación clara. 2) Complejidad media: Requiere revisar más documentación. 3) Complejo: Requiere revisar bases de datos de conocimiento.
	4.3	Aplicar criterios de decisión para la evaluación	Módulos: 1. Gestión de Clúster 2. Carga de Datos 3. Gestión de usuarios 4. Visualización de datos	Elaborar tabla que detalle todos los resultados
5	**Concluir la Evaluación**			
	5.1	Revisar el resultado de la evaluación	Módulos: 1. Gestión de Clúster 2. Carga de Datos 3. Gestión de usuarios 4. Visualización de datos	Reuniones de revisión de cada módulo.
	5.2	Crear el informe de evaluación		Se documenta el proceso en un informe de operación de cada módulo
	5.3	Revisar la evaluación de la calidad y proporcionar retroalimentación a la organización.		Elaborar informe de Resultados. Se analizar cada objetivo con los resultados y se plantea conclusiones
	5.4	Realizar la disposición de los datos de evaluación		El análisis se documenta en dos informes: 1) Informe del caso de aplicación 2) Artículo de difusión de resultados

Fuente: Elaboración propia.