USO DE ANALITICAS PARA PREDECIR LOS COMPUTADORES AFECTADOS POR MALWARE, EN UNA INSTITUCIÓN FINANCIERA EN COLOMBIA, 2017

Autor

GERARDO MAYORGA GARCIA

ESCUELA COLOMBINA DE INGENIERIA

MAESTRIA DE GESTION DE INFORMACIÓN

Bogotá. Febrero de 2017

La información y opiniones enunciadas en este proyecto, son responsabilidad del autor dentro del marco de un ejercicio académico y no comprometen ni reflejan las opiniones de la Institución Financiera, en donde se realizó la prueba de concepto, objeto de este proyecto.

Gerardo Mayorga García

gerardo.mayorga@mail.escuelaing.edu.co

gmayorga@banrep.gov.co

Tabla de contenido

LISTA DE FIGURAS 4

RESUMEN 6

1. PLANTEAMIENTO DEL PROBLEMA 8

1.1. Enunciado del problema 8

1.2. Formulación del problema 10

2. OBJETIVOS 12

2.1. Objetivo General 12

2.2. Objetivos Específicos 12

3. JUSTIFICACION Y DELIMITACION 12

4. MARCO TEORICO 14

5. HIPOTESIS 46

6. PRUEBA DE CONCEPTO 50

7. PRESUPUESTO DE INVERSION 109

8. PRODUCTOS GENERADOS 110

9. APORTES DEL AUTOR 110

10. TRABAJOS FUTUROS 114

11. CONCLUSIONES 117

EQUIPO INVESTIGADOR 120

REFERENCIAS BIBLIOGRAFICAS 121

LISTA DE FIGURAS

Figura 1 Tiempo transcurrido en detectar la infección 8
Figura 2 - Brecha de Seguridad Informática 16
Figura 3 Ataque informáticos más efectivos 21
Figura 4 Diferencias entre Business Intelligence vs Big Data - Analytics 28
Figura 5 Servicios en la nube 33
Figura 6 Trabajos realizados con servicios en la nube 37
Figura 7 Tipos de Análisis 43
Figura 8 Tipos de Análisis para responder la pregunta 44
Figura 9 Eventos semanales y representativos de seguridad informática 51
Figura 10 Ecosistema de Big Data 54
Figura 11 Análisis de Gartner 58
Figura 12 Análisis de Forrester 59
Figura 13 Componentes de Analytics y Big Data 62
Figura 14 Componentes estándar de la arquitectura de la POC 63
Figura 15 Arquitectura de la POC – Institución Financiera 64
Figura 16 Fuente de datos del antivirus 66
Figura 17 Calidad de datos – Modelo ETL 67
Figura 18 Modelo de Analytics - ARIMA 84
Figura 19 Resultado del modelo ARIMA 92
Figura 20 Predicción de Virus y Troyanos 95

Figura 21 Comparación del modelo predictivo vs la realidad 96
Figura 22 Predicción Acertada, para número de eventos 99
Figura 23 Ocurrencia Acertada, para número de eventos..... 100
Figura 24 Predicción Desfasada, para el número de eventos..... 101
Figura 25 Ocurrencia Desfasada, para el número de eventos 101
Figura 26 Configuración del reporte para verificar la no presencia de virus..... 103
Figura 27 Verificación que no continuaron eventos de infección..... 104
Figura 28 Cronograma 108
Figura 29 Presupuesto 109
Figura 30 Propuesta: Adicionar fuentes de datos al modelo 115

RESUMEN

Las Sociedades de la Información y del Conocimiento se encuentran soportados por datos informáticos que son considerados activos estratégicos en todas las empresas a nivel mundial. Los datos representan dinero y por tal motivo es necesario brindar diferentes niveles de protección; uno de estos mecanismos son las herramientas de seguridad informáticas diseñadas para ofrecer la seguridad digital contra diferentes tipos de ataques informáticos, siendo el más representativo el malware o software maligno, término utilizado para hacer referencia cualquier programa informático diseñado para realizar acciones no autorizados con la finalidad de cometer un fraude informático, a través de diferentes vectores de ataques.

