# Análisis de Performance de Base de Datos Sql y NoSql aplicado a Datos de Entidades Públicas.

Mercedes Barrionuevo[1] and Mariela Rodriguez[2]

[1] Universidad Nacional de San Luis, San Luis, Argentina
[2] Universidad Nacional de Jujuy, Jujuy, Argentina
mdbarrio@unsl.edu.ar
mariela.rodriguez@fi.unju.edu.ar

**Resumen** En los últimos años hemos sido testigos de revoluciones tecnológicas que se suceden a un ritmo tan acelerado que parecen imperceptibles. La era del Big Data ha traído consigo una gran cantidad de datos que necesitan ser almacenados de la forma más eficiente posible y ser recuperados en un tiempo considerablemente rápido. Contar con herramientas de administración de base datos tanto relacional como no relacional es de vital importancia. Esta área es de constante interés para determinar cuál de ellas se comporta mejor u obtiene una mayor performance en un dominio de datos en particular. Finalmente, es preciso concluir si para un dominio de interés, es necesario poner énfasis en la optimización del espacio de almacenamiento o en el tiempo de respuesta de una consulta.

**Palabras claves:** Base de Datos Relacional. Base de Datos No Relacional. Postgress. MongoDB. ElasticSearch. DNRPA

## 1. Introducción

La cambiante situación económica actual del país va dejando ciertas incógnitas del poder económico de sus habitantes. Analizar la compra y los hurtos de autos nuevos o usados, es una alternativa de interés a considerar, la cual nos permitirá obtener algunos indicadores de problemas o tendencias actuales. Utilizar diversas tecnologías de preprocesamiento, almacenamiento y visualización de datos de manera conjunta nos pueden ayudar a tal fin.

La gestión de las bases de datos es fundamental para todos los trabajos de estas áreas. Un sistema de gestión de bases de datos (SGBD) es un programa que permite a uno o varios usuarios acceder a una base de datos [2]. Permite manejar los accesos diferenciados (identificación, seguridad) y permite interpretar las búsquedas para ingresar, modificar o suprimir datos. Se pueden diferenciar 2 grandes familias de SGBD: los SQL y los NoSQL. Para saber cuál tecnología elegir, en este trabajo vamos a modelar consultas y determinar su velocidad de respuesta para calcular la performance de cada una de ellas.

Las bases de datos SQL (acrónimo de Structured Query Language), también llamadas bases de datos relacionales, están constituidas por un conjunto de tablas

en las que los datos están clasificados por categorías. Ejemplo de este tipo de base de datos son: *Oracle [3], PostgreSql [4] y MySql [5].*

Por otro lado, las bases de datos NoSQL son no relacionales. Éstas no necesitan un esquema fijo y son fácilmente modulares. El objetivo es recuperar los datos de un mismo lugar sin necesidad de pasar por las relaciones entre tablas. Ejemplo de estas bases de datos son: *MongoDB [6] [7] , Elasticsearch [8] [9] y Neo4J [10].*

Distintos Ministerios, Secretarías y Organizaciones dependientes del Poder Ejecutivo Nacional Argentino han abierto sus datos. Entre ellos, el Ministerio de Justicia y Derechos Humanos [11] publica datos, actualizados mensualmente, referidos a Estadística de trámites de maquinarias, vehículos, embargos, inscripciones, bajas, transferencia, prendas, robos y recuperos de autos, entre otros. Todos estos datos publicados poseen licencia Creative Commons Attribution 4.0 y tienen una frecuencia de actualización mensual.

Por lo tanto, el objetivo planteado en este trabajo es determinar el SGBD adecuada o recomendable para gestionar datos del registro de automotores de la DNRP, para lograrlo se implementan consultas directas a las bases de datos mediante la interfaz de usuario correspondiente.

Este documento está organizado como sigue: la sección 2 describe la metodología involucrada en el desarrollo de este trabajo y los pasos realizados dentro de esta metodología. La sección 3 detalla la visualización de los resultados obtenidos en las consultas junto con la comparativa de los tiempos de ejecución de las consultas en cada uno de los SGBD. Finalmente se detallan las conclusiones y líneas futuras de trabajo.

