Arquitectura referencial de Big Data para la gestión de las telecomunicaciones

Referencial architecture of Big Data for the management of telecommunications

Lieter Plasencia Moreno[1]* Caridad Anías Calderón[1]

Recibido 4 de diciembre de 2015, aceptado 28 de noviembre de 2016
Received: December 4, 2015 Accepted: November 28, 2016

RESUMEN

El alto desarrollo alcanzado en tecnologías de la información a nivel global y el intenso uso por parte de los usuarios de las mismas han provocado el incremento de los volúmenes de datos que se transportan mediante las redes. La búsqueda de nuevos métodos para gestionar dichos datos conllevó al surgimiento del término Big Data, surgiendo un nuevo paradigma en la gestión de los mismos.
Actualmente son numerosas las empresas que se encuentran integrando dicha tecnología a sus plataformas y redes. Este artículo presenta una arquitectura referencial de Big Data aplicable a la gestión de las redes, servicios y aplicaciones de telecomunicaciones, permitiendo la optimización de los mismos. Además, la arquitectura se aplicó en dos casos: un caso de estudio enfocado en la seguridad de una red y un caso práctico aplicado en una empresa del sector de las telecomunicaciones.

Palabras clave: Big Data, gestión, telecomunicaciones.

ABSTRACT

The high development reached in information's technologies to the global level, and the intense use of the users of the same ones have caused the increment of the volumes of data that through the networks are transported. The search for new methods to manage this data bore to the emergence of the term Big Data, arising a new paradigm in the management of the same ones.
Now, they are numerous the companies that are integrating this technology into their platforms and networks. This article presents a referential architecture of Big Data applicable to the management of the networks, services and applications of telecommunications, allowing the optimization of the same ones. In addition, the architecture was applied in two cases: a study case focused on network security and a practical case applied in a company of the sector of telecommunications.

Keywords: Big Data, management, telecommunications.

INTRODUCCIÓN

En los últimos años se ha apreciado una evolución acelerada de las Tecnologías de la Información y las Comunicaciones (TIC), destacándose el incremento de la interacción de los usuarios con las mismas, lo que ha provocado la generación de grandes cantidades de datos que hasta hace poco no existían. Además, se ha presenciado la aparición de nuevos tipos de datos, muchos de los que no presentan una estructura adecuada para procesarlos y almacenarlos.

[1] Departamento de Telecomunicaciones y Telemática. Instituto Politécnico Superior José Antonio Echeverría (IPSJAE). Calle 114 Nº 11901, e/ 119 y 127, Marianao, La Habana, Cuba. C.P. 19390. E-mail: lieter.pm@gmail.com; cacha@tesla.cujae.edu.cu
* Autor de correspondencia

De esta forma surge el término Big Data, implicando una nueva forma de gestionar el alto nivel de datos que existen y que se generan a nivel global en la actualidad y aprovecharlos en función de lograr las metas que se trazan las distintas empresas y organizaciones de diferentes sectores. Debido al nuevo panorama que presenta Big Data en la gestión de las redes y servicios se han desarrollado diferentes herramientas y plataformas para el procesamiento de datos masivos que se generan desde diferentes fuentes en las redes y obtener de los mismos, información valiosa que pueda ser utilizada en la oferta de diferentes servicios.

La cantidad de datos que se generan producto de la gestión de las telecomunicaciones, incluyéndose datos de los dispositivos, del consumo de los usuarios, del tráfico de las redes y del control de los servicios, entre otros; además, el hecho de que estos datos incrementan con la adopción masiva de las TIC por parte de los usuarios y sin la existencia de métodos adecuados que faciliten su integración, han limitado la obtención de información valiosa que pueda ser empleada para mejorar el funcionamiento de las redes y para ofrecer nuevos y optimizados servicios. Por ello es necesario evolucionar a nuevas soluciones que permitan integrar los datos que se procesan en los sistemas de gestión de telecomunicaciones.

En este trabajo se realiza una propuesta para emplear Big Data a la obtención e integración de información valiosa para la gestión de las redes y servicios de telecomunicaciones.

ANÁLISIS TEÓRICO

Big Data es la combinación de viejas y nuevas tecnologías que ayudan a las empresas a obtener una mejor visión del procesamiento de su información. También se puede conceptualizar como la capacidad de manejar un inmenso volumen de datos que se generan de forma caótica, que a la velocidad y temporización correctas, permite el análisis en tiempo real y la definición de las acciones asociadas necesarias [1].

Big Data es utilizado para referirse a aquellos grupos de datos que por su elevado volumen, diversidad y complejidad no se pueden analizar, visualizar y almacenar con las herramientas y almacenes de datos tradicionales, por lo que también se le asocian los conceptos de "Datos masivos o "Datos a gran escala". Son muchas las características o definiciones que existen de Big Data, no obstante, la mejor forma de describir esta tecnología es a partir de parámetros específicos denominados las V de Big Data. Estas son "Volumen", "Velocidad", "Variedad", "Veracidad", "Valor", "Validez" y "Visualización", algunas de las que se muestran en la Figura 1 [2].

Figura 1. Las "V" de Big Data.

El "Volumen" hace referencia al tamaño de datos, los que van creciendo exponencialmente. Estos se van generando a distintas velocidades, de ahí el parámetro "Velocidad". La "Variedad" viene dada por dos razones principales: la primera se debe a que los datos se generan desde diferentes fuentes geográficamente distribuidas, y la segunda, a la existencia de diferentes tipos de datos: estructurados, semiestructurados y no estructurados. La "Visualización" constituye una parte muy importante de cualquier entorno Big Data. El adecuado empleo de los métodos de visualización de datos puede conllevar a la obtención de excelentes resultados aplicables a los objetivos que se trazan.

