|  |
| --- |
| detalle de manos con tijeras, marcadores, trabajandoINFORME CALIDAD DE DATOS |

|  |  |  |
| --- | --- | --- |
| JURISPRUDENCIA | **Noviembre de 2023** |  |
| Correo electrónico: biblioteca\_areatecnica@jusformosa.gob.ar | Dirección de Biblioteca e Informática Jurisprudencial – Área Técnica |  |

CONTENIDO

[INFORME CALIDAD DE DATOS 3](#_Toc150759840)

[OBJETIVO 3](#_Toc150759841)

[ALCANCE 3](#_Toc150759842)

[ESQUEMA DE DATOS 3](#_Toc150759843)

[TABLA DE FALLOS 1](#_Toc150759844)

[TABLA DE JURISPRUDENCIA 2](#_Toc150759845)

[TABLA DE TRIBUNALES 3](#_Toc150759846)

[ANÁLISIS DE DATOS 1](#_Toc150759847)

[VALORES NULOS 1](#_Toc150759848)

[VALORES DUPLICADOS 1](#_Toc150759849)

[VALORES VACIOS 1](#_Toc150759850)

[SUMARIOS SIN RELACIÓN CON FALLOS 3](#_Toc150759851)

[EXPLORACIÓN DE DATOS 1](#_Toc150759852)

[ANALISIS DE PERFORMANCE DE LA BASE DE DATOS 1](#_Toc150759853)

[RECOMENDACIONES TÉCNICAS 2](#_Toc150759854)

[ENTIDAD DE FALLOS 2](#_Toc150759855)

[ENTIDAD DE JURIPRUDENCIA 3](#_Toc150759856)

[ENTIDAD DE FALLOS 3](#_Toc150759857)

[CONSIDERACIONES GENERALES 3](#_Toc150759858)

# INFORME CALIDAD DE DATOS

## OBJETIVO

Presentar un análisis de calidad de datos basados en los procesos de ETL (Extracción, Transformación y Carga de Datos), EDA (Análisis Exploratorio de Datos) y testeo de los tiempos de respuesta de las consultas. Para lo cual se tomó como objeto de estudio la base de datos de jurisprudencia, almacenada en una base de datos relacional: Mysql versión 8.0.21

## ALCANCE

Este informe es presentado al Departamento de Jurisprudencia para su análisis y oportunamente corrección de los datos si así lo considerasen necesario. Los datos son analizados al mes de noviembre de 2023.

## ESQUEMA DE DATOS

Antes de ofrecer un informe de la calidad de los datos analizados es necesario describir su estructura de datos para dar al lector un panorama más preciso del contexto analizado.

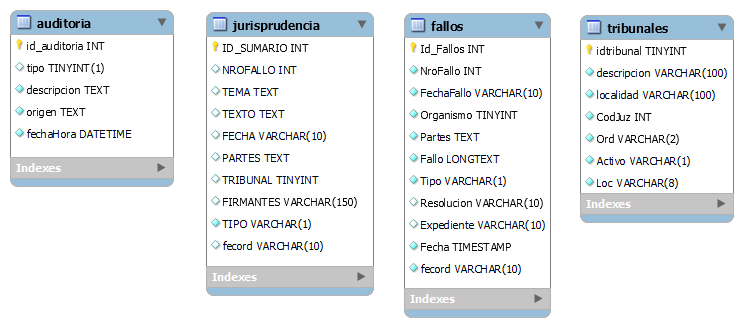


Figura . Diagrama Entidad Relación de Jurisprudencia

En las descripciones de las tablas se adicionan algunos comentarios acerca de la calidad de datos y más adelante en recomendaciones técnicas se ofrecen buenas prácticas a implementar sobre los datos.

### TABLA DE FALLOS

Tabla . Entidad de fallos

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nombre campo | Tipo de campo | Longitud del campo | ¿Es dato obligatorio? | Observaciones |
| Id\_Fallos | int | 11 | si | Clave primaria. |
| NroFallo | int | 11 | si | Número del fallo. |
| FechaFallo | varchar | 10 | si | Fecha del fallo. Tipo de dato no adecuado. |
| Organismo | tinyint | 2 | si | Clave referencial a la tabla de tribunales. No se usan claves foráneas. |
| Partes | text |  | si | Carátula del fallo. Se observan campos vacíos. |
| Fallo | longtext |  | si | Texto plano del fallo. |
| Tipo | varchar | 1 | si, pero tiene registros vacíos cuando el tipo es 0 sentencia. | Según comentario de tabla se nutre de valores 0 sentencia y 1 auto interlocutorio. |
| Resolucion | varchar | 10 | no | Campos vacíos. |
| Expediente | varchar | 10 | no | Campos vacíos. |
| Fecha | timestamp |  | si | Fecha de sistema. Fecha de alta del registro. |
| fecord | varchar | 10 | si | Esta fecha se corresponde y es la misma que la fecha del fallo. Se observa campos duplicados entre FechaFallo y fecord. |