## 2. Metodología

Para llevar adelante el proceso de comparación de motores de bases de datos y extraer información de forma sistemática se hará uso de la metodología CRISP-DM [15], la cual permite entender el proceso de descubrimiento de conocimiento. CRISP-DM es una metodología creada para trabajar con proyectos de minería de datos, pero de acuerdo a sus fases se adapta al actual proyecto explorando información para la concreción del objetivo propuesto. El ciclo de vida del proyecto de minería de datos, en esta metodología, consiste en seis fases: *Comprensión del negocio, Compresión de los datos, Modelado de datos, Evaluación e Implementación* [17] . Cada una de estas etapas se abordan en las siguientes secciones.

### 2.1. Comprensión del Negocio

En esta fase se debe comprender los objetivos generales y determinar los objetivos técnicos del proyecto. [16] El objetivo del negocio es medir la performance de las bases de datos estructurales versus las no estructurales y, en base a ello, poder recomendar la tecnología más adecuada para trabajar con datos del Registro Nacional del Automotor (DNRPA) [11].

Para concretar el objetivo del negocio, se tiene como objetivo técnico, realizar las consultas acordes y necesarias a ser ejecutadas en las diversas bases de datos mencionadas anteriormente, como así también, visualizar las consultas y determinar la existencia de posibles patrones de comportamiento mencionando las posibles causas.

### 2.2. Herramientas de software y hardware utilizadas

El software utilizado para la ejecución de las consultas es el siguiente:

1. **Sistema Operativo**: Windows 10 - 64 bit y Ubuntu 20.04 Linux
2. **MongoDB Server version 5.0**: Bases de datos NoSQL
3. **Studio 3T version 2022.6.1**: GUI de MongoDB con funciones de consulta visual utilizada para la exportación e importación de colecciones, vistas o consultas.
4. **PostgreSql versión 14.4**: Motor de Base de Datos Relacional.
5. **PgAdmin versión 4.0**: GUI de administración de PostgreSql.
6. **Pentaho Data Integration [14]:** Herramienta de la suite de Pentaho de las que se denomina ETL (Extract – Transform – Load).
7. **Spoon**: Spoon es una Interfaz Gráfica de Usuario (GUI), que permite diseñar transformaciones y trabajos de ETL.
8. **Logstash 7.6:** Herramienta para manejo de grandes volúmenes de archivos entre gestores de bases de datos y sistemas de archivos. Permite transformar datos y enviarlos a diversas bases de datos como MongoDB y ElasticSearch.
9. **ElasticSearch** 7.6 - Base de datos NoSQL.
10. **Power Bi 2.106**: Herramienta de visualización de Datos

Mientras que el hardware utilizado y las características de la máquina sobre la que se realizaron las consultas son:

1. Notebook: HP 15-dw2xxx
2. RAM: 8 GB
3. Disco: SSD 250 GB
4. Procesador: Intel(R) Core(TM) i7-3630QM CPU @ 2.40GHz. 10th generación.

### 2.3. Comprensión de los datos

La recolección de datos inicial corresponden a *"robos y recuperos de autos"* e *"inscripciones iniciales de autos"* de la DNRPA correspondientes al periodo 2018-2022. Cada archivo posee alrededor de 10.000 registros, los cuales representan datos recolectados en un mes, y 25 atributos con información relacionada al tipo de trámite, datos del registro donde se realiza el trámite, datos del vehículo y de su propietario. La cantidad total del dataset es de aproximadamente 2 millones de registros.

Los archivos se dividen en dos grupos bien diferenciados: *Incripciones de autos* [12], el cual posee toda información relacionada a las inscripciones de

vehículos y datos de su primer propietario; *Robos y recuperos de automotores* [13], contiene datos referidos a los trámites de robos y recuperos de automotores.