La "Veracidad" hace referencia al nivel de fiabilidad asociado a ciertos tipos de datos, tratándose de eliminar la incertidumbre que estos puedan aportar. La "Validez" es un concepto que tiende a confundirse con la "Veracidad". Los datos pueden ser fiables y no presentar errores pero si no son correctamente

comprendidos no son válidos. Y, por último, el objetivo final del empleo de Big Data es generar "Valor" a partir de la información almacenada que se obtiene de manera eficiente y con menor costo posible por medio de distintos procesos.

Dado el gran potencial que poseen las tecnologías Big Data para el procesamiento y gestión de grandes volúmenes de datos, se ha despertado un gran interés por aplicar dicha tecnología en muchos sectores. La aplicación de Big Data en las telecomunicaciones contribuye al desarrollo de nuevos y potentes sistemas de gestión de las redes y servicios, gracias a nuevas técnicas y herramientas para procesar los datos.

Algunas de las aplicaciones de Big Data que se pueden mencionar en este sector son:

- El óptimo almacenamiento de datos masivos en la nube mediante almacenes virtualizados [3].
- Los sistemas de gestión de seguridad y monitoreo de redes.
- Las técnicas de reubicación de información en nodos con gran ancho de banda, con el objetivo de disminuir el tiempo de transmisión de la información y la obtención de rutas más óptimas para el transporte de datos.
- La integración de los datos que circulan por las redes.
- El análisis de la información para la detección de fallos [4].
- La visualización de diferentes tipos de datos [5].
- La creación de *frameworks* para mejorar los servicios de comunicaciones móviles mediante la evaluación de la calidad de experiencia (QoE: por sus siglas del término en inglés *Quality of Experience*) de los usuarios.

Empleo de la virtualización en los entornos Big Data

Actualmente los sistemas *frameworks* y arquitecturas de gestión se caracterizan por el alto nivel de virtualización. Por ello, se ha considerado necesario el empleo de la virtualización en el contexto de la arquitectura propuesta. La virtualización provee las bases para el acceso, almacenamiento, análisis y gestión de los sistemas de datos distribuidos.

Una de las principales razones por la que se utiliza la virtualización es para optimizar el desempeño y la eficiencia de procesamiento de un sistema pues permite un mejor uso y control de los recursos físicos siendo necesario un alto nivel de seguridad. El empleo de la virtualización en un entorno Big Data incrementa la escalabilidad, al añadir eficiencia en cada capa de la infraestructura, favoreciendo el análisis de los datos masivos. Además, las herramientas de análisis y gestión de datos masivos trabajan de forma más eficiente en ambientes virtualizados.

RESULTADOS

A partir de la investigación realizada, se procedió a elaborar la arquitectura referencial de Big Data para la gestión de las telecomunicaciones. Algunos de los principios que formaron la base de la propuesta fueron:

1. Es necesario conocer qué datos son relevantes a los objetivos que se persiguen al emplear Big Data.
2. Se requieren los procesos de extracción, transformación y carga que garantizan la captura y almacenamiento de todo tipo de datos relevantes.
3. Se debe garantizar la transformación de los datos que no presentan una estructura adecuada para su posterior análisis.
4. Es posible emplear tanto sistemas de almacenamiento físicos como la nube para el almacenamiento de los datos masivos, pero siempre se debe garantizar que se almacenen todos los datos que se capturen y se procesen.
5. Las herramientas de análisis de datos a emplear estarán determinadas por los servicios que se desean proveer.
6. Se debe emplear la virtualización debido a las ventajas que proporciona.
7. Es necesario seguir un modelo de gestión distribuido, puesto que las fuentes de las que se obtendrán los datos se encuentran geográficamente distribuidas.
8. Se debe garantizar en todo momento la seguridad de los datos, siendo este uno de los retos de la gestión de datos masivos.

La arquitectura referencial de Big Data para la gestión de las telecomunicaciones propuesta se muestra en la Figura 2. Como se aprecia en ella, en el nivel más bajo de la arquitectura, se encuentran las fuentes que generan grandes flujos de datos a

diferentes velocidades y desde distintos puntos geográficos. En un segundo nivel, aparecen los procesos de extracción, transformación y carga (ETL: por sus siglas del término en inglés *Extraction, Transformation and Load*) de los datos masivos [6].

El objetivo es extraer los datos de distintas fuentes y enviarlos a los repositorios donde se almacenan. Los procesos de transformación y carga pueden ocurrir de dos formas principales: la primera, los datos inicialmente son cargados en las bases de datos que los almacenarán y dentro de estas se hacen las transformaciones necesarias, lo que facilita que las herramientas de análisis de datos los procesen y entreguen información clara y entendible, y; la segunda, los datos son transformados previamente al almacenamiento de los mismos.

En el tercer nivel de la arquitectura se considera el almacenamiento de datos masivos. Este nivel puede variar de una implementación a otra de la arquitectura, puesto que existen herramientas ETL que no solo transforman los datos sino que presentan espacios de almacenamiento para grandes volúmenes de información, no requiriéndose emplear bases de datos adicionales. Además, cada empresa u organización donde se aplique la arquitectura que se propone, puede determinar, de acuerdo a los tipos de datos con los que va a trabajar, cómo almacenarlos.

En el cuarto nivel de la arquitectura se considera el análisis de datos, en el que se emplean herramientas que se encargarán de obtener información de alto nivel de impacto, útil para la gestión de las redes y servicios de telecomunicaciones. En este nivel se emplean herramientas de análisis predictivo de datos, algoritmos que establecen puntos de interrelación dentro de grandes volúmenes de datos, herramientas de visualización que permitan representar información de interés sobre las redes y servicios para una empresa de telecomunicaciones.

