Automatización y computación distribuida para laboratorios de informática forense

Leopoldo Sebastián Gómez[*,†], Hernán Horacio Herrera[*], Federico Martín Uribe[*]

[*]Poder Judicial del Neuquén
[†]Universidad Nacional de Río Negro
{sebastian.gomez, hernanhoracio.herrera, federico.uribe}@jusneuquen.gov.ar

Resumen. El crecimiento del volumen digital en las pericias informáticas es uno de los factores críticos que lleva al colapso de los laboratorios de informática forense y al aumento de las listas de espera de casos en trámite. Urge entonces comenzar a transitar un nuevo camino en lo que refiere a aspectos metodológicos, junto con la aplicación de nuevas técnicas y herramientas de informática forense apoyadas en recursos tecnológicos que provean una razonable capacidad de cómputo y de almacenamiento masivo de información digital. Este trabajo presenta la implementación y experimentación sobre una infraestructura de cómputo distribuido llevada adelante en el Gabinete de Pericias Informáticas del Poder Judicial del Neuquén, integrando diversas aplicaciones de informática forense con el objeto de automatizar y agilizar aquellas actividades operativas que demandan tiempos elevados durante el proceso forense digital. A la luz de los cambios en los plazos procesales para la investigación penal que operan luego de la última reforma al Código Procesal Penal de la Provincia del Neuquén se procura maximizar la disponibilidad de recursos computacionales para el procesamiento de evidencia digital.

Palabras clave: laboratorios de informática forense, computación distribuida, automatización

1 Introducción

La primera década del siglo XXI ha sido descripta como la Era Dorada en el desarrollo de la informática forense. Durante estos años los delincuentes dejaban indicios digitales que eran fácilmente detectables para los peritos. La evidencia digital a localizar era trivial y el software utilizado por los sospechosos era ampliamente conocido. Pero lo más importante era que el material digital a ser analizado y la cantidad de tareas operativas que practicaba el perito sobre los elementos probatorios eran razonablemente manejables en el tiempo.

Desde mediados de los 90 comenzó a desarrollarse software específicamente diseñado para informática forense como Encase[1] y FTK[2]. En los años sucesivos se pensó que se había logrado un estado relativamente aceptable de madurez en la adopción de este tipo de entornos integrados para el trabajo pericial. Sin embargo, la evolución propia de Internet y los avances de la tecnología provocaron cambios sustanciales.

El planteo clásico de trabajo en que el caso contenía un único sospechoso, una computadora, un disco rígido, y un único perito para hacer el análisis y la elaboración del informe pericial ha quedado en el recuerdo. Hoy en día los requerimientos periciales son acompañados con múltiples elementos probatorios, eventualmente conteniendo evidencia digital de más de un sospechoso y un elevado volumen de información digital para ser analizado.

Como resultado de este cambio, el desarrollo de tareas operativas para el análisis de la evidencia digital se ha vuelto complejo y se extiende en el tiempo. Este fenómeno conlleva a un crecimiento excesivo de la lista de espera de pericias en trámite e impacta negativamente en la investigación que deben llevar a cabo los organismos jurisdiccionales. Por el aumento de elementos probatorios en casos que demandan la ejecución de pericias informáticas para el esclarecimiento de los hechos, es posiblemente en el ámbito de la Justicia penal donde se visualiza el problema con mayor claridad.

Para otras disciplinas, y con más énfasis en el fuero civil o laboral, tradicionalmente ha resultado suficiente con la designación de un profesional de la materia para la ejecución de un peritaje, ya sea un perito oficial o de oficio, quien realiza las actividades periciales con el respectivo control de parte. El esquema de trabajo que conlleva la asignación de una pericia a un único perito para que sea quien efectúe todas las actividades operativas sobre el material probatorio ha mostrado sus debilidades en materia de informática forense ya que no es escalable. La relación clásica de trabajo uno-a-uno entre el perito y la pericia atenta contra la ejecución de múltiples actividades concurrentes en el tiempo sobre elementos probatorios de uno o varios casos, que pueden ser realizadas por diferentes miembros de un equipo profesional de laboratorio, optimizando el tiempo y posibilitando además un mayor control de calidad interno de las operaciones efectuadas.

Las pericias informáticas deben ser consideradas como labores de media o alta complejidad. Un abordaje de labores periciales informáticas a gran escala demanda una infraestructura tecnológica y de recursos humanos especializados más propia de

[1] Encase es una suite de productos para investigaciones digitales desarrollada por la empresa Guidance Software entre los que se encuentra Encase Forensic, una aplicación destinada a profesionales forenses que deben realizar una recolección eficaz de datos válidos a efectos legales e investigaciones mediante el uso de un proceso defendible y repetible. Contiene herramientas para varias áreas del proceso forense digital, como adquisición, análisis y presentación de informes.