Los atributos de los archivos de inscripciones tanto como de robos y recuperos de automotores son: *tipo de trámite, fecha del trámite, fecha de la inscripción inicial, código de la seccional del registro, registro seccional descripcion, registro seccional provincia, automotor origen, automotor anio modelo, automotor tipo código,automotor tipo descripción, automotor marca código, automotor marca descripción, automotor modelo código, automotor modelo descripción, automotor uso código, automotor uso descripción, titular tipo persona, titular domicilio localidad, titular domicilio provincia, titular genero, titular anio nacimiento, titular país nacimiento, titular porcentaje, titular domicilio provincia id, titular país nacimiento id.*
Se consideró apropiado realizar la exploración de datos y la verificación de la calidad de datos en las fases posteriores.

### 2.4. Preparación de los datos

En esta fase se procede a preparar los datos para que luego se apliquen las técnicas necesarias para el proyecto. En esta sección también, se describe aspectos de la calidad de los datos. La limpieza de datos se realizó mediante una herramienta de ETL, denominada Spoon [14].

La extracción de datos se realizó desde los archivos *csv* obtenidos de la DNRP y cargados a Spoon, se cargaron 106 archivos que contienen información de Enero de 2018 a Mayo de 2022.

Las transformaciones se realizan para unificar y corregir los nombres de las marcas, tipos de vehículos, tipos de trámites y registros encontrados. Las transformaciones de limpieza más relevantes se detalla a continuación:

1. Los archivos de *robos y recuperos de autos* poseen muchos valores nulos en diversos atributos como automotor_modelo_año. En algunos casos los valores nulos son reemplazados con información encontrada en otra columna y en otros casos son ignorados por no tener valores de referencia.
2. Respecto a los tipos de trámites realizados en las seccionales de la DRNPA se identificaron 9 tipos de inscripciones distintas las cuales fueron unificadas en una única categoría.
3. Los atributos correspondientes a Marcas, Tipo de vehículos, Tipo trámites, Registro del automotor contienen datos ingresados de forma manual, la cual es necesario realizar una limpieza, a fin de unificar estos datos para obtener resultados consistentes en los pasos posteriores. En la Fig. 1 siguiente se muestra la cantidad de registros antes y después de la limpieza de datos.

La construcción de datos se realizó mediante la agregación de las variables mes y año, que representan el mes y año de realización del trámite, para ser tratados como elementos independientes y no como un valor agrupado.

Finalmente, los datos son enviados preprocesados a los motores de base de datos de PostgresSql, MongoDB y ElasticSearch. La elección de los mismos se

| Limpieza | Marcas | Tipo de Vehículo | Tipo Tramite | Registro A |
|---|---|---|---|---|
| Inicial | 1790844 | 1090 | 11 | 868 |
| Limpio | 468 | 47 | 3 | 343 |

**Figura 1.** Limpieza de Marcas, Registro, Tipo Automotor y Tipo trámite

fundamenta en sus características de código abierto, buenas posiciones en rankings de bases de datos más utilizadas, y en particular, las bases de datos no estructuradas por ser escalables y tolerantes a fallos en ambientes de datos masivos.

Las transformaciones se hicieron de acuerdo a las características propias de las base de datos estructuradas o no estructuradas. Para la carga de Postgres fue necesario realizar la disgregación y normalización de los datos. En el caso de MongoDB se generó un archivo único para la generación de la colección. El envío de datos a ElasticSearh se realizó mediante un archivo de texto (csv). Esto último es porque la herramienta Spoon permite enviar datos a diversas bases de datos como mongoDB y PostgreSql incorporando distintos drivers, pero no se encontró el driver adecuado para la base de datos ElasticSearch. Por lo tanto, para la carga de datos a la base de datos de ElasticSearch se utilizó la herramienta Logstash a través del archivo *pipeline_elasctic.conf*.

### 2.5. Modelado, Evaluación e Implementación de los datos

En esta fase es necesario seleccionar la técnica con la que se desarrollará el proyecto. El objetivo del proyecto es determinar la base de datos más adecuada o recomendable para gestionar datos del registro de automotores de la DNRP, para lograrlo se puede implementar consultas directas a la base de datos mediante la interfaz de usuario que cuente cada uno de ellos. En la siguiente sección se detallan como se implementaron las consultas, las respuestas obtenidas y los tiempos de ejecución de las mismas.