Finalmente, en el último nivel de la arquitectura referencial propuesta, se encuentra la gestión de las redes y servicios de telecomunicaciones, la que se ve optimizada gracias al análisis de los datos masivos para, por ejemplo, lograr la configuración eficiente de los dispositivos de interconexión de redes, la mejora en los servicios telefónicos, una mayor calidad de las ofertas a los clientes y la optimización de las redes.

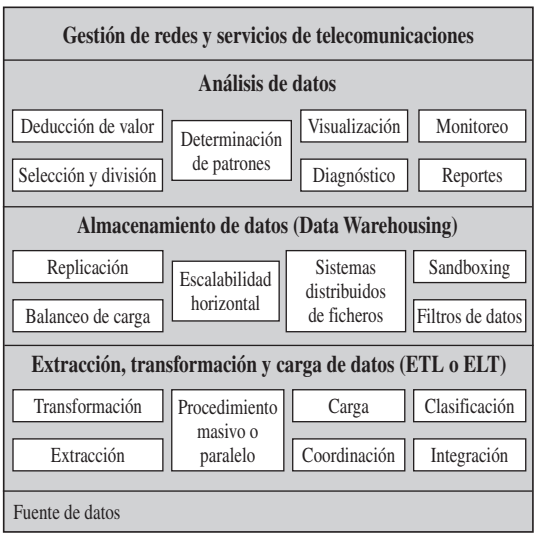

Figura 2. Propuesta de arquitectura referencial de Big Data para la gestión de las telecomunicaciones.

En los procesos que ocurren en los niveles de extracción, transformación y carga, almacenamiento y análisis de datos, se emplea el término de datos masivos, pues solo después que los datos salen del nivel de análisis de datos, es que estos se consideran información relevante, es decir, información que ya puede ser aplicada en la gestión de redes y servicios de telecomunicaciones.

Nivel de extracción, transformación y carga de la arquitectura propuesta

Uno de los niveles de la arquitectura referencial de Big Data para la gestión de las telecomunicaciones que se propone es el de extracción, transformación y carga, cuyos procesos son pilares al planificar y diseñar una infraestructura de manejo de datos, que implique la integración de diferentes y variadas fuentes. Estos procesos son los responsables de recopilar la información de las fuentes de origen de datos y de adaptarla, filtrarla e integrarla en un repositorio digital como, por ejemplo, una base de datos.

Se precisan, en la Figura 2, los principales procesos a ejecutarse en el nivel de extracción, transformación y carga, cuyas características son:

– Proceso de extracción: es el proceso inicial, en el que se obtienen los datos de las fuentes de origen. Habitualmente, con el objetivo de evitar saturación en los servidores donde finalmente se almacenarán los datos, se suelen implementar repositorios intermedios, conocidos como bases de datos operacionales o almacenes de datos operacionales, que actúan de pasarelas entre las fuentes de datos y el sistema destino de la información.
– Proceso de transformación: cuando los datos proceden de distintas fuentes, lo normal es que no coincidan en formato. Debido a esto, resulta imprescindible realizar tareas de transformación para, entre otros problemas, evitar duplicidades innecesarias de datos o impedir que se establezcan grupos de datos que no presentan conexiones entre ellos. En este proceso se llevan los datos extraídos a una estructura lógica común necesaria para su procesamiento y análisis posterior.
– Proceso de carga: este es el proceso en el que se cargan los datos, ya estructurados en el formato deseado, en el sistema de almacenamiento destino donde posteriormente serán procesados y analizados.
– Proceso de clasificación: es el proceso que permite la clasificación de los datos que se extraen en diferentes dimensiones para la simplificación de futuros procesamientos.
– Proceso de integración: es el proceso mediante el que se logra la armonización de datos de distintas fuentes y su integración en un grupo único antes de ser transformados y reducidos en un formato común.
– Proceso de coordinación: se refiere al proceso que mantiene y controla a todos los demás procesos que se realizan en este nivel de la arquitectura.
– Procesamiento masivo paralelo (MPP): proceso que realiza la división de tareas que serán procesadas al mismo tiempo y de forma aislada. De esta forma, el sistema es más eficiente en el procesamiento de datos [7].

Nivel de almacenamiento de datos de la arquitectura propuesta

El concepto de almacenes de datos se originó hace varias décadas. Inicialmente se concibió para que fuesen utilizados por usuarios que administraban sistemas operacionales que necesitaban almacenar información para apoyar la toma de decisiones. Con la llegada de Big Data, el concepto de almacén de datos ha evolucionado, no obstante, los almacenes de datos tradicionales siguen siendo usados debido a que son eficientes en el análisis de datos operacionales antiguos.

Los almacenes de datos tradicionales soportan datos estructurados, están optimizados para propósitos específicos y generalmente son centralizados. Con la aparición de Big Data, se ha pensado en almacenes de datos híbridos, en los que se encuentren tanto los datos estructurados como los no estructurados procesados por las herramientas ETL.

En la Figura 2 se muestran las principales características que deben cumplirse en el nivel de almacenamiento de datos de la arquitectura propuesta, es decir, cualquier tecnología que se utilice para la implementación del nivel de almacenamiento de datos debe poder cumplirlas. Estas son:

– Replicación: permite la redundancia de información en las bases de datos, con lo que, si una base de datos deja de funcionar, la información se asegura pues se encuentra replicada en otras bases de datos.
– Balanceo de carga: realiza la adecuada distribución de los datos en múltiples servidores.
– Escalabilidad horizontal: permite que los datos se puedan almacenar en varios servidores. A mayor cantidad de información, más servidores se emplearán.
– Sistemas distribuidos de ficheros: opera con una red o clúster de servidores interconectados y configurados para trabajar con un sistema de ficheros lógico. El tamaño del sistema de ficheros puede variar, aumentar o disminuir, de acuerdo con las necesidades y sin afectar el rendimiento general del sistema.
– *Sandboxing* o establecimiento de almacenes de datos temporales: técnica que permite la creación de almacenes de datos temporales para la experimentación, procesamiento y análisis de datos. Los datos que contienen son copiados desde la fuente donde se encuentran almacenados y se puede escoger libremente cómo se van a tratar los mismos y qué hacer con ellos.
– Presencia de filtros de datos: permiten obtener datos específicos que se desean tratar o asegurar en el sistema de almacenamiento.