### TABLA DE JURISPRUDENCIA

Tabla . Entidad de jurisprudencia

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nombre campo | Tipo de campo | Longitud del campo | ¿Es dato obligatorio? | Observaciones |
| ID\_SUMARIO | int | 11 | si | Clave primaria. |
| NROFALLO | int | 11 | si | Número del fallo. Dato que ya se encuentra en la tabla de fallos. |
| TEMA | text |  | si | Descriptores del sumario. |
| TEXTO | text |  | si | Texto del sumario. |
| FECHA | varchar | 10 | si | Fecha del sumario. Tipo de dato no adecuado. |
| PARTES | text |  | si | Carátula del fallo. Dato que ya se encuentra en la tabla de fallos. No se tiene campos separados por actor, demandado y sobre o tema. |
| TRIBUNAL | tinyint |  | si | Clave referencial a la tabla de tribunales. No se usa claves foráneas solo referencial. Sin embargo, algunos registros están con el valor 0. |
| FIRMANTES | varchar | 150 | si | Cadena con los nombres de los firmantes. |
| TIPO | varchar | 1 | si, pero tiene registros vacíos cuando el tipo es 0 sentencia. | Según comentario de tabla se nutre de valores 0 sentencia y 1 auto interlocutorio. |
| fecord | varchar | 10 | si | Esta fecha es la misma que el campo Fecha. Campos duplicados. |

### TABLA DE TRIBUNALES

Tabla . Tabla de tribunales

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nombre campo | Tipo de campo | Longitud del campo | ¿Es dato obligatorio? | Observaciones |
| idtribunal | tinyint |  | si | Clave primaria. |
| descripcion | varchar | 100 | si | Nombre del tribunal. |
| localidad | varchar | 100 | si | Nombre de la localidad donde Se encuentra el tribunal. Hay registros vacios. |
| CodJuez | int |  | si | Código de Juzgado. |
| Ord | varchar | 2 | si | Orden para ordenar los tribunales. |
| Activo | varchar | 1 | si | Indica si el registro está activo o no. |
| Loc | varchar | 8 | si | Campo redundante, almacena la localidad en formato abreviado FSA. |

## ANÁLISIS DE DATOS

No presentan valores duplicados a nivel de registros, pero sí se observan valores nulos en las columnas de **Resolucion** y **Expediente** en los fallos, que implica un 11,64% para el total de fallos consultados al día de la fecha 18.108 registros. En estas mismas columnas se observan registros con valores vacíos siendo el 88,35%, por lo que podemos sostener que estas columnas en casi el 100% (18.107 registros) tienen sus campos con datos vacíos o nulos.

Siguiendo con la entidad de los fallos el campo **Tipo** determina si el registro es sentencia o AI, en este caso los registros en vacío son Sentencias. Por último, se debe revisar aquellos registros que no poseen **Partes** y **Fallo** cargados en sus registros. -

En la revisión de esos datos con **Partes** en valor vacío se observan los registros de la *Figura 2. Consulta fallos con columna Partes en valores* . En relación a la columna Fallos se visualiza la *Figura 3. Consulta fallos con columna Fallo en valores vacíos*. Véase las recomendaciones técnicas para estos casos.

Observándose la entidad de jurisprudencia (también denominada sumarios) se observa que **fecord** tiene tanto nulos como vacíos, en las recomendaciones técnicas se sugiere su eliminación pues Esta fecha es la misma que el campo Fecha para jurisprudencia.

En el campo **TIPO** sucede lo mismo que en la entidad de fallos, este determina si el sumario es de sentencia o AI, en este caso los registros en vacío son Sentencias. Si se deben revisar aquellos registros con **TEMA** y **PARTES** en vacío *(Figura 4. Consulta de sumarios con columna TEMA vacío* y *Figura 5. Consulta de sumarios con columna PARTES en vacío*). -

Por último, de los tribunales podemos observar 4 registros con localidad en valor vacío lo que se recomienda completar este dato.

Se analizaron además cuantos registros de sumarios se relacionan a los fallos según el número de fallo y el tribunal, resultando la consulta con 197 registros a revisar esta inconsistencia.