[2] FTK, acrónimo de Forensic Tool Kit, es un software para informática forense desarrollado por la empresa Access Data que cuenta con capacidades de procesamiento de evidencia digital y permite correlacionar grandes volúmenes de información digital obtenida de distintas fuentes, bien sean computadoras, discos rígidos, dispositivos móviles, capturas de tráfico de red o almacenamiento en Internet.

un laboratorio, en el que actúa un equipo profesional, que de la simple actuación de un perito informático en forma independiente. El trabajo riguroso que se realiza en un laboratorio de informática forense, con estricto apego a los protocolos de actuación y procedimientos operativos estandarizados, asegura la cadena de custodia, la inalterabilidad de la evidencia digital, y promueve un proceso forense digital repetible y reproducible que permite desacoplar la actuación concurrente en el tiempo entre la histórica dupla conformada por el perito de oficio y el perito de control. Finalizadas las labores periciales sobre la evidencia digital en respuesta a los puntos de pericia requeridos, se remite el material probatorio junto a los resultados obtenidos en el laboratorio de informática forense para cualquier otra intervención que sea oportuna. Este dictamen y los respectivos elementos de prueba pueden ser escrutados y controlados por las partes para emitir sus observaciones con los resguardos propios del manejo de prueba digital.

La alta complejidad y numerosa cantidad de casos que requieren de especialistas en informática forense está demandando la creación de laboratorios o gabinetes dedicados exclusivamente y en forma permanente a la actividad pericial informática. En los últimos años ha comenzado a abandonarse el paradigma el trabajo aislado del perito informático, evolucionando hacia un esquema de trabajo segmentado por roles en el que se distribuyen tareas operativas conforme la experiencia y experticia de cada profesional del laboratorio. Alineado a esta nueva forma de organización de la actividad pericial informática, la adopción de un modelo de trabajo escalable requiere la planificación, implementación y el mantenimiento de una infraestructura tecnológica que sea capaz de soportar sin pérdida de performance la ejecución concurrente de múltiples tareas operativas y el procesamiento intensivo de grandes volúmenes de información digital.

La extinción de la acción por su insubsistencia en función del vencimiento del plazo para investigar se funda en primer término en expresos mandatos legislativos extraídos de los Códigos Procesales. El artículo 158 del Código Procesal Penal de la Provincia de Neuquén establece plazos perentorios para la investigación penal, contándose con cuatro meses en la etapa preparatoria. Transcurrido este plazo se produce la extinción de la acción penal y debe dictarse el sobreseimiento del imputado. El fiscal o el querellante podrán solicitar una prórroga de la etapa preparatoria cuando la pluralidad de víctimas o imputados, o las dificultades de la investigación hagan insuficiente este plazo. Las investigaciones que conllevan el análisis de evidencia digital usualmente tienen demoras en función de la lista de espera de casos en trámite y de la complejidad misma de cada pericia informática. Tal como sucede en mayor escala a nivel global, el Gabinete de Pericias Informáticas del Poder Judicial del Neuquén se enfrenta al constante desafío provocado por el aumento del volumen digital y la cantidad de elementos de prueba que pueden estar asociados a una investigación penal. En este escenario el Fiscal o el querellante pueden pedir una prórroga y el Juez de Garantías fija el plazo, el que no excederá de cuatro meses. Si aún no fuera suficiente para concluir un acto concreto de investigación, es posible que el Fiscal o querellante recurran en última instancia al Colegio de Jueces para solicitar una última prórroga que puede ser concedida con un plazo máximo de cuatro meses.

En la actualidad el laboratorio recibe requerimientos periciales con un importante volumen digital para ser examinado, habiéndose alcanzado hasta 10 terabytes en un

solo caso. Teniendo presente las reglas vigentes del proceso penal, uno de los objetivos del laboratorio pericial informático es procurar nuevas formas de trabajo que permitan a futuro disminuir el tiempo de espera asociado a los requerimientos de los organismos jurisdiccionales. El aumento de la capacidad operativa del laboratorio puede llevarse a cabo mediante: a) la incorporación gradual de recursos humanos calificados, b) las mejoras en la metodología de trabajo y técnicas de informática forense, c) la productividad lograda en las tareas operativas mediante la introducción de tecnología informática avanzada y nuevas herramientas de informática forense. Atendiendo a los plazos impuestos para la investigación penal en el código de rito, la escasez de recursos humanos calificados en el organismo pericial restringe las posibilidades de poder dar respuesta a todos los requerimientos de los operadores judiciales. Las mejoras en cuanto a metodología y técnicas de informática forense han ayudado en los últimos años pero no son suficientes.

Planteado el escenario en la Justicia provincial neuquina, el objetivo de este trabajo es abordar la experimentación con nuevas tecnologías que posibiliten la automatización, la ejecución en paralelo de actividades operativas y un mejor aprovechamiento de los recursos informáticos del Gabinete de Pericias Informáticas del Poder Judicial del Neuquén.

2 Problemática actual y abordajes metodológicos y tecnológicos

Hace una década [Roussev & Richard III, 2004] se vaticinó el efecto del crecimiento de la masa de datos a ser analizada en un entorno digital y su impacto en el costo y tiempo de análisis de evidencia digital, pronosticando el colapso en el procesamiento de información con la utilización de un único equipo informático como plataforma para el trabajo pericial.

En Reino Unido [Parsonage, 2009] se informó que era común que las unidades periciales informáticas tuvieran una lista de espera que podía llegar a doce meses. En Estados Unidos [Gogolin, 2010] se recopilaron estadísticas que revelaron que este fenómeno era exponencial, con laboratorios de informática forense que mantenían casos en lista de espera por lapsos que excedían los dos años. En los últimos años hay estudios que indican que estos tiempos se han reducido pero en general siguen existiendo grandes demoras para la ejecución de pericias informáticas.

