Beneficios del uso de Big Data en la Justicia. Análisis de su aplicación sobre el software INVESTIGA

Autores: Ariel Podestá*, Sergio Viera+, Sabrina Lamperti*, Diego Caparra+, Pedro Asis+, Ana Di'Iorio*

**Facultad de Ingeniería - UFASTA - (ufasta.edu.ar) {apodesta, slamperti, diana}*
+Facultad de Ingeniería - UTN FRD - (frd.utn.edu.ar) {sviera, diego.caparra, pasis}

Resumen

El presente trabajo introduce los notables beneficios que el sistema judicial puede obtener del uso de las técnicas de Big Data vinculados al análisis forense de datos operando sobre la copiosa cantidad de información que la sociedad genera día tras día, a través de los distintos medios de comunicación y sistemas informáticos que utiliza. Se presenta un caso testigo desarrollado por dos Universidades Argentinas para el Ministerio Público de la Provincia de Buenos Aires.

1. Introducción

La Agencia Española de Protección de Datos Personales definió al Big Data como a "las gigantescas cantidades de datos digitalizados que son controlados por las empresas, autoridades públicas y otras grandes organizaciones que poseen la tecnología para realizar un análisis extenso de los mismos basado en el uso de algoritmos." [1]

Big Data es un concepto que cobra fuerza progresivamente con el paso del tiempo. Se sustenta sobre dos pilares fundamentales. La información y la capacidad de cómputo.

La información, como primer pilar, es un activo cuyo crecimiento aumenta en forma exponencial al transcurrir los años. Nuevos sistemas de comunicación, diferentes formas de capturar momentos, recursos de entretenimiento, redes sociales de distintas finalidades y sistemas administrativos al servicio de la sociedad son algunos de los ejemplos de recursos que se van integrando al mundo tecnológico, que todo ciudadano tarde o temprano comienza a utilizar con cotidianeidad.

Esta creciente tendencia al uso de las tecnologías de información digital genera este incremento masivo de información, que debe ser almacenada en algún medio y que puede ser aprovechada para realizar análisis de una infinidad de índoles. Es por ello que el protagonismo de Big Data va cobrando relevancia a medida que evoluciona la sociedad junto a la tecnología, dándole el soporte para operar sobre esta significativa cantidad de datos.

La capacidad de cómputo, como segundo pilar del Big Data, es una característica de la tecnología actual, que según la "Ley de Moore" [2] se duplica cada 18 meses. Este aumento, también exponencial, genera el escenario correcto para poder procesar tamaña cantidad de datos generados por el ser humano [3]. Sin este poder de cálculo sería inviable el análisis tanta información.

Así, es como hoy en día el análisis de Big Data tiene todo el sustento necesario para tener lugar y ofrecer toda su potencia.

1.1. Fuentes de Información

El origen de los datos corresponde a una incontable cantidad de recursos. Técnicamente cualquier medio de almacenamiento, sistema o red de comunicaciones puede proveer datos útiles para un análisis de tipo Big Data. Todo recurso es válido. Toda red social (Facebook, Twitter, Instagram, etc.), sistema de gestión, medio de comunicación (telefonía fija, celular, de radio, SMS, WhatsApp, etc.) y datos integrados en cualquier archivo son algunos de los ejemplos de recursos de datos viables para integrar a un sistema de análisis de Big Data. Con lo cual, ningún registro de información queda exento.

1.2. Síntesis de Beneficios

Big Data ofrece infinidad de beneficios que no siempre son comúnmente conocidos. Este tipo de análisis puede dar el sustento para una correcta toma de decisiones a futuro; puede servir para reconstruir sucesos (aplicable a la justicia) y encontrar posibles causas de los mismos; o incluso puede brindar una visión muy precisa del comportamiento de individuos, sociedades u otras entidades. Es la percepción directa de lo que ocurre con el mayor nivel de síntesis y precisión que se puede tener, si los datos son compilados adecuadamente.

Por otra parte, las tecnologías de Big Data ofrecen la posibilidad de procesar cantidades de información inmanejables por otros sistemas para reducirlas a un volumen representativo que sí pueda ser tratado por los mismos. Este es el caso de Big Data – INVESTIGA, el sistema que se presenta en este trabajo.

2. Problemática – Big Data en la Justicia

El uso de Big Data representa un desafío cada vez mayor para quienes buscan incorporarla a sus mecanismos de tratamientos de datos personales. En este sentido, quienes tienen el rol de investigación de los delitos, se enfrentan no sólo a los hechos, rastros, y redes criminales, sino también al crecimiento exponencial de las fuentes de información.