Este análisis se realizó a fecha 06 de noviembre de 2023.-

### VALORES NULOS

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Entidad | Número de registros | Número de columnas | Nulos totales (en todas las columnas) | Número de columnas con valores nulos | Cantidad máxima de nulos encontrados sobre columna | Porcentaje de cantidad máxima de nulos encontrados sobre columna | Columnas con nulos |
| fallos | 18108 | 11 | 4216 | 2 | 2108 | 11,64% | Resolucion (2108), Expediente (2108) |
| jurisprudencia | 6683 | 10 | 659 | 1 | 659 | 9,86% | fecord(659) |
| tribunales | 37 | 7 | 0 | 0 | 0 | 0,00% | - |

### VALORES DUPLICADOS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Entidad | Total de registros duplicados | Porcentaje de duplicados | Columnas con duplicados | Proporción de duplicados |
| fallos | 0 | 0% | - | 0 |
| jurisprudencia | 0 | 0% | - | 0 |
| tribunales | 0 | 0% | - | 0 |

### VALORES VACIOS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Entidad | Vacíos totales (en todas las columnas) | Número de columnas con valores vacíos | Cantidad máxima de vacíos encontrados sobre columna | Porcentaje de cantidad máxima de vacíos encontrados sobre columna | Columnas con valores vacíos |
| fallos | 49569 | 5 | 15999 | 88,35% | Partes(12), Fallo(14), Tipo (17545), Resolucion(15999), Expediente(15999) |
| jurisprudencia | 7704 | 4 | 6415 | 95,98% | TEMA(1), PARTES(9), TIPO(6415), fecord(1279) |
| tribunales | 4 | 1 | 4 | 10,81% | localidad(4) |

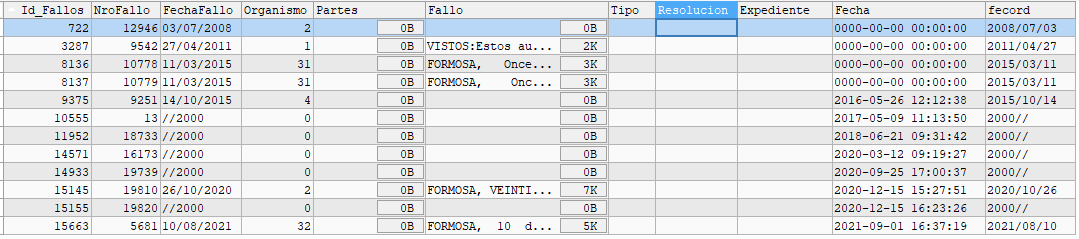


Figura . Consulta fallos con columna Partes en valores vacíos

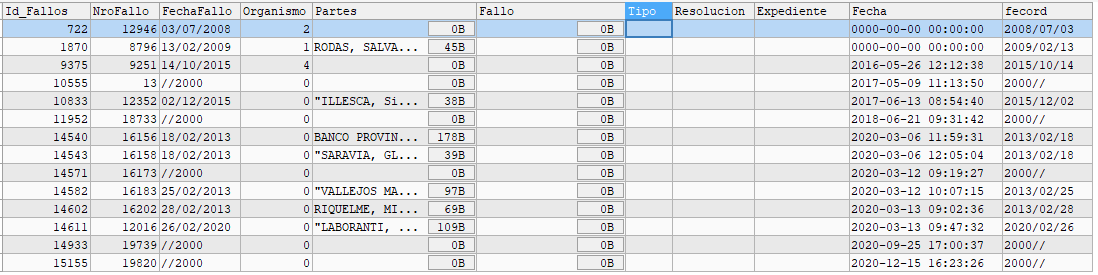


Figura . Consulta fallos con columna Fallo en valores vacíos



Figura . Consulta de sumarios con columna TEMA vacío

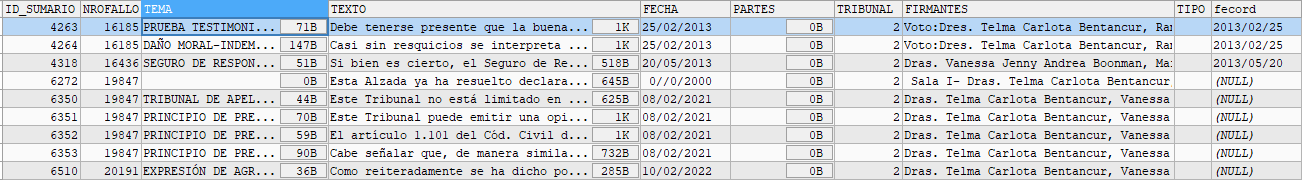


Figura . Consulta de sumarios con columna PARTES en vacío



Figura . Consulta de sumarios con TRIBUNAL en valor 0

### SUMARIOS SIN RELACIÓN CON FALLOS

Se ofrece un fragmento de los 3491 registros de sumario que se consultaron y no tienen relación con sus fallos.