En una encuesta realizada a investigadores y peritos informáticos [Al Fadhi et al., 2013] se continúa evidenciando que el aumento volumen de datos y el tiempo que conlleva el procesamiento de la información digital son las principales limitaciones de la actividad pericial informática.

El fenómeno conocido como Big Data[3] tiene en cuenta tres ejes de cambio conocidos como las 3Vs, a saber: velocidad, variedad y volumen, siendo quizás este

[3] Big Data en términos generales refiere a la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semiestructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a una base de datos relacional para su análisis. El concepto de Big Data aplica para toda aquella

último el que provoca el mayor impacto en la labor pericial informática. Hace menos de una década se llevaba adelante una investigación digital sobre dispositivos de almacenamiento que contenían un par de decenas de gigabytes. El volumen del corpus digital que reúne los datos contenidos en los elementos probatorios continúa incrementándose. Actualmente en un laboratorio pericial informático la unidad de medida habitual para casos típicos está en el orden del Terabyte y continúa en aumento.

El escenario actual que conlleva al desbordamiento de las listas de espera ha sido correctamente caracterizado por un incremento del número de dispositivos que son secuestrados por caso, el aumento de casos en los que la evidencia digital es relevante y el volumen de datos que contienen potencial evidencia digital por dispositivo incautado[Choo, 2014].

Para abordar el problema planteado, en algunos casos se pueden utilizar técnicas de triage[4] para priorización de dispositivos, pero con el ordenamiento y procesamiento selectivo se corre el riesgo de perder la posibilidad de hacer un análisis exhaustivo sobre evidencia digital crucial para el caso, bien sea inculpatoria o exculpatoria. Algunas alternativas desde el punto de vista metodológico han incluido la posibilidad de realizar un indexado de los datos y que el triage sea efectuado por los mismos investigadores [Polastro & Eleuterio, 2015], para luego realizar un análisis exhaustivo limitado sólo a aquellos dispositivos que los investigadores hayan indicado como relevantes en función de los hallazgos resultantes de consultas sobre los índices creados a partir de la evidencia digital.

La reducción de datos es un campo promisorio en materia de investigación científica para uso en informática forense, ya que no hay duda que un archivo DLL tiene menos posibilidades de convertirse en evidencia que un archivo JPG. Si bien comienzan a surgir avances para su aplicación intentando minimizar el corpus digital que debe ser sometido a análisis, a la fecha sólo ha tenido un uso acotado en técnicas de hashing negativo para excluir del análisis a aquellos archivos comunes a todos los sistemas operativos o aplicaciones conocidas. Es viable obtener mejoras en el procesamiento de datos mediante data mining e inteligencia computacional, aunque estas técnicas sólo se han aplicado para casos muy puntuales.

Como contraparte al crecimiento exponencial del volumen de datos a ser analizado quizás podría pensarse que un perito informático es capaz de localizar rápidamente los elementos clave en una investigación digital concreta mediante su experiencia y experticia. Dando por descontado que esta simplificación del escenario planteado es propia del bien conocido "efecto CSI"[5], el problema real subyace en la escasez de

información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales.

[4] Triage es un término utilizado en la medicina para priorizar la atención de pacientes en función de su estado de gravedad. Llevado al ámbito de la informática forense, esta técnica se aplica para la selección de aquella evidencia digital que debe ser priorizada para llevar a cabo un posterior análisis forense exhaustivo, en función de diversos indicadores o características que pueden ser determinadas de forma inmediata.

[5] El "efecto CSI" puede ser definido como el aumento de las expectativas de los jurados en los juicios, debido a las avanzadas técnicas científicas que se muestran en esta serie televisiva CSI: Crime Scene Investigation. Esta serie presenta un equipo ficticio de investigadores que

expertos y el tiempo que conlleva la formación de recursos humanos en esta especialidad.

La escalabilidad hacia arriba utilizando una computadora poderosa con una gran cantidad de núcleos y dispositivos de estado sólido conformando RAID[6] es muy costosa y acotada en cuanto a capacidad de almacenamiento, por lo que esta solución tiene alcance limitado. Asimismo, no es posible esperar a que se desarrollen computadoras más rápidas que reduzcan significativamente el tiempo de procesamiento del importante volumen de evidencia digital que debe manejarse en una pericia informática. Los arreglos de discos RAID han ayudado a mejorar la velocidad de lectura y escritura de datos pero deben ser complementados con otro tipo de mejoras en la infraestructura tecnológica. El uso de GPU[7] requiere técnicas de programación específicas y asimismo existen diferencias de velocidades entre la GPU y el acceso a disco que limitan la performance.

El software AD Lab de Access Data es una aplicación para análisis de evidencia digital con procesamiento distribuido que toma las imágenes forenses a ser procesadas desde un servidor de archivos. Ha de tenerse en cuenta que este tipo de solución también tiene una escalabilidad limitada ya que genera un tráfico intensivo de red. El costo de mover datos desde un repositorio central a los nodos de cómputo muchas veces excede a las ventajas del procesamiento distribuido. Existen otras soluciones tecnológicas que involucran el uso de sistemas de archivos distribuidos a fin de reducir el tráfico de red. Es probable que sea conveniente utilizar algún sistema de archivos especial como FUSE[8] o alguna variante similar para optimizar la performance en la automatización de tareas operativas en informática forense.