Es innegable que la posibilidad de unir distintas informaciones provenientes de fuentes de datos abiertas, tales como publicaciones en redes sociales, imágenes digitales, vídeos y fotos, registros de transacciones comerciales y bancarias, señales GPS de los móviles, registros de servidores web, imágenes de satélites, contenido de las páginas web, constituyen un potencial enorme que puede coadyuvar a los investigadores judiciales.

Desde esta perspectiva, se busca que los operadores judiciales o de organismos de investigación, optimicen el tratamiento de grandes volúmenes de datos reduciendo el tiempo para descubrir puntos de interés que contribuyan a una oportuna toma de decisiones. Dentro de esta información, proveniente de fuentes heterogéneas, se espera que los operadores judiciales sean capaces de detectar los patrones delictivos, reconstruyan secuencias temporales, descubran a los partícipes de hechos y determinen sus roles.

2.1. Necesidades puntuales

En el marco de las investigaciones judiciales y el tratamiento de datos, se hacen evidentes ciertas necesidades que pueden ser cubiertas por procesos informáticos automatizados.

- **Almacén único de datos**: Durante toda investigación, el operador judicial recopila información cuyo volumen puede ser significativo. La posibilidad de cargarla masivamente con celeridad, en un único lugar, en forma organizada y disponible para ser rápidamente analizada es sin duda una clara necesidad.

- **Integración de información en diferentes formatos**: Ciertamente la información puede provenir de distintos orígenes en formatos disímiles. El ejemplo más emblemático son los registros de llamadas provistos por las distintas operadoras de telefonía. La ausencia de estándares y protocolos para la entrega de la información hace que el formato de estos archivos sea heterogéneo.

- **Normalización de datos**: La expresión de ciertos tipos de datos puede no ser siempre la misma. Por ejemplo, el número telefónico podría o no contener el código de área. La forma de expresar estos datos depende de quién los entregue. Ergo, es responsabilidad de quien los procesa el interpretar correctamente la información a fin de generar efectivamente las interrelaciones existentes.

- **Trazabilidad completa**: En toda investigación es fundamental poder determinar el origen exacto de cada dato procesado. Se debe poder conocer de dónde proviene la información que se observa en cualquier momento. El objetivo es garantizar que todo dato analizado provenga de un recurso oficial.

- **Confidencialidad de los datos**: En el ámbito de la justicia es menester mantener una estricta discreción con los datos manipulados. En toda investigación es posible que se sospeche de la conducta de determinados individuos hasta que se determina su participación en el caso o no. Pero el hecho de que estén siendo investigados no significa que hayan cometido un delito. Sin embargo la percepción desde la sociedad habitualmente es que seguramente algo tienen que ver, perjudicándolos en su vida cotidiana. Esta condena social a destiempo y posiblemente injustificada es lo que se debe evitar, manteniendo en secreto los datos de la investigación.

- **Interrelación de datos**: Una de las tareas de todo investigador judicial es descubrir vínculos entre las entidades involucradas en un caso (por ejemplo personas en contacto con redes delictivas). Con toda la información surgida de una investigación es posible realizar cruzamientos de datos, cuyo fruto podría resultar en el descubrimiento de relaciones inesperadas o difícilmente visibles si los datos se procesaran manualmente.

3. Solución propuesta - Caso de Aplicación INVESTIGA + Big Data

El proyecto Big Data – INVESTIGA, utilizando tecnologías orientadas al Big Data (como lo es “Apache Solr” [4]) da el soporte técnico necesario al proyecto INVESTIGA (que se tratará a continuación) para satisfacer estas necesidades.

3.1. Proyecto base: INVESTIGA

INVESTIGA es un software de soporte a las investigaciones judiciales desarrollado por el Laboratorio de Investigación y Desarrollo en Informática Forense (InFo-Lab) [5]. Su principal utilidad es brindar una visión gráfica de interrelaciones que pueden existir entre datos provenientes de diferentes fuentes de información. Su propósito fundamental es proveer al investigador judicial una percepción clara del caso y la interacción entre las entidades involucradas.

En la Figura 1 se presenta una captura del sistema donde se observa la representación gráfica de un entrecruzamiento de datos.