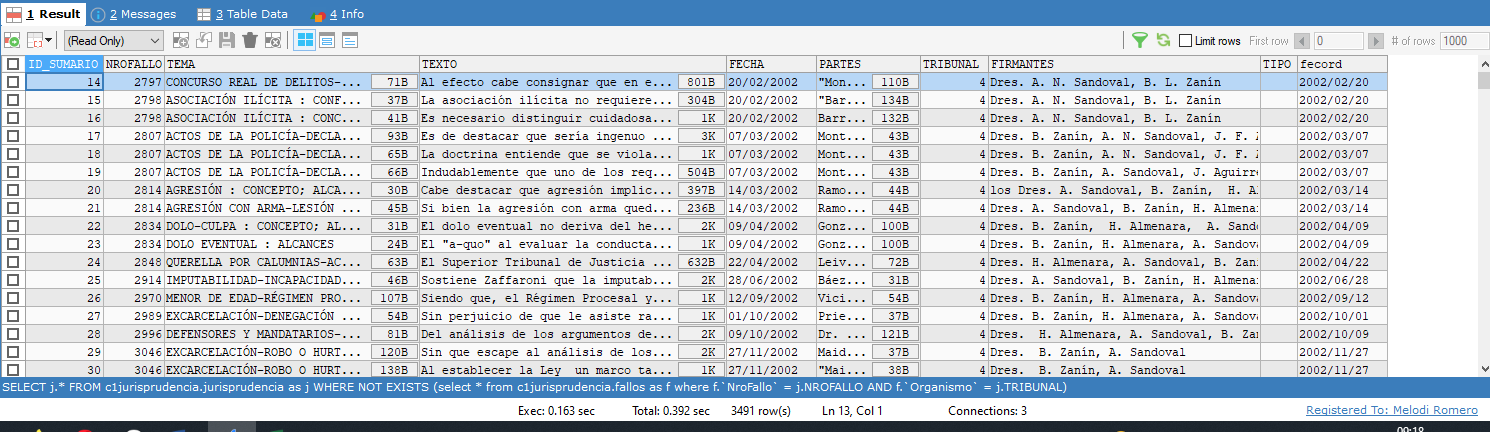


Figura . Sumarios sin relación a sus fallos

## EXPLORACIÓN DE DATOS

De la exploración de los datos se puede aportar que el año 2018 se registra la mayor cantidad de fallos entre los años 1994 - 2023, los días lunes y jueves son las fechas de fallos más predominantes (coincidiendo con los días de pre despacho de los juzgados).

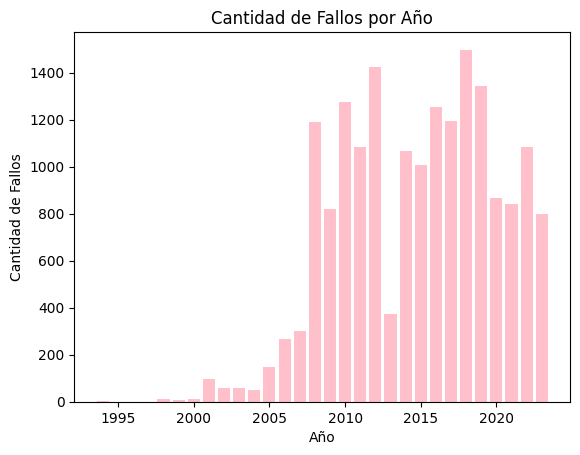


Figura . Cantidad de fallos por año

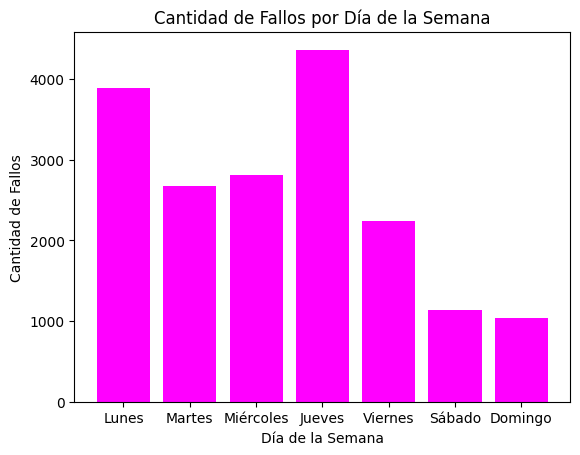


Figura . Cantidad de fallos por día de la semana

El organismo que más fallos (7296) registra es el id 2 corresponde a la Excma. Cámara Civil y Comercial.