---

llegan a la escena del crimen para resolver los asesinatos cometidos en el área metropolitana de Las Vegas. La ficción de la televisión ha hecho creer al espectador que pruebas analíticas que demoran días o semanas pueden lograrse en pocas horas. CSI:Cyber es sido un derivado de la serie CSI orientado a los delitos informáticos. Esta secuela de CSI ha tenido fuerte impacto sobre las expectativas de los operadores judiciales y de la sociedad en general, de que los especialistas en informática forense pueden obtener en forma simple y rápida resultados periciales que esclarezcan un hecho delictivo, simplificando cualquier complejidad subyacente en el objeto de estudio.

[6] RAID es un acrónimo de Redundant Array of Inexpensive Disks, refiere a un sistema de almacenamiento de datos en tiempo real que utiliza múltiples unidades de almacenamiento de información digital, entre los que se distribuyen o replican datos. Los beneficios que tiene un RAID respecto a un único dispositivo de almacenamiento son mejoras en uno o varios de los siguientes aspectos, a saber: integridad, tolerancia a fallos, rendimiento y capacidad.

[7] GPU, cuyo significado es Graphic Processing Unit, es un coprocesador dedicado al procesamiento de gráficos u operaciones de coma flotante. Mientras que un CPU (Central Processing Unit) usualmente tiene pocos núcleos de cómputo y está diseñado para procesamiento serial, un GPU tiene una arquitectura de procesamiento masivo en paralelo que consiste en una gran cantidad de núcleos de cómputo diseñados para manejar múltiples tareas en forma simultánea.

[8] FUSE es un sistema de archivos en espacio de usuario (Filesystem in User Space) que permite crear un sistema de archivos virtual, ya que en realidad no almacena datos propios sino que actúa como una visualización o traducción de un sistema de archivos existente o dispositivo de almacenamiento. Algunos ejemplos del sistema de archivos FUSE son GMailFS y también GooFS, siendo este último el sistema de archivos creado por Google para manejar

La práctica habitual de realizar una imagen forense completa en grandes archivos no es compatible con la naturaleza concurrente del procesamiento distribuido. Desafortunadamente una imagen forense es en efecto un gran archivo con acceso aleatorio. Ello atenta contra la utilización de sistemas distribuidos como Hadoop[9], que dividen los archivos en bloques y los distribuyen en varios nodos con almacenamiento local. En este caso, al tener que procesar una imagen forense completa se requeriría transferir grandes cantidades de datos entre diferentes nodos de almacenamiento conforme se fuera requiriendo cada parte que compone la imagen forense, lo que podría limitar las ventajas de performance que se obtienen con el algoritmo MapReduce[10].

4 Selección de la infraestructura de cómputo distribuido

Los primeros abordajes al problema desde la academia [Roussev & Richard III, 2004], [Richard III & Roussev, 2006], avanzaron en el desarrollo de un prototipo de un sistema distribuido para informática forense llamado DELV, que estaba implementado usando un pequeño clúster Beowulf[11] de ocho nodos Linux, un servidor de archivos y un sistema de control interconectado mediante una red Ethernet de alta velocidad. Las limitaciones tecnológicas del equipamiento informático tradicional utilizado para labores periciales plantean la necesidad de una segunda generación de herramientas para informática forense [Ayers, 2009].

Habiéndose detectado la problemática inherente al procesamiento de evidencia digital, en el año 2006 [Gómez, 2006] se expuso la necesidad de sintetizar y ordenar en forma sistemática los componentes involucrados en tareas periciales para reducir tiempos de trabajo mediante la ejecución paralela de actividades forenses, siempre que se cuente con recursos humanos especializados y herramientas adecuadas. La automatización y ejecución concurrente de tareas operativas en el ámbito pericial informático fue referida en otro trabajo previo en esta línea de investigación [Gómez, 2007] y se delinearon actividades operativas que eran susceptibles de ser realizadas en

archivos grandes de manera rápida y eficiente, que soporta toda su infraestructura informática de procesamiento de información en nube.

[9] Hadoop es un proyecto de software de código abierto que ofrece una plataforma para el procesamiento distribuido de grandes volúmenes de datos entre grupos de computadoras (clúster) utilizando modelos simples de programación. Ha sido diseñado para permitir escalabilidad desde un único servidor hasta miles de equipos informáticos, cada uno utilizando su capacidad local de cómputo y almacenamiento.

[10] MapReduce es una técnica de programación para permitir el procesamiento de grandes volúmenes de datos en un sistema distribuido. Los programas escritos en este estilo funcional son automáticamente paralelizados y ejecutados en grandes clústeres de cómputo. Asimismo, se reduce el tráfico de datos en la red ya en lugar de tener que transferirse la totalidad de los datos a un determinado recurso de cómputo, cada nodo procesa un subconjunto de estos datos en forma local.

[11] Beowulf es un clúster de recursos de cómputo conformado usualmente por computadoras personales que trabajan en forma dedicada para la ejecución de tareas de cómputo de alto rendimiento.

paralelo mediante la incorporación de tecnología, destacándose: a) preservación de evidencia digital (adquisiciones simultáneas de dispositivos), b) búsquedas de patrones (cadenas de texto, imágenes a través de hashing), c) descifrado de contraseñas. Sobre este último punto, años atrás se implementó una infraestructura de cómputo distribuido utilizando máquinas virtuales [Gómez, 2009]. Se utilizaron nodos implementados con máquinas virtuales que contenían una distribución Linux reducida especialmente preparada para implementar un clúster con Openmosix[12] y se realizaron pruebas de rendimiento utilizando John the Ripper[13] con MPICH[14] para descifrado de contraseñas, logrando una importante reducción del tiempo que demanda esta tarea. En la actualidad, a nivel comercial existe software especializado que utiliza computación distribuida para descifrado de contraseñas desarrollado por empresas como Passware y Elcomsoft.