Uno de los dilemas a los que se enfrentan muchas empresas es que no pueden costearse la infraestructura física necesaria para almacenar grandes volúmenes de datos no estructurados. En la actualidad, muchos proveedores de almacenamiento de datos ofrecen soluciones para la nube como parte de su gama de productos y las comercializan entre los clientes como soluciones asequibles y accesibles.

El almacenamiento en la nube permite que solo se necesite alquilar potentes servidores equipados con sofisticadas aplicaciones diseñadas especialmente para manejar grandes volúmenes de datos, a los que pueden acceder permanentemente. Los clientes de estas soluciones obtienen rápidos resultados. Existen varias ventajas del uso de la nube en entornos Big Data, entre ellas la escalabilidad, la elasticidad, la creación eficiente de recursos compartidos, la reducción de costos y la tolerancia a fallos [8].

Nivel de análisis de datos de la arquitectura propuesta

Teóricamente, una de las ventajas de Big Data es que mientras más datos se analicen, mayor será la amplitud de visiones que se puedan establecer en torno a los objetivos que se persigan con su empleo por parte de una organización o empresa. La primera pregunta que hay que formularse para seleccionar las herramientas de análisis de datos es: ¿qué o cuáles problemas se están tratando de resolver y en qué sector o área de la sociedad se encuentran enfrascados? Además, en la selección de herramientas de análisis también se debe tener en cuenta el nivel de complejidad del problema a resolver.

En la Tabla 1 se recoge el uso de algunas herramientas analíticas para varios tipos de análisis a realizar. Para el nivel de análisis de datos de la arquitectura referencial para la gestión de redes y servicios de telecomunicaciones que se propone solo son de interés los dos primeros tipos de análisis.

- Analíticas básicas: son utilizadas para explorar grandes volúmenes de datos. Permiten la división de enormes volúmenes de datos en pequeños grupos que son más fáciles de explorar, monitorear en tiempo real permitiendo identificar en ellos anomalías e incidentes. En los procesos de la gestión de las redes y servicios de telecomunicaciones estas herramientas son de gran importancia para: el monitoreo del desempeño de los dispositivos y los servicios de la red, la detección de anomalías y la visualización de las configuraciones.
- Analíticas avanzadas: proveen algoritmos para análisis complejos de distintos tipos de datos, permitiendo la obtención de patrones de los mismos, la predicción y prevención de eventos y el procesamiento de datos complejos. Estas herramientas analíticas se pueden utilizar en el análisis de textos para extraer información valiosa o en el desarrollo de algoritmos y modelos predictivos que ayuden a la minería de datos. Es decir, se emplean en la obtención de mayor cantidad de información sobre los servicios o aplicaciones que se desean suministrar.

Tabla 1. Casos de uso de las herramientas analíticas [9].

Tipo de análisis	Uso
Analíticas básicas	Selección y división de datos, reporte, visualizaciones simples y monitoreo básico
Analíticas avanzadas	Análisis complejos como modelos predictivos o técnicas de establecimiento de patrones
Analíticas operacionalizadas	Análisis de procesos de negocios
Analíticas monetizadas	Análisis monetarios

En la Figura 2 se muestran los principales procesos que deben realizarse en la capa de análisis de datos de la arquitectura propuesta. Estos son:

- Proceso de deducción de valor: determina qué información es relevante a los objetivos que se persiguen.
- Proceso de selección y división de datos: se seleccionan y se dividen los altos volúmenes de datos en pequeños grupos para su análisis más óptimo.
- Proceso de determinación de patrones: establece interrelaciones entre los grupos de datos.
- Proceso de visualización: muestra en una interfaz gráfica el análisis que se realiza de los datos.
- Procesos de diagnósticos y reportes: permite obtener información útil resultante del análisis de los datos, aplicable a los objetivos que se persiguen.

– Proceso de monitoreo: proceso que permite conocer el desempeño y comportamiento del sistema.

En este punto se debe precisar que los procesos de los niveles de extracción, transformación y carga, almacenamiento de datos y análisis de datos de la arquitectura propuesta son similares para cualquier aplicación de Big Data en la gestión de las telecomunicaciones. Los procesos del nivel de gestión de redes y servicios de las telecomunicaciones dependerán de los objetivos específicos de gestión de cada empresa u organización.

Nivel de gestión de redes y servicios de la arquitectura propuesta

La gestión de redes está compuesta por cinco áreas funcionales de la gestión: configuración, desempeño o prestaciones, fallos, seguridad y contabilidad. Algunos casos del empleo de Big Data en las áreas funcionales de la gestión de redes se explican en esta sección.

La obtención de información sobre las redes y sus servicios se ha visto beneficiada con el surgimiento de Big Data, que ha incorporado un amplio número de herramientas y oportunidades para el tratamiento de grandes cantidades de datos, estructurados y no estructurados. Las herramientas analíticas de datos masivos pueden ser empleadas en el análisis de transacciones financieras, archivos de *logs*, tráfico de las redes lo que permite identificar anomalías y actividades sospechosas y correlacionar coherentemente múltiples fuentes de datos.