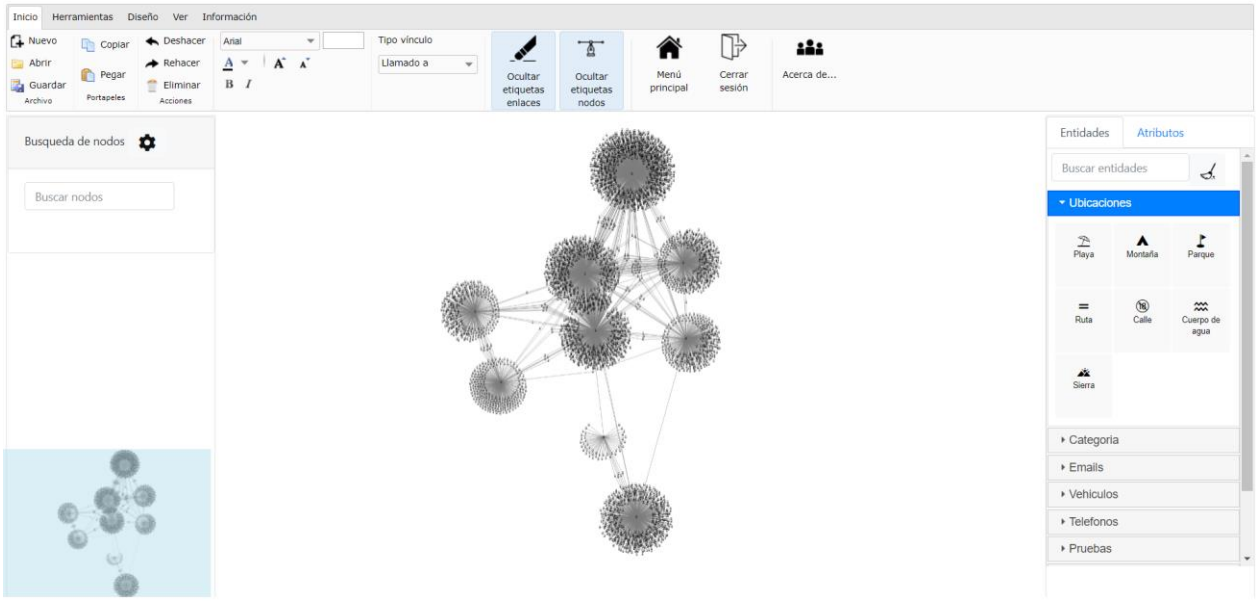

Figura 1: Captura de pantalla de sistema INVESTIGA.

Cada investigación es un escenario diferente para INVESTIGA. Los datos recopilados difieren, así como también las modalidades de investigación y el uso que se le da. Esto se vio reflejado, al momento de realizar las primeras pruebas piloto, donde surgieron necesidades específicas que requerían aplicar una solución tecnológica apropiada al procesamiento de grandes datos.

Toda plataforma de análisis e interrelación de datos, como lo es INVESTIGA, naturalmente tendrá un límite en la cantidad de entidades que puede procesar en un determinado momento. Sin importar la evolución en la potencia de procesamiento, siempre existirá un umbral tras el cual el sistema deja de responder en el tiempo esperado. Como INVESTIGA no es la excepción, para trabajar con un número ilimitado de datos, se requirió de otra solución que le dé el soporte necesario: Big Data – INVESTIGA.

3.2. Proyecto Big Data – INVESTIGA

El proyecto Big Data - INVESTIGA aborda esta problemática mencionada creando un módulo de pre-procesamiento de la información a graficar, que permite filtrarla y resumirla según el criterio del usuario. De este modo, el usuario puede desentenderse de la limitante del tamaño de los datos recolectados.

Las herramientas que permiten hacer esta reducción en el volumen de información son los “filtros” que Big Data – INVESTIGA ofrece. La aplicación de estos filtros permite descartar masivamente datos que pueden no son ser relevantes para la investigación.

Por ejemplo, algunos de los casos típicos investigados son los que comúnmente se llaman “piratas del asfalto”, dada la modalidad en que asaltan vehículos de transporte en viaje. En estos casos habitualmente se analizan las llamadas que ocurren en las cercanías de determinadas antenas de telefonía celular. Los listados de comunicaciones suelen ser de magnitudes considerables, dificultando así su tratamiento. Por ejemplo, un gráfico representativo de uno de ellos podría ser el presentado en la Figura 2.

Este tipo de gráficos, no solo es un problema para el usuario, dado el exceso de información contenida, sino que también lo es para el sistema. El operar con tal cantidad de entidades en un gráfico demanda un poder de procesamiento enorme que afecta notablemente el rendimiento de la interfaz de usuario, lo que reduce en utilidad final del sistema percibida por el usuario final.