Aunque en forma aislada, se han presentado implementaciones exitosas aplicadas a informática forense para indexado de datos utilizando HTCondor[15], así como también con Hadoop [Silva, 2012], y se han logrado mejoras en los tiempos de hashing utilizando BOINC[16] [Revenga del Toro et al., 2009].

Para la implementación de un sistema distribuido que permita aumentar la performance del laboratorio pericial informático se consideraron diferentes soluciones tecnológicas como Hadoop, BOINC y HTCondor, finalmente optando en una primera instancia por este último en función de las facilidades que brinda para la ejecución directa de múltiples aplicaciones de informática forense sin requerir modificaciones en el código de los programas.

HTCondor es un planificador de trabajos que utiliza la capacidad ociosa de las computadoras y puede ejecutarse concurrentemente sobre computadoras no dedicadas conectadas a través de una red de datos. Este software permite aprovechar recursos de cómputo subutilizados a través de la integración y conformación de plataformas computacionales heterogéneas que son capaces de procesar una gran cantidad de tareas por unidad de tiempo.

Las aplicaciones pueden ser lanzadas en cualquier equipo de computación que forme parte del pool de recursos de HTCondor sin que sea necesario realizar

[12] Openmosix es un sistema de clúster para Linux que permite a varias máquinas actuar como un único sistema multiprocesador.

[13] John the Ripper es una aplicación de criptografía que realiza una búsqueda exhaustiva en un espacio de posibles combinaciones para descifrar contraseñas.

[14] MPICH es una implementación portable y de alto rendimiento del estándar de paso de mensajes MPI para aplicaciones de memoria distribuida que utilizan computación paralela.

[15] HTCondor es un sistema de manejo de carga de trabajo especializado para la ejecución de tareas de cómputo intensivo desarrollado y mantenido por la Universidad de Wisconsin Madison. El software, código fuente y documentación están disponibles en forma gratuita bajo una licencia ASL. La licencia Apache (ASL) es una licencia de software gratuito permisiva escrita por la Apache Software Foudation (ASF). La licencia Apache requiere la preservación de la información de Copyright y de delimitación de responsabilidades. Disponible en: https://research.cs.wisc.edu/htcondor/ Último Acceso: 06-04-2016

[16] BOINC, acrónimo de Berkeley Open Infrastructure for Network Computing, es una plataforma libre que permite crear un proyecto de computación distribuida utilizando los ciclos de CPU o GPU ociosos que no son consumidos por el usuario de un equipo informático.

modificaciones al código del programa. Si se cuenta con el código fuente de una aplicación, es posible incorporarle bibliotecas adicionales que permiten potenciarlo con el uso de checkpoints o puntos de control que permiten resguardar el progreso de la tarea y si es necesario poder suspender la ejecución de la misma en un nodo y reasignarla a otro del pool para completar el trabajo.

HTCondor es ideal para la gestión de tareas que son completamente desatendidas y paralelizables, en el sentido de que pueden ser vistas como sistemas cerrados que no requieren interacción con el usuario y que no dependen de la ejecución de ninguna otra tarea. Podemos verlo entonces como un gestor oportunista de tareas, el cual permite utilizar computadoras conectadas a través de una red de datos y que busca ciclos de CPU subutilizados para llevar a cabo la ejecución de aplicaciones en paralelo.

Entre las características destacadas de HTCondor es dable mencionar que:

- Utiliza un modelo de cómputo oportunista que permite integrar todo tipo de equipos de computación, desde servidores hasta computadoras de escritorio;
- Gestiona y asigna en forma automática tareas a recursos computacionales;
- Contiene un lenguaje propio de definición llamado ClassAds que permite realizar el emparejamiento (matchmaking) entre de tareas y nodos;
- Ha sido diseñado para proveer un ambiente de ejecución confiable, asegurando la ejecución de una tarea aún ante la presencia de fallas en el entorno, como cortes de energía y fallas en la red de datos;
- Permite utilizar diversos ambientes de ejecución, conocidos como Universos, que ofrecen flexibilidad y capacidad de adaptación para múltiples tipos de tareas. El universo Vanilla acepta como tarea a aquellas aplicaciones que se ejecutan desde la línea de comandos. El universo Standard permite crear tareas que utilicen puntos de control con capacidad de detener y luego reanudar aplicaciones o bien ser migradas entre nodos o, siempre que puedan enlazarse algunas bibliotecas adicionales al código fuente de la aplicación. Asimismo, otros ambientes como el universo Java, Parallel, Grid, VM, Local y Scheduler brindan prestaciones más específicas sobre el clúster.

5 Automatización de tareas de informática forense

Está claro que la automatización no irá en detrimento de la calidad de las investigaciones digitales [James & Gladyshev, 2013] si ésta es aplicada en las etapas adecuadas e implementada en forma correcta, contribuyendo a reducir el tiempo de la investigación, lo que finalmente permite mitigar el crecimiento de la lista de espera de casos en trámite.