Uno de los usos de las herramientas analíticas es en la gestión de la seguridad de las redes. En un caso de estudio publicado, Zions Bancorporation anunció el empleo de Hadoop [10] y herramientas de análisis inteligente que permiten trabajar con gran cantidad de datos en menor tiempo que con las herramientas de análisis de datos tradicionales (se empleó de 20 minutos a una hora para analizar grandes volúmenes de datos utilizando las herramientas tradicionales, mientras que con Hadoop se realizó en un minuto aproximadamente). Se aumentó la seguridad de sus redes gracias al análisis efectivo de información proveniente de distintas fuentes como *firewalls*, dispositivos de redes, tráfico por la red, procesos de negocio y transacciones diarias.

Otro ejemplo de cómo Big Data puede ser empleado en la gestión de la seguridad de las redes es el trabajo realizado por HP Labs para identificar dispositivos infectados con *malware* (tipo de *software* malintencionado que daña los dispositivos) en las redes empresariales. Para ello se tomó millones de datos de solicitudes del protocolo de transferencia de hipertexto (HTTP: de sus siglas del término en inglés *Hypertext Transfer Protocol*), del sistema de nombres de dominio (DNS: por sus siglas del término en inglés *Domain Name System*) y de los sistemas de alerta de intrusos [11].

Con la evolución de Big Data se han logrado establecer mejores estrategias y métodos en la detección de amenazas persistentes avanzadas (APT: por sus siglas del término en inglés *Advanced Persistent Threat*) [12], que es uno de los problemas más serios que enfrentan las empresas y organizaciones en cuanto a la seguridad de la información. Las APT, en contraste a otros tipos de *malware* como los troyanos y los gusanos, son agresores de las redes que trabajan en modo "*low-and-slow*", es decir, *low* pues mantienen un perfil bajo en la red haciendo muy difícil su detección y *slow* porque están activos durante un largo período. La detección de estas amenazas se basaba en la experiencia humana, lo que provocaba que fuera una labor intensiva, difícil de generalizar y no escalable. Con el empleo de arquitecturas más escalables y de mayor nivel de procesamiento en la detección APT, el análisis de grandes grupos de datos ya no constituye un desafío, pues se ha logrado establecer un método que emplea algoritmos de monitoreo que permite determinar prácticamente todos los posibles ataques a las redes.

Diversos estudios se refieren al empleo de sistemas Big Data en la mejora de la calidad de experiencia de los usuarios. Para ello, se utilizan técnicas de extracción y análisis de datos actualizados sobre las opiniones de millones de usuarios de diversos servicios, lo que permite que las empresas conozcan cómo sus servicios son aceptados y responder a las nuevas necesidades de los clientes. Otra forma de mejorar la QoE, es optimizar la calidad de servicio (QoS: por sus siglas del término en inglés *Quality of Service*) que brindan las redes y los servicios que se ofrecen, puesto que ambos conceptos se encuentran estrechamente relacionados.

Además, se han trazado diversos acercamientos a la gestión de las redes en los nuevos entornos Big Data, entre ellos la gestión de la red basada en el valor (VBNM: por sus siglas del término en inglés *Value-Based Network Management*). La VBNM se basa en el análisis del comportamiento de los clientes y del consumo de los recursos de la red por parte de estos. También se basa en la extracción de información de la red, o sea, no solo considera los datos que circulan por ella, sino que tiene en cuenta la información brindada por los dispositivos de la misma, para lograr disminuir el consumo de recursos y el tiempo de retardo de la información en la red, aumentar la eficiencia de los dispositivos, mejorar la configuración de los mismos, disminuir la congestión de la red y reubicar los recursos disponibles donde su utilización sea más productiva [13].

IBM (*International Business Machines*, por sus siglas en inglés), una de las empresas más destacadas en tecnología y consultoría, opina que Big Data está hecho para la industria de las telecomunicaciones. Gracias al desarrollo de las redes y la proliferación de dispositivos inteligentes, los proveedores de servicios de telecomunicaciones tienen acceso a un gran cúmulo de información sobre los comportamientos y las preferencias de sus clientes. Además, actualmente, a nivel internacional, muchas empresas que brindan servicios de telecomunicaciones se encuentran enfrascadas en el desarrollo de alternativas para emplear Big Data en su gestión [14], siendo esta una de las razones por la que se considera de vital importancia la investigación que en este artículo se presenta.

Aplicación de la arquitectura referencial de Big Data propuesta en un caso de estudio enfocado en la gestión de la seguridad una red

En esta sección se muestra la aplicación de la arquitectura propuesta a la gestión de seguridad en una red y en particular en la detección de intrusiones en la red. La seguridad es uno de los aspectos más importantes a considerar dentro de cualquier ambiente o entorno de red, siendo este uno de los principales retos en el empleo de la tecnología Big Data.

Los principales puntos de interés para aplicar la propuesta de arquitectura definida a la gestión de una red son:

– Definir las principales fuentes de dónde serán extraídos y almacenados los datos.
– Determinar las herramientas necesarias para la extracción, transformación y carga de los datos, desde las fuentes que los generan hasta los sistemas de almacenamiento de datos.
– Definir un sistema de detección de intrusiones basado en herramientas de gestión de seguridad de la red y entornos de gestión de datos masivos para la correcta detección de anomalías en la red.
– Establecer un sistema de almacenamiento para almacenar los datos capturados.
– Definir los procesos de análisis y las herramientas necesarias para ejecutarlos. Esto permite obtener información aplicable a la optimización de la gestión de la seguridad de la red.

Para el establecimiento del sistema de detección de intrusiones en la red se empleará a Snort [15] y a Hadoop. Snort es un analizador de paquetes y detector de intrusos (IDS: por sus siglas del término en inglés *Intrusion Detection System*) que ofrece capacidades de almacenamiento tanto en archivos de texto como en bases de datos *open source*. Implementa un motor de detección de ataques y monitoreo de puertos que registra, alerta y responde a las anomalías previamente definidas. Posibilita, entre otras funciones, la observación del funcionamiento de la red y el tráfico en la misma en tiempo real.