Para este caso filtrar la información irrelevante suele ser la alternativa indicada. Un filtro útil podría ser: “filtrar aquellos números telefónicos cuyo código de área no corresponda a Mar del Plata”. De ese modo, solo serán contempladas las llamadas que involucren a números telefónicos con código de área “0223”. En la Figura 3 se presenta el gráfico con dicho filtro aplicado.

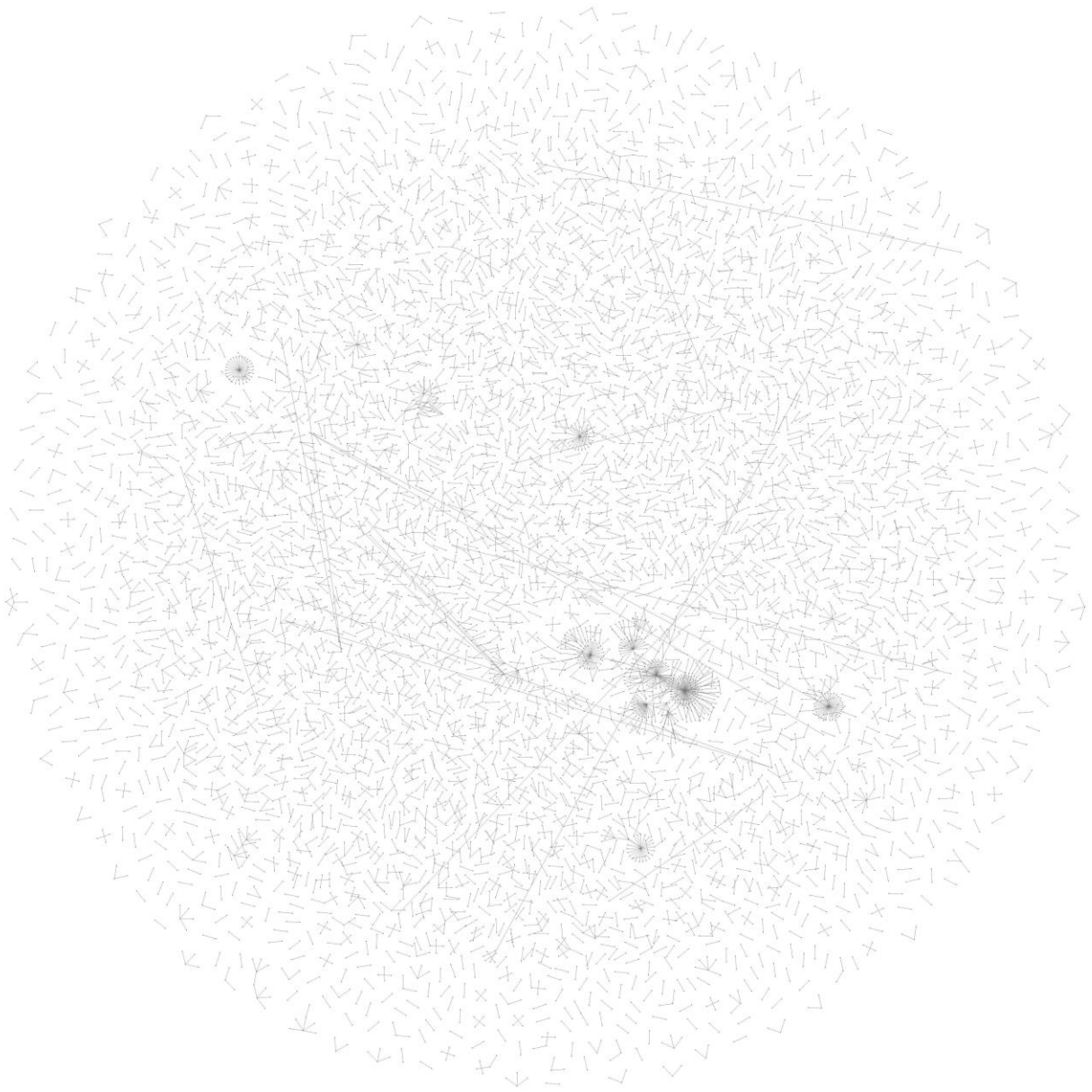

Figura 2: Representación gráfica de listado de llamadas sin procesar.