Muchas tareas periciales que involucran un elevado tiempo de procesamiento son pasibles de ser automatizadas, como la detección de imágenes sospechosas de contener pornografía, la generación de índices para realizar consultas dinámicas sobre el corpus digital, la creación de diccionarios personalizados para descifrado de contraseñas en base a la información digital contenida en los dispositivos de almacenamiento, la generación de listas de valores hash para análisis de firmas, la

detección de malware o localización de archivos relevantes a la investigación, y la búsqueda de evidencia digital mediante palabras clave.

Actualmente las herramientas comerciales para informática forense automatizan una serie de tareas operativas que son habituales durante el análisis de evidencia digital, sin embargo estas aplicaciones no hacen uso de múltiples procesadores por lo que tienen un bajo rendimiento cuando se trabaja con importantes volúmenes de información digital. Encase y FTK [Access Data, 2015] poseen versiones que utilizan procesamiento distribuido. Este tipo de tecnología es excesivamente onerosa y por otra parte todavía no hay una gran aceptación en la comunidad forense, fundamentalmente basadas en críticas en cuanto a la madurez y fiabilidad de estas aplicaciones con capacidades de cómputo distribuido. En los últimos años han surgido nuevas alternativas para el procesamiento masivo de información digital como Xiraf[17] [Van Baar et al., 2014] y Hansken[18], ambos desarrollados en el Netherland Forensic Institute, pero no están disponibles para ser implementados por otros laboratorios de informática forense.

A fin de poner en producción el entorno de cómputo distribuido se crearon como pruebas de concepto los scripts necesarios para el envío de trabajos al cluster, priorizando aquellas tareas de preprocesamiento y de postprocesamiento de evidencia digital que resultan de mayor utilidad durante el desarrollo de actividades operativas. El objetivo final apunta a la integración y escalabilidad de las herramientas de informática forense disponibles en el laboratorio, permitiendo un mejor aprovechamiento de los recursos tecnológicos. A fin de posibilitar la planificación de trabajos en el cluster es necesario que las aplicaciones admitan la ejecución directa por línea de comandos sin recurrir a una interfaz gráfica. Principalmente se ha utilizado Python, Java y AutoIt como lenguajes de programación para las tareas de automatización.

Básicamente existen dos bloques de código que permiten automatizar las actividades operativas en el clúster HTCondor, siendo uno de ellos un archivo de tarea con extensión .sub y el otro un archivo de procesamiento por lotes, con extensión .bat en el caso de efectuar la ejecución sobre un sistema operativo

[17] Xiraf es una aplicación de software avanzada para informática forense desarrollada en el Netherland Forensic Institute que puede analizar en forma automatizada grandes cantidades de datos provenientes de diversos dispositivos. Manteniendo un alto rendimiento en el procesamiento de evidencia digital, permite que los resultados queden disponibles en forma inmediata para que los responsables de las investigaciones judiciales puedan realizar búsquedas y seleccionar información relevante. Disponible en: https://www.forensicinstitute.nl/products_and_services/forensic_products/xiraf/ (Último acceso: 06-04-2016).

[18] Hansken es el sucesor de Xiraf, la herramienta de software que comenzó como un proyecto de investigación en el año 2006 en el Netherland Forensic Institute y fue transferida a la Policía Nacional de Holanda en el 2010. Xiraf fue el primer software de informática forense holandés que permitió el procesamiento computarizado de la información digital recolectada en procedimientos judiciales. Según manifiestan sus desarrolladores, la mejora de Hansken es sustancial, ya que lo que ha Xiraf le insumía 24 horas para completar el procesamiento de evidencia digital, Hansken lo hace en 30 minutos. Disponible en: https://www.forensicinstitute.nl/research_and_innovation/innovation_dev_programmes/hansken/ (Último acceso: 06-04-2016).

Windows, que será finalmente el responsable de invocar a la herramienta de informática forense con sus parámetros requeridos y eventualmente otras tareas adicionales de preprocesamiento y postprocesamiento de datos.

Sintéticamente se describen las funcionalidades de aquellos trabajos que son automatizados y gestionados por el planificador de HTCondor:

- Hasher.sub: obtiene hashes de un conjunto de archivos, los que luego serán utilizados para autenticar la evidencia digital. Posee un alto grado de paralelismo ya que cada slot de HTCondor efectúa procesamiento sobre archivos individuales.
- Indexer.sub: genera un índice sobre el corpus digital para permitir consultas dinámicas. Para la implementación efectuada mediante el uso de una herramienta comercial para indexación se ha obtenido un grado medio de paralelismo, teniendo en cuenta que cada slot de HTCondor es responsable de indexar la totalidad de archivos que componen una imagen forense completa, la que previamente debe ser montada en un repositorio compartido.
- VideoExtractor.sub: se encarga de extraer un conjunto de fotogramas partiendo de un lote de archivos de video digital. Posee un alto grado de paralelismo ya que cada slot de HTCondor efectúa procesamiento sobre archivos individuales.
- Mover.sub: transfiere evidencia digital entre repositorios digitales garantizando la integridad de los datos mediante el cálculo de MD5 en origen y en el destino. Posee un alto grado de paralelismo ya que cada slot de HTCondor efectúa procesamiento sobre archivos individuales.