A pesar que las organizaciones disponen de herramientas tradicionales de seguridad informáticas, en algunos casos las funcionalidades no son efectivas debido a que los ataques informáticos son más sofisticados con el pasar de los días, permanentemente se generan datos estructurados y no estructurados que crecen de forma incremental, intentando comprometer los sistemas informáticos en el momento menos esperado, no se presenta un patrón de comportamiento, las intromisiones se pueden presentar períodos cortos de tiempo o a largo plazo para intentar pasar desapercibidos y estudiar el comportamiento del usuario. Estas herramientas de seguridad enfocan sus funcionalidades en realizar análisis para la detección y diagnóstico.

Con los avances de las Tecnologías de la Información y la Comunicación es posible realizar nuevos tipos de análisis, como: predictivos, prescriptivos y preventivos, a partir del examen de enormes cantidades datos, correlacionando diferentes tipos de datos estructurados y no estructurados, provenientes de diversas fuentes internas de las empresas o fuentes externas como las redes sociales, sensores, etc. Adicionalmente, es viable la construcción de prototipos en

herramientas de software, que son ejecutados en modelos estadísticos, permitiendo realizar predicciones con un alto grado de probabilidad con respecto a la realidad.

Los resultados de los nuevos análisis permiten apoyar las labores de gestión de las soluciones anti-malware y proporcionan herramientas para que los dueños de los procesos, enfoquen de una mejor forma su estrategia de negocio.

Una característica importante radica en que luego de contar con el software, hardware y conocimientos necesarios, relacionados con nuevos conceptos y tecnologías de punta, es posible realizar nuevos análisis para diferentes herramientas de seguridad informática, así como para otras áreas o procesos dentro de las empresas.

Términos clave

Seguridad informática, malware, datos estructurados, datos no estructurados, Prueba de Concepto, “Analytics”, Big Data, Business Intelligence, Modelo Arima

1. PLANTEAMIENTO DEL PROBLEMA

1.1. Enunciado del problema

De acuerdo al estudio realizado por la empresa Verizon, el intervalo de tiempo ocurrido desde el momento en que se compromete de forma negativa un activo informático hasta que la víctima se entera del incidente son muy largos. La siguiente estadística muestra el porcentaje de las empresas con sus respectivos intervalos de tiempo: “el 27% se demora días, el 24 % se demora semanas, el 39 % se demora meses y el 9% tarda años” (Verizon, Data Breach Investigations Report, 2012)

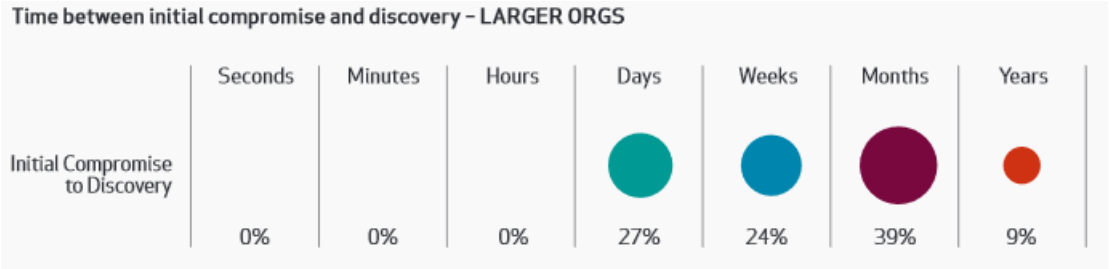


Figura 1 Tiempo transcurrido en detectar la infección

A partir del momento en que se materializa la infección por malware, inmediatamente aparece una serie de consecuencias negativas para la organización, algunas de ellas son: los trabajadores no tienen las condiciones para poder cumplir sus funciones asignadas, existen actividades adicionales por parte del personal de la mesa de ayuda para solucionar el problema, en las organizaciones afectadas el principal problema es la pérdida de credibilidad, de igual forma es muy probable el incumplimiento de compromisos establecidos con anterioridad, etc.