Si bien es un filtro simple, el verdadero potencial radica en la concatenación de múltiples filtros, en base a criterios o indicios apropiados de la investigación. Entonces, bien podría utilizarse el filtro anterior, en combinación con "filtrar toda llamada producida fuera del horario de 6:00 AM y 8:00 AM". Con lo cual, para el ejemplo solo se tratarían llamadas de teléfonos celulares no pertenecientes a Mar del Plata producidas entre las 6:00 AM y las 8:00 AM cualquier día dentro del registro obtenido. En la Figura 4 se observa el resultado obtenido.

Luego de la aplicación de esta concatenación de filtros se percibe una visión más limpia de la información, evidenciándose ciertos grupos con actividad elevada en el periodo en cuestión. Este grupo se observa ubicado en el centro de la Figura 4.

Este tipo de concatenación de filtros puede realizarse en forma indefinida, aislando cada vez más los datos significativos para la causa.

Esta es la diferencia fundamental entre contar con un módulo de procesamiento de grandes cantidades de datos y no: es el poder operar con la información recolectada o

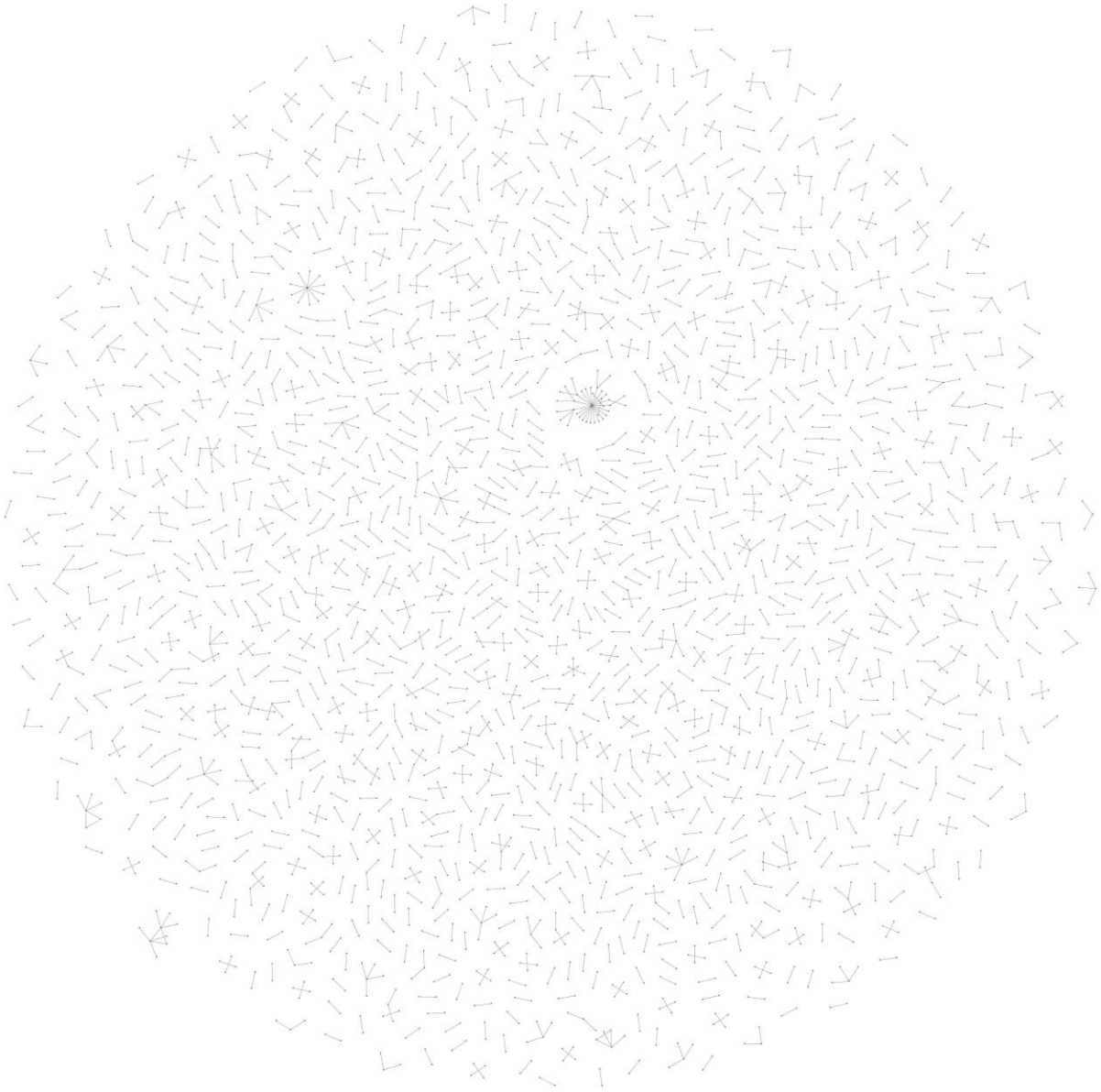

Figura 3: Listado de llamadas filtrado por código de área distinto de "0223".