Por su parte, Hadoop es una herramienta de código abierto con un alto desempeño en el procesamiento de datos masivos, la que fue seleccionada para la extracción, transformación y carga de los datos que captura Snort. Cuenta con distintos componentes que se encargan de la transformación, almacenamiento y carga de datos no estructurados, que en muchas herramientas de procesamiento de datos no existen, permitiendo la extracción masiva de datos en cuestiones de segundos.

La mayoría de los sistemas de detección de intrusos identifican rápidamente ataques a partir de una serie de reglas. Los paquetes entrantes son analizados y comparados con las reglas definidas y si uno de los paquetes no cumple con las reglas establecidas, entonces acciones especificadas se realizarán. Es obvio que a mayor cantidad de reglas que se definan, mayor número de amenazas se podrán identificar. La mayor desventaja de los sistemas de detección de intrusos es que no son capaces de identificar

ataques desconocidos, es decir, distintos eventos que no se encuentran definidos en sus reglas.

La principal fuente de la cual se extraerán los datos hacia Hadoop será de Snort. Dicha herramienta presenta varios modos de ejecución. Uno de estos modos es el *Packet Logger*, en el que Snort analiza el tráfico de la red, captura los paquetes de interés y los almacena temporalmente en el HDFS (sistema distribuido de archivos de Hadoop, por sus siglas del término en inglés *Hadoop Distributed File System*).

Los principales datos de interés serán los paquetes que circulan desde o hacia los distintos nodos que se encuentran distribuidos en la red. Estos datos son generados por los dispositivos de la red, los usuarios internos y externos de la red, las aplicaciones, etc. Existen ataques que se caracterizan por el envío de un gran número de paquetes hacia un dispositivo como son los ataques del protocolo de mensajes de control de internet (ICMP: por sus siglas del término en inglés *Internet Control Message Protocol*), los *pings* de la muerte, los ataques *smurf*, los ataques del protocolo de datagramas de usuario (UDP: por sus siglas del término en inglés *User Datagram Protocol*), entre otros. En la detección de este tipo de amenazas se centra principalmente el caso de uso, ver Figura 3.

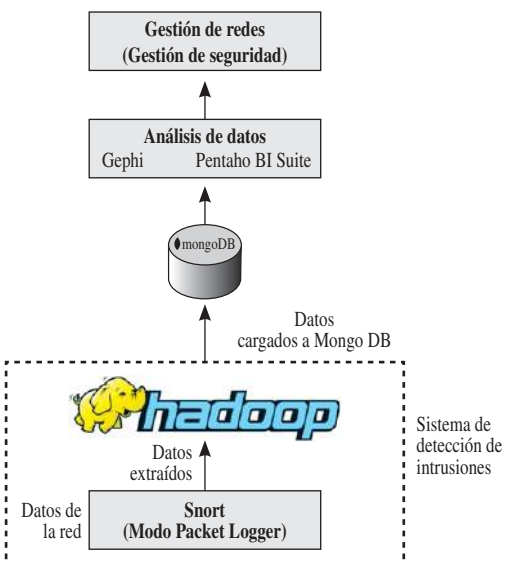

Figura 3. Arquitectura del sistema de detección de intrusiones.

A medida que Hadoop vaya extrayendo los datos de Snort, será capaz de identificar y clasificar desde y hacia donde están dirigidos los paquetes y determinar la cantidad que estos son, empleando las banderas y los campos del protocolo IP de los paquetes (IP: de sus siglas del término en inglés *Internet Protocol*) como la dirección fuente y destino y el número de puerto. El procesamiento de los datos dentro de Hadoop se muestra en la Figura 4.

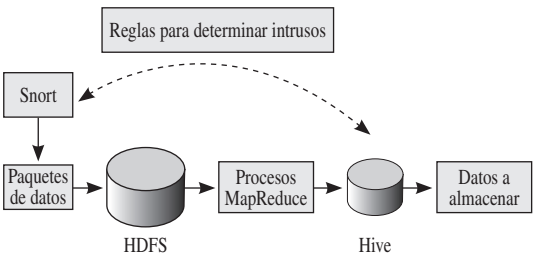

Figura 4. Procesamiento de los datos en Hadoop.

Además, Hadoop realizará procesos de mapeo y reducción para reducir los paquetes que presentan la misma información, además de entregar la información en un formato adecuado para su posterior análisis. Los datos serán enviados a Hive [9], componente de Hadoop, el que a partir de las reglas implementadas en Snort, será capaz de identificar cuando existe un ataque hacia un nodo o dispositivo, logeando los paquetes identificados para reconocer que son un ataque. Las reglas definidas en Snort Hive las tendrá implementadas en su lenguaje de consulta HiveQL. Además, dada la facilidad del lenguaje de programación de Hive, permite elaborar nuevas reglas que pueden ser añadidas a Snort para la detección de nuevas anomalías.

La principal ventaja que presenta la integración de Hadoop y Snort es que se pueden establecer las reglas en Snort previo y posterior al análisis de los datos. Generalmente, Snort analiza los paquetes para determinar posibles amenazas, guiándose por las reglas que previamente fueron establecidas. Esto puede generar un alto número de falsas alarmas ante la llegada de paquetes desconocidos. Con el empleo de Hadoop esto puede mejorarse, ya que se pueden ir analizando los paquetes en Hadoop y elaborarse nuevas reglas de acuerdo con los datos encontrados que representan amenazas, los que si no fueron identificados por Snort, mediante Hive se detectarán.