Diversas estadísticas muestran las grandes *pérdidas económicas* para ciudadanos, así como para las grandes empresas, a nivel mundial “El costo de la

delincuencia cibernética para la economía mundial se ha estimado en US$ 445 billones (£ 266 billones) al año" (Williams, 2014), panorama que se repite en menor proporción en Colombia y de acuerdo a lo publicado que "El cibercrimen representa el 15 por ciento de los ilícitos cometidos a empresas en Colombia y generó un daño económico cercano a 600 millones de dólares en el último año" (Tecnósfera, 2016).

En todas las instituciones financieras a nivel mundial, el malware causa grandes pérdidas económicas por el traslado no autorizado de grandes cantidades de dinero hacia cuentas que no corresponden, cometiendo fraudes informáticos. También se presenta la fuga de información de datos confidenciales, presentado un grave problema debido a que se está incurriendo en faltas graves a la ley de protección de datos personales, habeas data, circular 052, entre otras.

Adicionalmente, si el malware logra dañar los datos alojados en las bases de datos de las aplicaciones corporativas, es posible incumplir con las operaciones propias del negocio, como las operaciones internacionales, las subastas, compensación de cheques, etc.

Es difícil medir las consecuencias por un ataque de virus informáticos, pero en términos generales se pueden apreciar las siguientes valoraciones brindadas por McAfee (IntelSecurity, 2014), son las siguientes: *"La pérdida de la confianza del cliente (62%) fue la consecuencia más reconocida de un fallo de seguridad, seguido por el daño a la marca y la reputación de la empresa (52%), desafíos regulatorios (41%), pérdida financiera a través de la pérdida de clientes y multas (40%) y la reducción de la confianza en la seguridad de los empleados (36%)"*. El anterior estudio fue realizado a 500 personas con las funciones de tomar decisiones en el sector de la tecnología en empresas representativas, de la siguiente forma: 200 en Estados Unidos, 100 en el Reino Unido, 100 en Alemania y 100 en Australia.

Adicionalmente, existe un “gap” o brecha de seguridad informática, que corresponde al espacio comprendido entre los avances constantes de los proveedores de herramientas de seguridad informática versus los avances incrementales de las amenazas informáticas especializadas, nuevos elementos de cumplimiento para las organizaciones y las herramientas de siguiente generación.

De igual forma, nos encontramos en un mundo cambiante y la tecnología avanza a grandes pasos para brindar soluciones a las nuevas necesidades de los usuarios y por tal razón los fabricantes diseñan las tecnologías de siguiente generación, caracterizada por conceptos como: “La Nube”, “Movilidad”, “Aplicaciones - Apps”, “Big Data”, “Business Intelligence”, “Datamining”, etc. Términos que serán explicados más adelante.

1.2. Formulación del problema

En virtud de lo anterior, el presente trabajo de grado se orienta a revisar la documentación existente con conceptos y tecnologías, que permitan ayudar a la gestión de amenazas informáticas originadas por malware, aprovechando las funcionalidades para realizar análisis predictivos de grandes volúmenes de datos y obtener respuestas casi en tiempo real; por lo tanto, el estudio busca responder el siguiente interrogante:

¿Es posible realizar análisis predictivos, con el apoyo de un modelo estadístico, para detectar posiblemente, cuáles computadores estarán comprometidos por malware?

De igual forma este documento busca responder a las siguientes preguntas específicas:

Primera, debido a que los análisis tradicionales para la detección y diagnóstico de malware no son efectivos, ¿existen otros tipos de análisis que permiten predecir el evento de infección por software maligno?

Segunda, como los ataques informáticos se presentan de diversas formas y en un momento inesperado, ¿se encuentra implementado un modelo estadístico en una herramienta informática, que permita realizar el pronóstico de ataques por malware?

Tercera, ¿cuáles son las características tecnológicas y conocimientos necesarios para realizar análisis predictivos?

Por último, ¿con la implementación de esta propuesta, es posible apoyar el proceso de gestión de malware para el manejo de ciberataques en una Institución Financiera?

2. OBJETIVOS

2.1. Objetivo General

Predecir mediante el uso de herramientas analíticas tecnológicas y en corto plazo de tiempo, los posibles computadores que pueden ser “infectados” por software maligno.

2.2. Objetivos Específicos

Investigar los nuevos conceptos y herramientas informáticas disponibles en el mercado, con funcionalidades de análisis tecnológicos de datos.