En datos erróneos o involuntarios se hallaron 14 registros con organismo en valor 0 dentro de la entidad fallos.

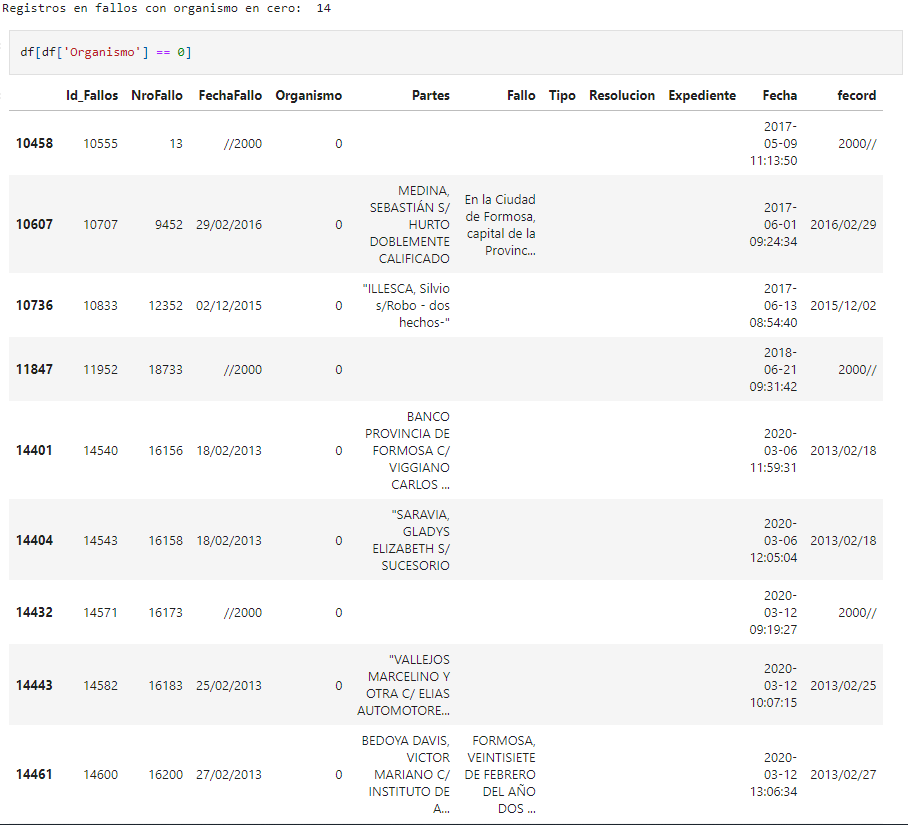


Figura . Registros en fallos con organismo en valor 0

También se analizó aquellos registros con el campo FechaFallo con fechas inválidas como ser “//2000” o un valor de fecha no válido como “31/09/2014”, “30/14/2015”, “31/02/2020”. Estos registros se dejan para su posterior corrección con el Departamento de Jurisprudencia.

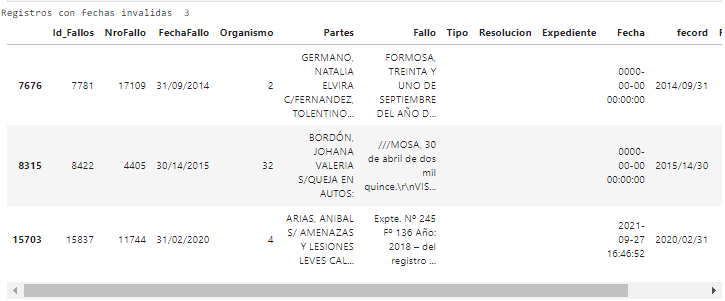


Figura . Registros en fallos con fechas inválidas



Figura . Registros en fallos con fecha "//2000"

Por último, se registran al día de la fecha “14 de noviembre de 2023”, 17574 sentencias y 563 auto interlocutorios.

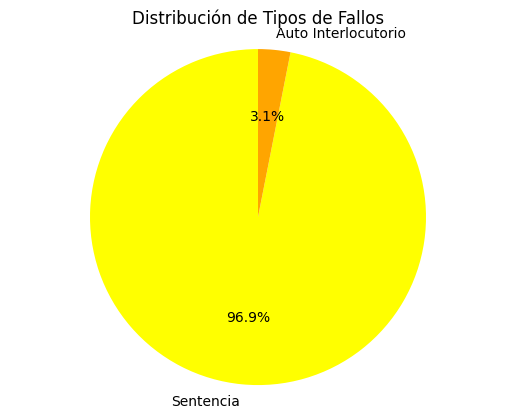


Figura . Distribución de fallos por tipo.