- **Consulta 1:** ¿Cuáles son los autos importados de empresas que sufrieron robos y han sido recuperados en el último mes?
  Esta consulta en los 3 SGBD devuelve 16 registros de autos importados que han sido robados y recuperados el último mes, tal como lo muestra la Fig. 2.
- **Consulta 2:** ¿Los compradores de automotores son mayormente hombres, mujeres o personas jurídicas?
  Esta consulta devuelve que la mayoría de los compradores son hombres con 1120667 registros.
- **Consulta 3:** ¿De qué marca y modelo (año) de auto son los más robados?
  Esta consulta devuelve que los autos mas robados son Volkswagen modelo 2011 con 1677 registros, tal como lo muestra la fig. 3.
- **Consulta 4:** ¿Cuáles fueron los meses de menor venta de autos?
  Esta consulta identifica los meses en los que la venta cayó a niveles más bajos en el periodo 2018 a 2022, siendo tales meses abril 2020, marzo 2020 y diciembre 2021.

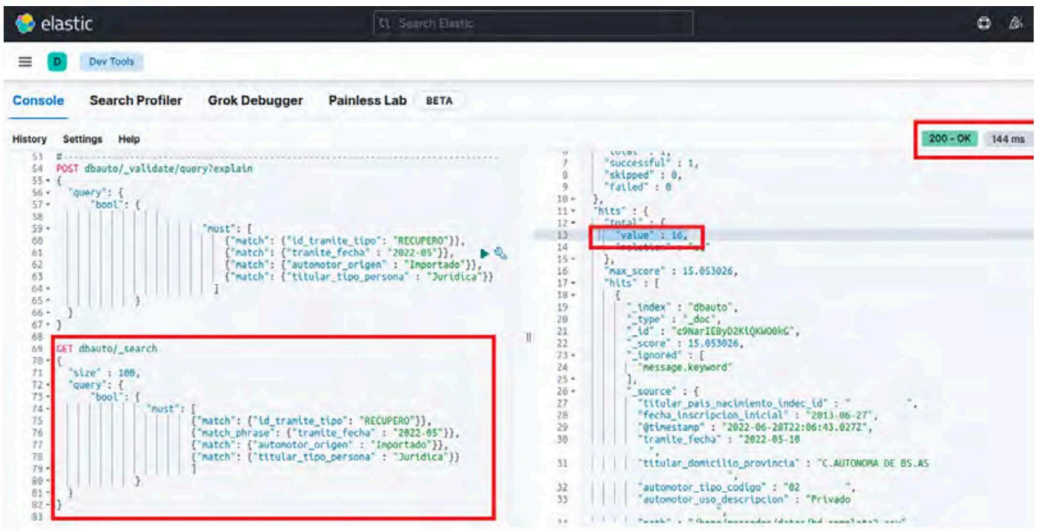


**Figura 2.** Consulta en ElasticSearch.

```
SELECT COUNT(anio_modelo) AS cantidad, anio_modelo, automotor_marca_descripcion
FROM tramites WHERE automotor_marca_descripcion = (
 SELECT marca.automotor_marca_descripcion AS marca_auto
   FROM tramites AS tram INNER JOIN marca_automotor AS marca
     ON tram.id_marca_automotor = marca.id_marca_automotor
   WHERE tram.id_tipo_tramite_sec IN (2, 3)
   GROUP BY marca_auto ORDER BY COUNT(tram.id_marca_automotor) DESC LIMIT 1 )
AND id_tipo_tramite_sec IN (2, 3)
GROUP BY anio_modelo, automotor_marca_descripcion
ORDER BY COUNT(anio_modelo) DESC LIMIT 1
```