no hacerlo. En el ejemplo el tamaño de los datos no es tan extenso, a fin de presentar el caso con claridad, pero bien podría haberse aplicado la misma mecánica con lotes de información decenas o cientos de veces más grande que el presentado, y aun así poder sido procesados.

4. Resultados de las Pruebas

Big Data – INVESTIGA no solo aporta la posibilidad de filtrar información, sino que permite abordar las necesidades mencionadas en el punto “2.2 Necesidades puntuales” y brindar los siguientes beneficios:

- **Enriquecimiento progresivo del conocimiento**: A medida que diversos casos van siendo tratados con un sistema de este tipo, la información acumulada aumenta. Esta información, puede tener correlación con nuevos casos que se aborden. La diferencia entre comenzar el estudio de un caso sin información previa y hacerlo con un conjunto de casos predecesores vinculados, puede ser

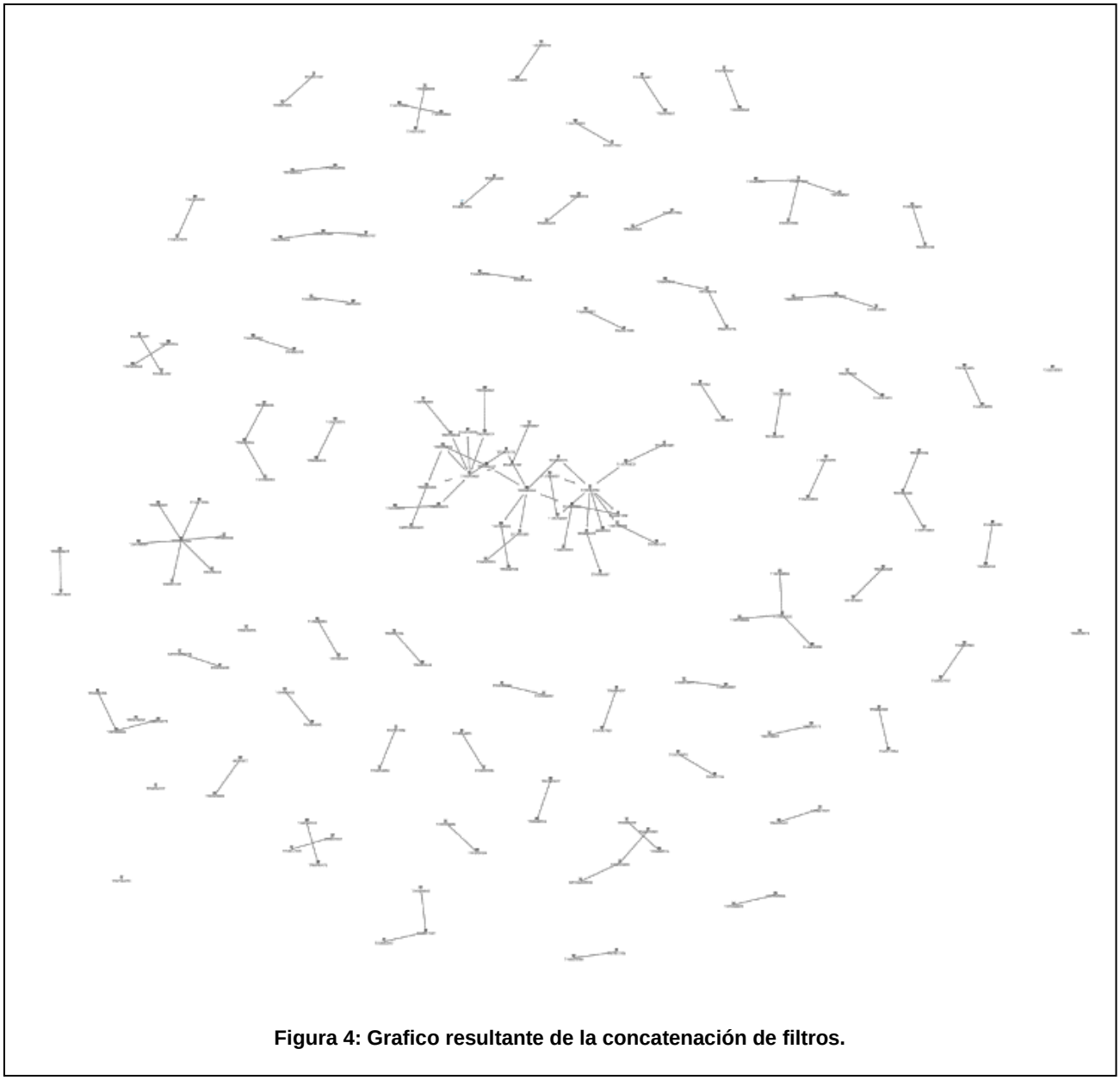

Figura 4: Grafico resultante de la concatenación de filtros.

determinante. El pre-conocimiento de conductas ya demostradas en individuos involucrados puede realmente agilizar la investigación actual.

- **Reportes estadísticos**: Pudiendo obtener reportes estadísticos sobre toda la información almacenada, en tiempos insignificantes para el proceso de investigación, podría poner en evidencia ciertos patrones aún no avistados. Por ejemplo: "obtener números telefónicos que han recibido más de 100 llamadas por día". Un dato que solo podría surgir por el entrecruzamiento de información de diferentes casos, podría ser de suma utilidad para nuevas investigaciones o aquellas en curso.