Además, en la implementación de la arquitectura propuesta se recomienda utilizar como base de datos a MongoDB [16], la que es una base de datos no relacional orientada a documentos y ampliamente utilizada a nivel internacional. La idea principal es que después de analizados los datos en Hadoop, estos sean almacenados en un sistema que sea escalable y altamente disponible para su uso posterior. Es importante precisar que en MongoDB se almacenarán todos los datos que sean necesarios para la gestión de la red, no solo los requeridos para la gestión de seguridad.

Las dos herramientas seleccionadas para el análisis de los datos almacenados en MongoDB fueron: Pentaho BI Suite [17] y Gephi [18]. Pentaho constituye un conjunto de programas *open source* para generar inteligencia en los negocios y posee una *Web* organizada en productos o componentes de reporte, análisis, minería de datos y *dashboards* y es altamente utilizado para el acceso, integración, visualización, exploración y minería de todo tipo de datos que puedan impactar en los negocios. Pentaho fue seleccionada para la aplicación de la arquitectura propuesta ya que soporta los principales procesos de análisis de datos que se desean implementar: minería de datos y análisis predictivo de fallas y amenazas.

Gephi se seleccionó principalmente por la alta capacidad de visualización de redes que provee. Es una herramienta *open source*, creada para facilitar que el usuario explore la red, la visualice y realice análisis en tiempo real. Además, por sus características es altamente aplicable a la gestión de los servicios de una red.

El empleo de estas herramientas de análisis representa grandes ventajas para la gestión de la seguridad de una red y, en particular, para la detección de intrusiones. Primeramente, mediante Gephi se puede realizar un esquema que visualice la red y sus elementos, lo que ayuda a determinar las principales zonas de riesgo, los elementos de la red más vulnerables y dónde han ocurrido mayor cantidad de amenazas. Esto facilita que se puedan llevar a cabo las acciones necesarias para optimizar la detección de intrusiones.

Con el empleo de Pentaho se pueden desarrollar diagramas donde se muestren los resultados del análisis del tráfico de la red, determinándose aquellos parámetros que más influyan en la optimización de la seguridad de la red y la gestión de la misma. También, mediante Pentaho, se puede trabajar en el análisis predictivo de amenazas a partir de los datos que se encuentran almacenados en MongoDB y de los obtenidos en tiempo real.

Aplicación de la arquitectura referencial de Big Data propuesta en un caso práctico implementado en una empresa

Posteriormente a la elaboración de la arquitectura propuesta, se desarrolló un método para la implementación de Big Data en las empresas, en el que se describen los pasos o procesos a desarrollar para garantizar el correcto establecimiento de la tecnología, de acuerdo con los distintos niveles de la arquitectura. De esta forma, se definieron los principales procesos a implementar en cada uno de los niveles de la arquitectura y se estableció un análisis de las distintas herramientas de código abierto existentes en el mercado, de acuerdo al nivel de potencialidades que ofrecen para cada caso en que se apliquen [19].

Se decidió entonces implementar la arquitectura y el método en una empresa dedicada a ofrecer servicios de comunicaciones, informática y automática [20], la que posee una estructura sólida y escalable. Se trabajó en la optimización de la seguridad de la red, recolectando flujos y capturando tráfico de la red en tiempo real, permitiendo analizar las potencialidades que nos ofrecen las herramientas de entornos Big Data con respecto a las tradicionales. Primeramente, se realizó un estudio de la red, el que estuvo enfocado en los distintos servicios que en la red de la empresa se ofrecen, la capacidad de procesar datos, los dispositivos de borde de la estructura del centro de datos, la cantidad de usuarios de la red y la cantidad de datos que se manejan en tiempo real. Mediante este estudio se pudo definir cuáles eran los principales procesos a definir de acuerdo a estas características de la empresa [20].

Las pruebas realizadas en la empresa fueron:

- La recolección de datos por lotes [21], análisis reactivo [9] empleando herramientas de *warehouse* y visualización de datos.
- Almacenamiento de datos.

– Recolección de datos en *streaming* [21], análisis proactivo [9] y aplicaciones de seguridad y desempeño.

Se trabajó con dos herramientas principales: los *frameworks* Hadoop y Big Data Storm [22], debido al gran número de herramientas de recolección y análisis de datos que presentan. Para la recolección de datos se empleó Flume [10], herramienta que permite la recolección de datos en lotes o en *streaming* y para el análisis reactivo se empleó Pig [10], mediante el que se desarrollaron *scripts* para el análisis de datos, siendo una herramienta sencilla y fácilmente entendible, donde se requieren pocas líneas de código. Además, para optimizar el análisis de datos se empleó Mahout [10], permitiendo la búsqueda de patrones entre datos, procesos de clasificación, filtros colaborativos, entre otras funcionalidades. Para la captura de datos se empleó la herramienta Wireshark y para el acceso vía *Web* a las distintas herramientas Google Chrome. Por último, para el almacenamiento de datos se empleó MongoDB y como herramienta de *warehouse* Hive.

Se instalaron dos máquinas virtuales, empleándose el software VMware Player 12.1.0. En la primera máquina virtual se instalaron Hadoop y Big Data Storm (Flume, MongoDB, Pig, Mahout, Hive), mientras que en la segunda Wireshark y Google Chrome. Mediante los *scripts* desarrollados en Pig y empleando Flume se pudo trabajar en la determinación de la capacidad necesaria de almacenamiento de los *logs* de la empresa y, posteriormente, realizar los análisis de los logs almacenados. Además, el *script* realizado permite filtrar una información específica, entre miles de *bytes* de datos en cuestiones de segundos.

Además, se realizó un proceso de capturas de paquetes de acuerdo a distintos puertos y direcciones IP, buscando analizar el tráfico de la datos por la red y los principales protocolos de red empleados. Lo anterior permite, desde las correctas líneas de código, buscar la existencia de ataques en la red. En el diagrama de la Figura 5 se observa la comparación entre la detección de ataques antes y después de la implementación de las herramientas de datos masivos. Es importante destacar que lo anterior corrobora la importancia de la aplicación de Big Data a la gestión de una red, en este caso, a la gestión de la seguridad.