**Figura 3.** Consulta 3 en pgAdmin 4.

- **Consulta 5:** ¿Cuál es la marca de autos mas vendidos por provincia? Esta consulta detalla por ejemplo que para la provincia de Bs. As la marca más vendida es Volkswagen, mientras que para Córdoba es Fiat, para Santa Fe es Toyota y así siguiendo para el resto de las provincias como se puede ver en la gráfica 5. La figura 4 muestra su resolución en MongoDB.

```
db.getCollection("inscripciones_robados").aggregate([
        //Etapa 1: filtro por registros de incripciones
        {"$match" : {"id_tramite_tipo" : {"$regex" : "INSC"}}},
        //Etapa 2: agrupo por marca y provincia y cuento la cantidad de registros
        {"$group" : {
                "_id" : {
                    "marca" : "$automotor_marca_descripcion",
                    "prov" : "$registro_seccional_provincia"
                },
                "suma" : {"$sum" : 1.0}
         }},
        //Etapa 3: ordeno de mayor a menor por los totales obtenidos en suma.
        {"$sort" : {"suma" : -1.0}},
        //Etapa 4:
        {"$group" : {"_id" : {"nombre" : "$_id.prov"},
                    "data" : {
                    "$push" : {
                        "provincia" : "$_id.prov",
                        "marca" : "$_id.marca",
                        "ventas" : "$suma"
                    }
         }},
        //Etapa 5: me quedo con el primer registro cuyo valor es el máximo.
        {"$group" : {
                "_id" : {
                    "p" : {"$first" : "$data.provincia"},
                    "m" : {"$first" : "$data.marca"},
                    "v" : {"$first" : "$data.ventas"}
         }}
]
```

**Figura 4.** Consulta 5: Marca más vendida por provincia.

## 3. Exploración y Visualización de los datos

De los datos recolectados se pudieron realizar diferentes gráficas para poder analizar si los resultados obtenidos con las distintas consultas se veían reflejados en dichos gráficos.

Realizadas las consultas, se hizo un análisis visual de los datos obtenidos en las diferentes consultas. La Fig. 5 visualiza los resultados de las consultas 1 a la 3. En principio se muestra los datos de los vehículos importados de las empresas que fueron recuperados en el mes de mayo de 2022. A continuación se visualiza la cantidad de vehículos comprados por género, siendo el 51 % de género masculino, 31 % de género femenino y por último 18 % corresponden a titulares jurídicos (empresas).

La Fig. 6 detalla la cantidad de ventas por mes desde el 2018 hasta la actualidad. En este gráfico se puede ver una baja significativa entre los años 2018 y 2019. Se estima que algunos de los factores de dicha caída fue el incremento de los precios y la escasa financiación, provocando la caída abrupta de ventas, sobre todo en Diciembre de 2021. Por otro lado, se puede ver que otros 2 picos mínimos de ventas ocurridos en Abril y Marzo del 2020 ocasionados por la pandemia Covid19.

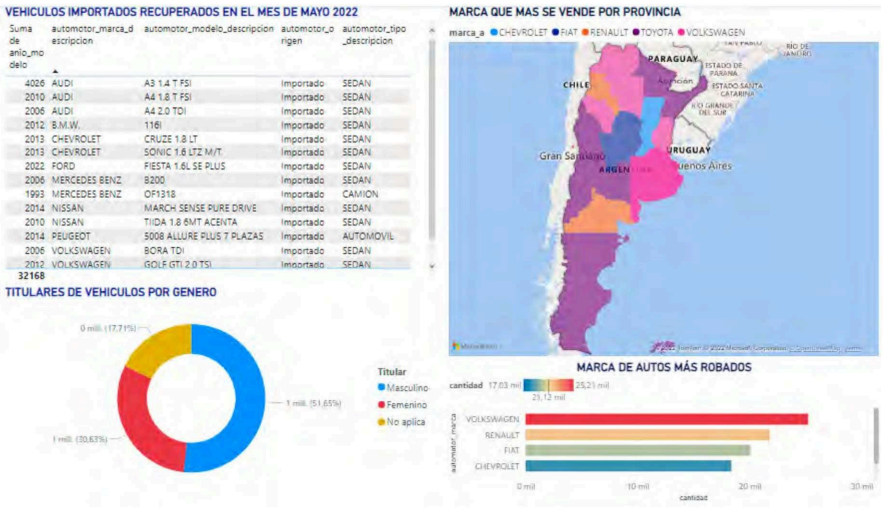


**Figura 5.** Vehículos recuperados, Titulares por Género y Marcas por provincia.


**Figura 6.** Histórico de Ventas de automotores.

### 3.1. Comparativa de Tiempos de Ejecución

En el cuadro 3.1 se muestran los tiempos de ejecución promedio de las consultas en las distintas base de datos, representadas gráficamente en la Fig. 7. Dichos valores son el resultado de 10 corridas por consulta.