De la exploración de los datos en los sumarios se puede aportar que el año 1994 se registra la mayor cantidad de sumarios entre los años 1993 - 2023, los días lunes y jueves son las fechas de fallos más predominantes (coincidiendo con los días de pre despacho de los juzgados, igual que en el caso de los fallos).

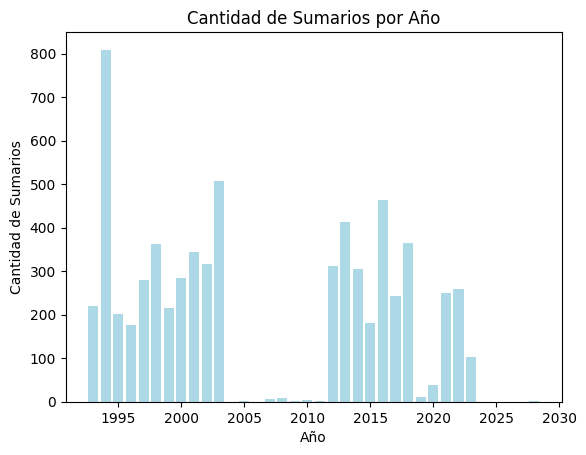


Figura . Cantidad de sumarios por año

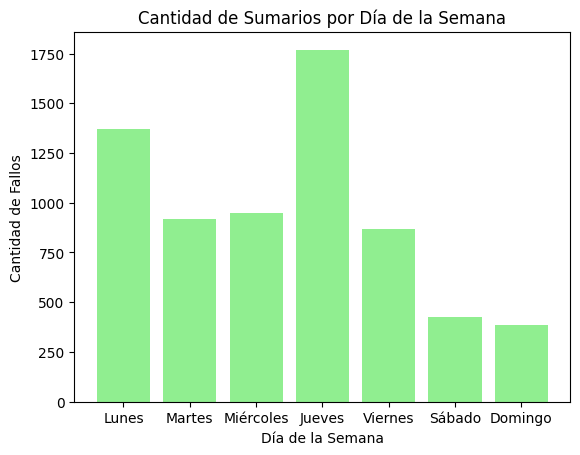


Figura . Cantidad de sumarios por día de la semana

Y en consecuencia al haber más sentencias que auto interlocutorios, los sumarios también son del tipo sentencia en su 96% (6417 registros) frente al 4% (268). Como dato atípico se hallaron sumarios del año 1993 cuando fallos hay registrados desde el 1994 y se supone que los sumarios están vinculados a los fallos.

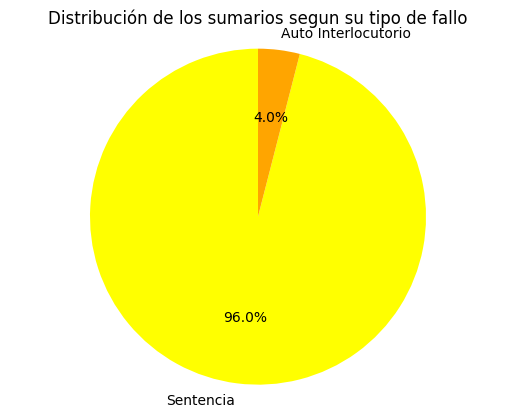


Figura . Distribución de sumarios por tipo de fallo

Como datos con carga errónea en los campos se observan registros de sumarios con tribunal en valor cero, con fechas erróneas, y años 1993 (220 registros) y 2028 (1 registro).