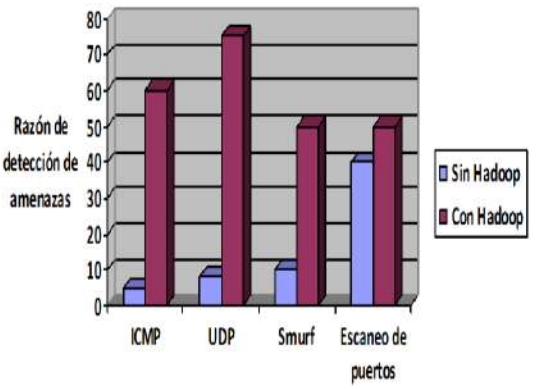

Figura 5. Comparación en la detección de amenazas en la red mediante Hadoop.

CONCLUSIONES

Como principales conclusiones de esta investigación se puede plantear que:

– La aplicación de las tecnologías Big Data en la gestión de las telecomunicaciones proporcionan la optimización y desarrollo de sus redes y servicios.
– La arquitectura referencial de Big Data para la gestión de las telecomunicaciones propuesta puede ser aplicada para la gestión de redes, servicios y aplicaciones de distintos escenarios, contando con la tecnología adecuada para la implementación de la misma.
– La integración de Snort y Hadoop para el establecimiento de un sistema de detección de intrusos permite disminuir las falsas alarmas generadas.
– Es importante destacar que mediante los casos de uso se pudo aplicar la arquitectura referencial de Big Data para la gestión de las telecomunicaciones, en un caso de estudio y en uno práctico que se puede implementar para la gestión de seguridad de una red, con lo que queda demostrado que la propuesta es aplicable a la gestión de redes y puede ser implementada con tecnologías existentes, las que son de código abierto y *software* libre.

REFERENCIAS

[1] J. Hurwitz. "Big Data for Dummies". John Wiley & Sons. New Jersey, Estados Unidos. 2013.

[2] P. Chandarana and M. Vijayalakshmi. "Big Data analytics frameworks". 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), pp. 430-434. Abril 2014. DOI: 10.1109/CSCITA.2014.6839299.

[3] Z. Liu, C. Hu, Y. Li and J. Hu. "DSDC: A Domain Scientific Data Cloud Based on Virtual Dataspaces". 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops PhD Forum (IPDPSW), pp. 2176-2182. 2012. DOI: 10.1109/IPDPSW.2012.269.

[4] R. Liu, Q. Li, F. Li, L. Mei and J. Lee. "Big Data architecture for IT incident management". 2014 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), pp. 424-429. 2014.

[5] M. Kaur. "Big Data Visualization Tool with Advancement of Challenges". International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 4 Nº 3, pp. 665-668. 2014. ISSN: 2277-128X.

[6] Z. Liu, P. Yang and L. Zhang. "A Sketch of Big Data Technologies". Seventh International Conference on Internet Computing for Engineering and Science (ICICSE), pp. 26-29. 2013.

[7] D. Kasibhotla. "Introduction to Massively Parallel Processing (MPP) database". URL: https://dwarehouse.wordpress.com/2012/12/28/ introduction -to-massively-parallel-processing-mpp-database. Fecha de Consulta: 4 de febrero de 2015.

[8] J.B. Clark. "Standars and Big Data. Big Data in the Cloud: Preparing for the Future". 2013.

[9] B. Schoenborn, Big Data Analytics Infrastructure for Dummies: John Wiley & Sons, Inc., 2014.

[10] T. White. "Hadoop: The Definitive Guide". O'Reilly Media, Inc. 2da Edición. Estados Unidos, pp. 625. 2010. ISBN: 978-1-449-38973-4.

[11] Big Data Working Group. Big Data Analytics for Security Intelligence. Cloud Security Alliance White Paper. 2013.

[12] P. Giura y W. Wang. "Using Large Scale Distributed Computing to Unveil Advanced Persistent Threats", pp. 13, 2012.

[13] J. Arias. "Value-Based Network Management for Telecoms". 2015, pp. 24.

[14] B. Fox, R. Dam y R. Shockley. "Analytics: El uso de Big Data en el mundo real aplicado a las telecomunicaciones". IBM Global Business Services, Business Analytics and Optimization, , pp. 20. 2013.

[15] P. G. Prathibha y E. D. Dileesh. "Design of a Hybrid Intrusion Detection System using Snort and Hadoop". International Journal of Computer Applications. Vol. 73 Nº 10. Julio 2013.

[16] MongoDB Documentation Release 3.0.0. 2015.

[17] N. Goodman. "Pentaho Data Integration". Bayon Technologies White Paper. 2009.

[18] C. B. Amat. "Análisis de redes y visualización con Gephi". REDES - Revista hispana para el análisis de redes sociales. 2014.

[19] A. Cordero García. "Método para la implementación de Big Data en la Gestión de las Telecomunicaciones". Tesis de Diploma en opción al Título de Ingeniero en Telecomunicaciones y Electrónica, Instituto Superior Politécnico José Antonio Echeverría, 2016.

[20] Tecnomática. Catálogo para productos y servicios, pp. 10. 2013.

[21] S. Siddiqui. "Big Data Implementation and Visualization". IEEE International Conference on Advances in Engineering & Technology Research (ICAETR-2014), pp. 10. 2014.

[22] Z. Chena, N. Chena and J. Gonga. "Design and implementation of the real time GIS data model and Sensor Web service platform for environmental big data management with the Apache Storm". Collaborative Innovation Center of Geospatial Technology, pp. 4. 2015.