| - | Consulta 1 | Consulta 2 | Consulta 3 | Consulta 4 | Consulta 5 |
|---|---|---|---|---|---|
| PosgreSql | 1.014 | 0.876 | 1.162 | 0.953 | 0.755 |
| MongoDB | 4.788 | 7.988 | 2.750 | 7.033 | 7.486 |
| ElasticSearch | 0.144 | 0.667 | 0.243 | 0.425 | 0.732 |

**Cuadro 1.** Tiempos Promedio de Ejecución en segundos

En el cuadro 2 se muestra la varianza entre las distintas mediciones para cada consulta, siendo las más significativas los tiempos de la consulta 1 y 2 en MondoDB. Esto se debe al hecho de tener un tiempo extra de preparación propio del motor de base de datos al iniciar las ejecuciones.

| - | Consulta 1 | Consulta 2 | Consulta 3 | Consulta 4 | Consulta 5 |
|---|---|---|---|---|---|
| PosgreSql | 0.212 | 0.143 | 0.256 | 0.122 | 0.053 |
| MongoDB | 10.03 | 20.96 | 0.011 | 0.055 | 0.063 |
| ElasticSearch | 0 | 0.001 | 0.002 | 0.001 | 0.003 |

**Cuadro 2.** Varianza de los Tiempos de Ejecución en segundos

Como se pueden observar en las gráficas anteriores los tiempos de ejecución de ElasticSearch y sus varianzas son considerablemente mejores.

## 4. Conclusiones

Este proyecto se encuentra enmarcado en la categoría de almacenamiento de grandes volúmenes de datos, que de acuerdo a la frecuencia de actualización de datos del Registro de la Propiedad del Automotor, irá creciendo de manera continua. Por lo tanto, decidir dónde y cómo almacenar esta información es de vital importancia.

Las base de datos SQL y NoSQL son dos tecnologías que tienen la misma finalidad: almacenar datos y ofrecer las herramientas para leer y manipular esos datos. Elegir la base de datos más adecuada es una tarea muy importante porque será la base de trabajo de todas las profesiones en el campo de la ciencia de datos. Sin embargo, esta elección no es fácil y la respuesta no siempre es evidente.

Las tareas realizadas al conjunto de datos fueron: preparación y limpieza de datos, la cual demandó la mayor parte del tiempo del proyecto, luego se

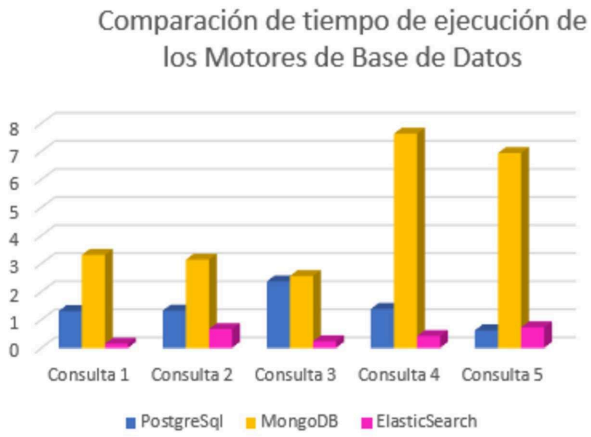


**Figura 7.** Resultados de consultas en distintas Base de Datos.

modelaron las consultas en PostgreSql, MongoDB y ElasticSearch, y finalmente, se obtuvieron los resultados y tiempos de respuesta.