- **Trazabilidad de la información**: Los sistemas de Big Data ponen especial atención al origen y consistencia de la información visualizada. De todo dato deben conocerse sus etapas de procesamiento transitadas. Esto garantiza la legitimidad de la información analizada. En el ámbito de la justicia esta capacidad una necesidad básica debido a la sensibilidad los datos y las posibles consecuencias de contemplar información errónea en una investigación judicial.

- **Velocidad de reacción**: Cuando grandes cantidades de datos deben ser analizadas, los tiempos de investigación pueden extenderse a un punto inaceptable, volviendo impracticable su procesamiento. Un sistema que responda siempre en tiempo y forma, sin importar la magnitud de la

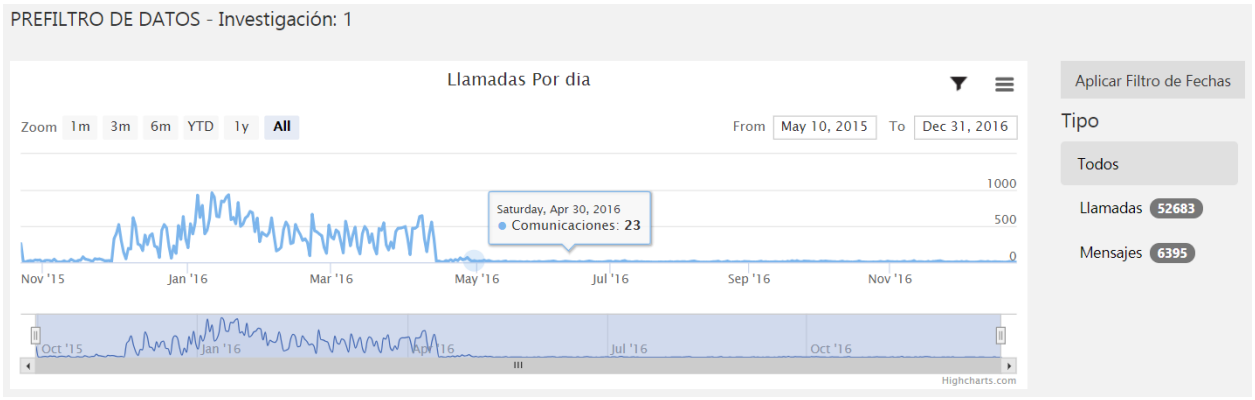

Figura 5: Análisis gráfico de frecuencia de aparición, filtrado a criterio del usuario.

información ingresada puede brindar la celeridad que la justicia necesita para evitar nuevos delitos.

- **Análisis gráfico masivo**: Uno de los beneficios más notables de este tipo de sistemas es la representación gráfica de los datos, en diferentes formas de visualización. Según la herramienta utilizada dichos gráficos podrían permitir el filtrado de datos en forma dinámica. Este es el caso de Big Data – INVESTIGA.

Esta percepción directa del comportamiento de la información es una ventaja significativa para los tiempos de investigación que sintetiza aspectos relevantes del caso abstrayendo al usuario del excesivo detalle innecesario al momento de hacer un análisis macro.