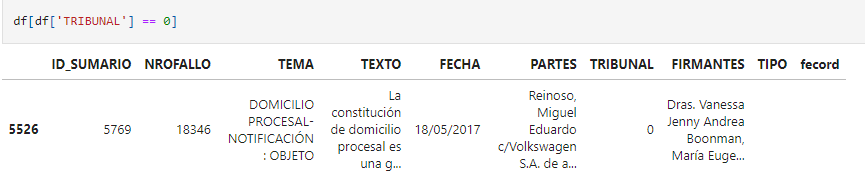


Figura . Sumarios con campo tribunal en cero

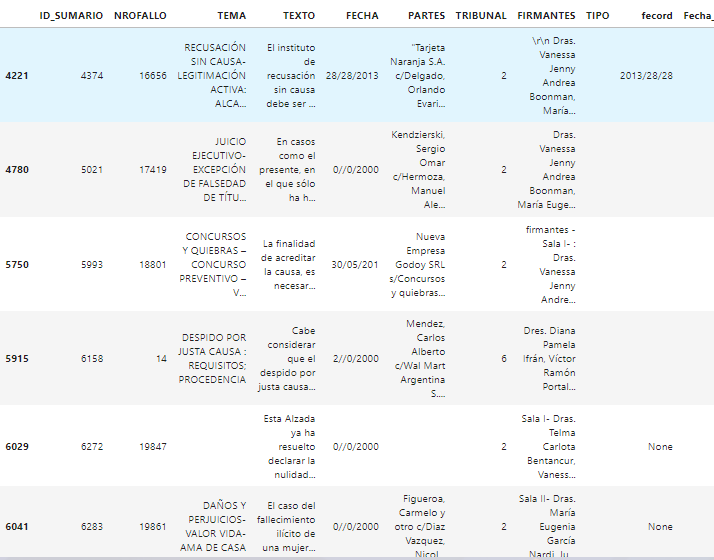


Figura . Sumarios con fecha errónea.

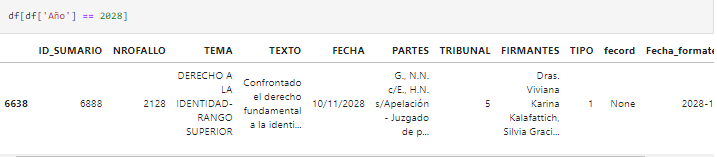


Figura . Sumario con registro de año 2028



Figura . Extracto de sumarios con registro año 1993

Se analizaron registros outliers o atípicos pero no se consideran ser incluidos como un problema en la calidad de datos.

## ANÁLISIS DE PERFORMANCE DE LA BASE DE DATOS

A continuación, se analizan el tiempo de respuesta de consultas realizadas desde la intranet del ámbito de estudio, sobre la base de datos analizando el tiempo de respuesta de las mismas.

Tabla 4. Performance de las consultas a la base de datos

|  |  |  |  |
| --- | --- | --- | --- |
| Consulta | Tiempo de Respuesta (Segundos) | Número de registros retornados | Cantidad de tablas consultadas |
| Lectura de todos los fallos | 13,874 | 18187 | 1 (tabla fallos) |
| Lectura de todos los sumarios | 0,623 | 6691 | 1 (tabla jurisprudencia) |
| Lectura de sumarios con temática: “violencia doméstica” | 0,083 | 91 | 1 (tabla jurisprudencia) |
| Lectura de sumarios con temática: “régimen jurídico” | 0.260 | 1420 | 1 (tabla jurisprudencia) |
| Lectura de jurisprudencia con firmante: Ariel Coll | 0,209 | 1492 | 1 (tabla jurisprudencia) |
| Lectura de todos los fallos del año 2019 | 1,443 | 1344 | 1 (tabla fallos) |
| Lectura de los fallos del tribunal Excma. Cámara Civil y Comercial, que más fallos tiene registrado. | 0,791 | 7297 | 2 (tabla fallos y tabla tribunales) |

En la labor cotidiana se recomienda que las consultas no tarden más de 0,05 segundos, lo que significa que las consultas ejemplificadas antes, exceden el tiempo de respuesta recomendado, aclarando que las mismas se realizaron directamente sobre la base de datos no sobre el sistema de gestión que consume la base, a esto habría que adicionar el tiempo de respuesta de la red y el sistema de gestión, pero para evidenciar ineficiente performance de las consultas es suficiente lo antes detallado en la tabla…

### ÍNDICES POR TABLA

Se observan índices sobre las tablas, los cuales son:

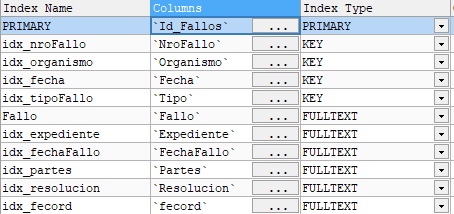


Figura . Índices de la tabla de fallos

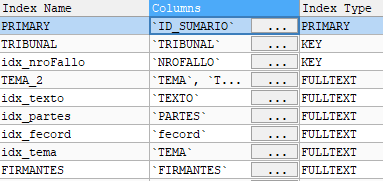


Figura . Índices de la tabla de jurisprudencia

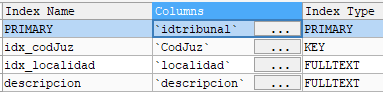


Figura . Índices de la tabla de tribunales

### CONEXIÓN DE RED

La base y la aplicación actual que consulta dicha base están en servidores distintos, por lo cual la velocidad y la congestión de la red también pueden afectar el tiempo de respuesta.