Identificar un modelo estadístico que permita establecer patrones para una predicción, con el fin de identificar los posibles computadores que serán atacados por malware.

Identificar los componentes tecnológicos y conocimientos requeridos para realizar una prueba de concepto, que permita el procesamiento de grandes volúmenes de datos, con tiempos reducidos de respuesta casi en tiempo real.

Presentar esta propuesta al Área de Seguridad Informática, de la Institución Financiera, para ayudar a la gestión de riesgos informáticos originados por software maligno.

3. JUSTIFICACION Y DELIMITACION

Este estudio se justifica por las siguientes razones: La primera, se debe a la preocupación existente en las Instituciones Financieras, debido a que los datos informáticos, considerados como un activo organizacional, se encuentran expuestos

a riesgos informáticos causados por la materialización de ataques por software maligno.

En segundo lugar, desde la aparición del primer virus informático en el año de 1949, en los laboratorios Bell Computer, éstos han ido evolucionando de forma paralela con la tecnología hasta nuestros días, con la gran inquietud que el malware no ha sido posible controlarlo a través los proveedores. Existe una gran preocupación originada por los riesgos informáticos a los que se encuentran expuestos nuestros datos informáticos y que prácticamente representan en gran porcentaje nuestra vida. Los problemas de software maligno (malware) se han presentado desde hace varios años atrás, tenemos el problema ahora y la tendencia es que continúe persistiendo a futuro.

El tercer aspecto, se relaciona con la extensa documentación y varios casos de éxito en donde se relacionan que nuevos conceptos y tecnologías, apoyan las problemáticas de diferentes negocios a nivel mundial, con excelentes resultados. Por lo anterior, se tiene presente la inquietud si existen mecanismos no tradicionales que puedan contribuir a la gestión de amenazas informáticas causadas por malware.

Para este trabajo se realizarían consultas del tema en páginas de Internet, como por ejemplo Ebsco, Google Scholar, etc., y libros digitales o físicos.

De otra parte, es importante señalar que para la Prueba de Concepto - POC, se realizar el “análisis predictivo” con los datos informáticos correspondientes a los archivos de log, de las acciones realizadas por el sistema antivirus, instalado en la Institución Financiera, para 2 meses del año 2.016.

Para la elección del proveedor de la POC, se tendrán como opciones aquellos que se encuentren clasificados entre las mejores, en tal caso la clasificación es realizada por empresas que muestran los estudios sobre las funcionalidades de los diversos productos para realizar análisis predicativos.

4. MARCO TEORICO

4.1. Condiciones para la existencia de malware

Actualmente nuestro mundo se encuentra gobernado por 2 tipos de sociedades, la primera es la “Sociedad de la Información” que se caracteriza por la implementación de las Tecnologías de la Información y de la Comunicación (TIC) con la finalidad de crear, modificar y distribuir información, que puede ser enviada a cualquier parte del mundo y en cualquier momento. La segunda es la “Sociedad del Conocimiento” que se encuentra “caracterizada por una estructura económica y social, en la que el conocimiento ha substituido al trabajo, a las materias primas y al capital como fuente más importante de la productividad, crecimiento y desigualdades sociales” (Drucker, 1994).

Las sociedades antes mencionadas utilizan como materia prima “los datos informáticos” y que en adelante llamaremos solamente datos, los cuales crecen de forma exponencial con el pasar de los días, según se puede apreciar en los resultados del estudio elaborado por OBS: “*en un minuto*, en Internet se generan 4,1 millones de búsquedas en Google, se escriben 347.000 twitts, se comparten 3,3 millones de actualizaciones en Facebook, se suben 38.000 fotos a Instagram, se visualizan 10 millones de anuncios, se suben más de 100 horas de vídeo a Youtube, se escuchan 32.000 horas de música en streaming, se envían 34,7 millones de mensajes instantáneos por Internet o se descargan 194.000 apps. En total, en un

minuto se transfieren más de 1.570 terabytes de información.” (OBS Business School, 2015).