En esta ocasión se presentan dos herramientas que Big Data – INVESTIGA contempla.

- En la figura 5 se observa el componente de "Análisis grafico de frecuencia". Allí el usuario puede observar la frecuencia de aparición de datos, filtrados por un rango de fechas y tipo de actividad (mensajes y llamadas para este caso).
- En la figura 6 se muestra la herramienta de "Tag Clouds" que expone los datos más protagónicos dimensionando su texto según su frecuencia de aparición.

Estas herramientas indudablemente son un recurso útil para los investigadores.

5. Posibilidades a futuro

- **Filtros semánticos**. El análisis de la información a través de los medios convencionales implica conocimiento detallado de los datos para poder aplicar filtros matemáticos, obteniendo de esta manera un subconjunto de datos, a los cuales se aplica un nuevo nivel de filtrado previo análisis. Esta metodología, si bien es efectiva, puede generar una complejidad muy elevada desde el punto de vista del usuario. Los filtros semánticos, transforman la lógica matemática en expresiones claras para los usuarios, disminuyendo la abstracción matemática que implican los filtros convencionales.

- **Integración con más sistemas**. Las tecnologías utilizadas permiten separar de manera coherente el análisis referido al manejo de gran cantidad de datos (Big Data) del análisis y uso posterior de los resultados obtenidos. De esta manera se puede pensar en el proceso de Big Data como un sistema independiente, pero con la facilidad de integrarlo a otros sistemas, ya que la esencia de Big Data es obtener una muestra representativa de una gran cantidad de datos para ayudar al usuario a tomar una decisión acertada.

- **Incorporación de ontologías**: Se puede extender y mejorar la búsqueda semántica si incorporamos ontologías para limitar la complejidad y para organizar la información. La ontología puede entonces ser aplicada para resolver problemas. Incorporando una descripción (como una especificación formal de un programa) de los conceptos y relaciones que pueden formalmente existir para realizar búsquedas dentro del dominio de los datos investigados, que a la vez permita persistir y utilizar la información de búsquedas ya realizadas. Esta definición es consistente con el uso de ontología como un conjunto de definiciones conceptuales, pero más generales.

Figura 6: Tag Clouds. Cada nube muestra los datos más protagónicos dimensionando su texto según su frecuencia de aparición.

6. Conclusiones

El uso del Big Data en las investigaciones judiciales crece en importancia cada día en forma ininterrumpida. Es necesario seguir incorporando nuevas técnicas y herramientas que saquen provecho de toda la información recabada en las diversas fuentes de datos accesibles por la justicia, atendiendo por supuesto a las consideraciones propias y pertinentes en cuanto a protección de datos y fines de dicha investigación. Big Data bien aplicado, puede ser desequilibrante en términos de agilidad de la justicia. Big Data - INVESTIGA es un sistema informático desarrollado según las necesidades de los investigadores judiciales de la provincia de Buenos Aires, y próximo a implantarse como prueba piloto.

Big Data es una temática relativamente moderna. Debido a esto existe una demanda de tecnología aún insatisfecha para dar soporte a necesidades puntuales. Por consiguiente es esencial desarrollar proyectos que se ajusten a estos requerimientos específicos. Big Data – INVESTIGA es un ejemplo emblemático de esto mismo.

7. Referencias

[1] Memoria AEPD 2014 - ISSN: 2254-691X – Link: https://www.prevent.es/Files/HtmlCustom/Documentos/Memoria_AEPD_2014.pdf (accedido 10/08/2018)

[2] Moore, Gordon E. "Cramming more components onto integrated circuits". 1965. Link: https://drive.google.com/file/d/0By83v5TWkGjvQkpBcXJKT1I1TTA/view?usp=sharing (accedido el 10/08/2018).

[3] John R. Mashey. "Big Data ... and the Next Wave of InfraStress". Presentación de charla invitada, Usenix. 1998. Link: http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf (accedido 10/08/2018).

[4] Apache Solr. Link: http://lucene.apache.org/solr/ (accedido 10/08/2018)

[5] Constanzo Bruno, Lamperti Sabrina, Lasia Sebastián, Podestá Ariel, Cistoldi Pablo, Haydée Di Iorio Ana. “El Análisis Automático de Datos, su Aporte a la Investigación Criminal”. 2017. Link: http://redi.ufasta.edu.ar:8080/xmlui/bitstream/handle/123456789/1596/JAIIO%20SID%202017-2556-Investigaci%C3%B3n-CR.pdf?sequence=1 (accedido el 10/08/2018)