Una característica típica de las actividades operativas automatizables que habitualmente son ejecutadas en una primera etapa del análisis pericial informático es que requieren poca o nula coordinación entre ellas, lo cual permite lograr una buena paralelización de trabajos dentro del clúster. Por ejemplo, es viable utilizar el HTCondor para indexar evidencia digital proveniente de diversos dispositivos, estar extrayendo fotogramas de archivos de video digital y a la vez mover datos pertenecientes a otros casos entre repositorios digitales.

Ha de tenerse presente que muchas de las aplicaciones de informática forense utilizadas en el laboratorio son compatibles con un único sistema operativo. Ello no se traduce en ninguna limitante operativa para HTCondor, ya que el planificador permite definir requerimientos específicos previos a la ejecución de trabajos en el clúster.

Por otra parte en HTCondor es posible utilizar aprovisionamiento dinámico, lo que permite asignar slots con mayor cantidad de unidades de cómputo y memoria para ciertas tareas planificada que demanden mayores recursos, como por ejemplo aquellas aplicaciones que implementan rutinas multihilo.

En el laboratorio se cuenta con equipos de computación que tienen sistemas operativos Windows y GNULinux. Teniendo presente que la mayor parte de las aplicaciones está orientada a Windows, para una primera integración de herramientas de informática forense se aprovecharon en el clúster aquellos recursos de cómputo que utilizan este sistema operativo. El clúster se ha implementado sobre tres servidores con sistema operativo Windows 2012 Server que cuentan con las siguientes características: 2 IBM Bladecenter HS21 con 2 procesadores Intel Xeon E5345 2.66 GHz de 4 núcleos y 32 GB de memoria RAM y 1 IBM Bladecenter HS22 con 2 procesadores Intel Xeon E5620 2.40 GHz de 8 núcleos y 38 GB de memoria RAM.

Bien sea que se sigan lineamientos planteados en guías de mejores prácticas y protocolos de actuación, así como también rigurosos procedimientos operativos estandarizados, una tarea usual en la actividad pericial informática es la generación de imágenes forenses. Las estadísticas de trabajo recabadas en el Laboratorio Pericial Informático del Poder Judicial del Neuquén indican que esta tarea conlleva una significativa cantidad de tiempo para ser completada.

Algunos investigadores han intentado brindar nuevas soluciones para mejorar la performance de esta labor proponiendo la generación de imágenes forenses selectivas que sólo resguarden aquellos contenidos digitales que no sean triviales para la pericia informática, partiendo de un conocimiento previo de la casuística. Otras ideas han apuntado a evitar la ejecución de actividades operativas en forma secuencial, sugiriendo que pueda paralelizarse la generación de una imagen forense y la extracción simultánea de información digital sensible para la etapa de análisis. Estas soluciones todavía no han sido extendidas e integradas con las herramientas de trabajo típicas de un laboratorio de informática forense.

En cualquier laboratorio de informática forense que tenga un grado de evolución alineado a los avances tecnológicos, la generación de imágenes forenses en dispositivos de almacenamiento individuales ha quedado en el olvido y sólo tiene lugar en circunstancias excepcionales. Aunque aún puede ser adecuada para actividades periciales practicadas por peritos informáticos que trabajan en forma autónoma y fuera de un laboratorio de informática forense, esta modalidad tradicional de trabajo tiene múltiples limitaciones en materia de escalabilidad y procesamiento masivo de información digital. Actualmente en un laboratorio de informática forense se utilizan grandes repositorios digitales interconectados mediante una red de datos de alta velocidad, lo que permite un acceso simultáneo al corpus digital y el trabajo concurrente del equipo profesional en diversas actividades operativas. A fin de evitar el congestionamiento de la red, la generación de imágenes forenses es una de las tareas automatizables que podría ser gestionada por el planificador HTCondor, pudiendo plantearse un nivel de prioridad para la ejecución automática en un horario determinado.

Habiendo comenzado a integrar y experimentar con algunas herramientas de informática forense, las posibilidades de automatización y escalabilidad de tareas operativas aumentan conforme se van procurando otras aplicaciones que admiten ser ejecutadas desde la línea de comandos.

Una ventaja adicional de la automatización y la ejecución de tareas en forma concurrente es que muchos de los resultados obtenidos pueden ser puestos a disposición de los operadores judiciales, posibilitando hacer avances en la investigación penal, como es el caso de los índices construidos sobre la evidencia digital o los archivos de imagen y video que son extraídos del corpus digital mediante aplicaciones para detección de desnudez o de rostros.

6 Conclusiones

Los laboratorios de informática forense deben prepararse para los desafíos actuales que conllevan al desarrollo de pericias informáticas con grandes volúmenes de

información digital, así como también un perfeccionamiento constante de sus profesionales para enfrentar los cambios y la velocidad con que llegan al mercado una amplia variedad de dispositivos electrónicos susceptibles de albergar evidencia digital. El tiempo disponible de los recursos humanos más calificados de un laboratorio de informática forense -en este caso los peritos informáticos- es uno de los elementos más valiosos a resguardar a fin de que puedan minimizar su intervención en tareas de rutina y abordar problemas de mayor complejidad.