### DISEÑO DE LA BASE DE DATOS

A lo largo del documento se puede observar que, si bien la base de datos es sencilla, no tiene una normalización adecuada y la selección de tipos de datos no son los correctos, esto en consecuencia afecta el rendimiento general del sistema. Véase las recomendaciones técnicas más adelante descriptas.

### VOLUMEN DE DATOS

La cantidad de datos en la base de datos puede afectar directamente el tiempo de respuesta de una consulta. Cuanto más grande sea la base de datos, es probable que las consultas requieran más tiempo.

# RECOMENDACIONES TÉCNICAS

## ENTIDAD DE FALLOS

**Cambiar tipo de dato en FechaFallo.**

Cambiar a tipo date para FechaFallo, ya que el tipo de dato varchar(10) no es el adecuado.

**Corregir las fechas inválidas en FechaFallo.**

Corregir las fechas de los fallos que son inválidas: “//2000”, “31/09/2014”, “30/14/2015”, “31/02/2020”.

**Eliminar la columna Resolucion**

Dado que solo un registro cuenta con esos campos completos, además Resolucion almacena el número y año del fallo, datos que se hallan en las columnas de NroFallo y FechaFallo.

**Considerar la columna Expediente como campo no nulo**

Se entiende que esta columna almacena el número y año del expediente donde tiene lugar el fallo, considerar obligatorio la carga de este campo para los futuros registros.

**Actualizar la columna Tipo valores nulos por 0**

Es mejor asignar un valor a este campo antes que quede como dato vacío para marcar sus registros Tipo Sentencia.

**Registros con los campos Partes y Fallo vacíos**

Dada la Figura 2. Consulta fallos con columna Partes en valores y Figura 3. Consulta fallos con columna Fallo en valores vacíos, se recomienda eliminar los registros con **FechaFallo** en **“//2000”**, también eliminar aquellos con los campos **Partes** y **Fallos** en vacío, ya que se consideran registros erróneos o que no aportan valor.

Los demás registros que tienen un valor ya sea en Partes o en Fallo se pueden mantener completando el campo vacío. El campo Partes almacena la caratula del expediente y Fallo el texto de la sentencia o AI, por lo cual es fundamental completar estos registros cuyos campos presentan valores vacíos.

**Valores duplicados en FechaFallo y fecord**

Esta fecha es la misma que la fecha del fallo. Se observa campos duplicados entre FechaFallo y fecord, se puede eliminar **fecord**.

**Adicionar fecha de última modificación**

Es necesario agregar un campo de última modificación del registro.

**Auditar cambios en los fallos**

Auditar cambios en el registro, ya sea por triggers, procedimientos, la forma que resulte más oportuna.

## ENTIDAD DE JURIPRUDENCIA

**Cambiar tipo de dato en FECHA.**

Cambiar a tipo date para FECHA, ya que el tipo de dato varchar(10) no es el adecuado.

**Valores duplicados en FECHA y fecord**

Se observa campos duplicados entre **FECHA** y fecord, se puede eliminar **fecord**.

**Corregir las fechas inválidas en FECHA.**

Corregir las fechas de los sumarios que son inválidas.

**Actualizar la columna TIPO valores nulos por 0**

Es mejor asignar un valor a este campo antes que quede como dato vacío para marcar sus registros Tipo Sentencia.

**Registros con los campos TEMA y PARTES vacíos**

Es importante marcar estos campos como obligatorios o en su efecto eliminar el campo PARTES ya que se obtiene de la relación de fallos y dejar el TEMA como columna obligatoria. –

**Tipificar el campo FIRMANTES**

Se puede normalizar la información de los firmantes de un sumario.

**Optimizar la relación de fallos con sumarios**

Actualmente el sumario va vinculado al fallo por el numero de fallo y el tribunal (campos NROFALLO y TRIBUNAL), es mejor reemplazar esta relación por la clave referencial Id\_Fallos.

## ENTIDAD DE TRIBUNALES

**Eliminar campo Loc.**

Campo redundante, almacena la localidad en formato abreviado FSA, pero en el campo localidad se almacena el mismo dato por lo que se recomienda eliminarlo.

**Completar campo localidad.**

Completar los registros que tienen la localidad en vacío.

**Normalizar campo localidad.**

Se puede crear la entidad de localidad para su relación con los tribunales.

## CONSIDERACIONES GENERALES

Para todas las entidades se recomienda:

**Adicionar fecha de última modificación**

Es necesario agregar un campo de última modificación del registro.

**Auditar cambios en los registros**

Auditar cambios en el registro, ya sea por triggers, procedimientos, la forma que resulte más oportuna.

**Unificar la nomenclatura de las descripciones de entidades y columnas**

Si es posible, unificar criterios de denominación de tablas y columnas en la base de datos.