De los tiempos de ejecución obtenidos se puede concluir que ElasticSearch tiene un mejor tiempo de respuesta con respecto a las otras dos bases de datos. Esto se debe a la naturaleza distribuida y su capacidad de manejo de datos estructurados y no estructurados. Por otra parte, se puede ver que PostgreSql tiene mejores tiempos de respuesta que MongoDB, dado que este último es más eficiente cuando se trata de datos no estructurados.

Por lo tanto, para el conjunto de datos trabajado, se recomienda el uso de Elasticsearch por permitir consultas mediante textos completos, autocompletado y realización de correcciones ortográficas, sin necesidad de realizar la tarea de preprocesado tan exhaustiva.

Como trabajo futuro, se propone optimizar las consultas para evaluar los nuevos tiempos de respuesta y verificar si ocurre alguna mejora en los mismos, como así también ejecutarlas en un hardware más potente.

**Agradecimientos.** Al profesor Mg. Javier Bazzocco, docente del Centro de Investigación LIFIA.

## Referencias

1. Learning PostgreSQL. Salahaldin Juba, Achim Vannahme, Andrey Volkov. 2015. ISBN: 9781783989188
2. Beynon-Davies, P.: Sistema de Base de Datos pp.33-53. Editorial Reverté (2014). España.
3. Oracle, `https://www.oracle.com/ar/database/`. Consultado en Junio de 2022.
4. PostgreSql, `https://www.postgresql.org/`. Consultado en Junio de 2022.
5. MySql, `https://www.mysql.com/`. Consultado en Junio de 2022.

6. MongoDB, https://www.mongodb.com/es. Consultado en Junio de 2022.
7. Data Modeling for MongoDB: Building Well-Designed and Supportable MongoDB Databases. Steve Hoberman. 2014
8. ELK - ElasticSearch, https://www.elastic.co/es/what-is/elasticsearch
9. Martos, C., Uso y Ventajas de Elasticsearch en Bases de Datos No Relacionales (2019). Universidad Autónoma de Madrid. Escuela Politécnica Superior. España.
10. Neo4J, https://neo4j.com/. Consultado en Junio de 2022.
11. Dirección Nacional de Registros Nacionales de la Propiedad Automotor y Créditos Prendarios. Datos públicos generados, guardados y publicados por organismos de gobierno de la República Argentina, Ministerio de Justicia y Derechos Humanos. https://datos.gob.ar/dataset?tags=dnrpa Último acceso 30 de Junio 2022.
12. Dirección Nacional de Registros Nacionales de la Propiedad Automotor y Créditos Prendarios (https://datos.gob.ar/dataset/justicia-transferencias-autos). Último acceso 30 de Junio 2022.
13. Dirección Nacional de Registros Nacionales de la Propiedad Automotor y Créditos Prendarios. (https://datos.gob.ar/dataset/justicia-robos-recuperos-autos). Último acceso 30 de Junio 2022.
14. Pentaho Data Integration. https://sourceforge.net/projects/pentaho/. Consultado en Junio de 2022.
15. Chapman, P., Clinton, J., Kerber, R., Khabaza, T. y otros: CRISP-DM V 1.0. pp.10 – 12. CRISP-DM consortium: NCR Systems Engineering Copenhagen, DaimlerChrysler AG, SPSS Inc and OHRA Verzekeringen (2000)
16. Galan Cortina, V., Castro Galan, E.: . Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el Entorno Universitario. Universidad Carlos III de Madrid. Escuela Politécnica Superior Ingeniería en Informática (2015)
17. Moine, J. M., Haedo, A. S., Gordillo, S.: . Estudio comparativo de metodologías para minería de datos (2011). XIII Workshop de Investigadores en Ciencias de la Computación. RedUNCI
18. Db-engines ranking of wide column stores. https://db-engines.com/en/ranking. Consultado en Agosto de 2022.
19. Duran-Cazar, Jhonatan, Tandazo-Gaona, Eduardo y cia. Rendimiento de base de datos columnares". Universidad Politécnica Salasiana. Revista de Ciencia y Tecnología. Ecuador. 2019.