El objetivo de este trabajo ha sido exponer una implementación concreta de una infraestructura de cómputo distribuido en el Gabinete de Pericias Informáticas del Poder Judicial del Neuquén, sobre la que se han desarrollado pruebas de concepto con scripts específicos que posibilitan la ejecución en paralelo de herramientas de informática forense. Sin perjuicio de culminar la experimentación con algunas pruebas de concepto adicionales sobre este clúster, es probable que a futuro se consideren otras alternativas que posibiliten mayores ventajas en cuanto a performance y ejecución en paralelo como OpenMPI[19]. Se prevé que las tareas automatizadas que vayan siendo implementadas e incorporadas al proceso forense digital, bien sea sobre este clúster o con cualquier otra tecnología de cómputo distribuido, puedan ser invocadas desde el sistema informático de gestión de casos del laboratorio a fin de mantener un control de gestión, generando métricas de esfuerzo y asegurando la trazabilidad de todas las actividades operativas sobre la evidencia digital.

Posibilitando la integración de recursos computacionales se ha iniciado el camino hacia una segunda generación de herramientas de informática forense que permitan paralelizar actividades operativas, intentando reducir la gran demanda de tiempo que conlleva el proceso forense digital y procurando el máximo aprovechamiento de la infraestructura tecnológica del laboratorio dedicada para procesamiento de evidencia digital.

Referencias

1. Access Data (2015), AD Lab: Reducing case backlogs through distributed processing and collaborative analysis, White Paper. Disponible en: http://marketing.accessdata.com/AD-LABX-WHT-DivConquer, Último acceso el : 07/04/2016.

2. Al Fahdi, M.,Clarke, N. & Furnell, S. (2013), Challenges to digital forensics: A survey of researchers & practitioners attitudes and opinions, *Information Security for South Africa*, IEEE, pp. 1–8.

3. Ayers, D. (2009), A second generation computer forensic analysis system, *Digital Investigation*, *6*, S34-S42.

[19] Open MPI refiere a Open Source Message Passing Interface. Se trata de una implementación de código abierto de la interfaz de paso de mensajes MPI que es desarrollada y mantenida por un consorcio que proviene de la academia, la investigación y la industria. Se caracteriza por su alta eficiencia y prestaciones, facilitando la programación paralela y la ejecución en sistemas distribuidos.

4. Gogolin, G. (2010), The Digital Crime Tsunami, *Digital Investigation*, Volume 7(1), pp.3–8 doi:10.1016/j.diin.2010.07.001

5. Gómez, L. (2006), La investigación de actividades delictivas con alta tecnología, *Simposio Argentino de Informática y Derecho*, JAIIO, Mendoza.

6. Gómez, L. (2007), Tecnologías, técnicas y estimadores de informática pericial aplicada a investigaciones judiciales, *Simposio Argentino de Informática y Derecho*, JAIIO, Mar del Plata.

7. Gómez, L., (2009), Investigación científica del delito: tecnologías de virtualización aplicadas en informática forense, Simposio Argentino de Informática y Derecho, JAIIO, Mar del Plata, Argentina.

8. James, J. & Gladyshev, P. (2013), Challenges with Automation in Digital Forensic Investigations, *Computers and Society*, p. 17.

9. Lillis, D., Becker, B., O'Sullivan, T. & Scanlon, M., Current Challenges and Future Research Areas for Digital Forensic Investigation, *The 11th ADFSL Conference on Digital Forensics, Security and* Law (CDFSL 2016), Daytona Beach, Florida, USA, May 2016.

10. Parsonage, H. (2009), Computer Forensics Case Assessment and Triage, Disponible en: http://computerforensics.parsonage.co.uk/triage/ComputerForensics CaseAssessmentAndTriageDiscussionPaper.pdf, Último Acceso: 10-03-2016.

11. Polastro, M. & Eleuterio, P. (2015), Um Modelo de Triagem de Dados Digitais Aplicado à Perícia Criminal em Informática, *Simposio Argentino de Informática y Derecho*, JAIIO, Santa Fe, ISSN: 2451-7526

12. Quick, D. & Choo, K.-K.R. (2014), Impacts of increasing volume of digital forensic data: A survey and future research challenges, *Digital Investigation*, p.doi:10.1016/j.din.2014.09.002

13. Revenga del Toro, P., Gardel Vicente, A. & García Jiménez, R. (2009), "Cluster de alta computación para aplicaciones de informática forense. Disponible en: http://www.iuisi.es/15_boletines/15_2009/Cluster_d...pdf, Último acceso el: 07/04/2016.

14. Richard III, G. & Roussev, V., (2006), Digital forensics tools: the next generation, *Digital Crime and Forensic Science in Cyberspace,* Idea Group Publishing, pp.75–90.

15. Roussev, V. & Richard III, G., (2004), Breaking the performance wall: The case for distributed digital forensics, *Proceedings of the 2004 Digital Forensics Research Workshop*.

16. Silva, M. (2012), Utilização da computação distribuída para o armazenamento e indexação de dados forenses. 2012. xxii, 138 f., il. Dissertação (Mestrado em Engenharia Elétrica)—Universidade de Brasília, Brasília, 2012. Disponible en: http://repositorio.unb.br/handle/10482/10864, Último acceso el: 07/04/2016.

17. Van Baar, R., Van Beek, H. &Van Eijk, E. (2014), Digital Forensics as a Service: A Game Changer. *Digital Investigation*, 11:S54–S62.