En nuestras sociedades los datos representan dinero y la información es considerada como un activo estratégico en las empresas, por tal razón es necesario realizar la Gestión de la Información, a lo largo de todo el ciclo de vida de los datos, teniendo en cuenta que existen actividades transversales que se encargan de protegerlos, con diferentes técnicas especializadas en el área de la seguridad informática. Desde hace muchos años las empresas, a nivel mundial, han invertido miles de millones de dólares diseñando diferentes productos, metodologías y procesos para proteger sus datos confidenciales de cibercriminales o piratas informáticos, pero no ha sido posible detener los ataques informáticos con las soluciones de seguridad informáticas tradicionales; en lugar de éstas disminuir, los ataques informáticos crecen con el pasar de los días y cada vez son más especializados.

A continuación, se relacionan algunas variables, que motivan a personas muy inteligentes, a la creación de malware:

4.1.1. Brecha de Seguridad Informática

Al revisar la extensa documentación en Internet, se identificaron las principales causas del problema antes mencionado y una excelente forma de explicar el por qué las herramientas de seguridad informática no han sido efectivas al 100% para controlar los problemas por malware, es la siguiente (Mangelsdorf, 2013), Gap, Information Security

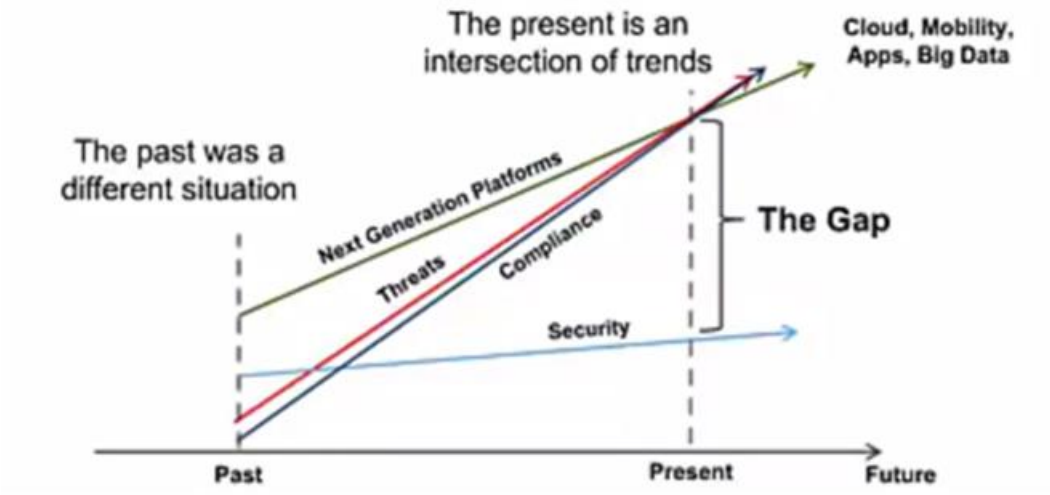


Figura 2 - Brecha de Seguridad Informática

Se utiliza el término “The Gap” o “Brecha de seguridad informática” para indicar la existencia de un problema ocasionado por un espacio existente entre los desarrollos de seguridad informática versus los avances tecnológicos para desarrollar amenazas informáticas, el cumplimiento de leyes para la protección de datos y las plataformas de siguiente generación que avanzan a grandes pasos, de forma incremental y se representan en la gráfica con un pendiente mayor a 45º; mientras que la línea que indica los avances de la seguridad informática tiene una pendiente levemente inclinada, representando avances más lentos, debido a los tiempos de investigación de fabricantes de nuevas tecnologías, la comercialización de los productos, la aceptación por las empresas y su implementación, con el respectivo afinamiento y mantenimiento de la solución informática.

El primer factor mencionado en la gráfica de Mangelsdorf corresponde a las *normas y leyes internacionales*, así como las regulaciones nacionales que aplican a los diferentes sectores. Para el caso de Colombia, entre algunas de las regulaciones se encuentran las siguientes:

Manual de Gobierno en Línea, es una estrategia del gobierno Colombiano y mediante Decreto Único Reglamentario 1078 de 2015, para el sector de las TIC's y comprende cuatro propósitos: " lograr que los ciudadanos cuenten con servicios en línea de muy alta calidad, impulsar el empoderamiento y la colaboración de los ciudadanos con el Gobierno, encontrar diferentes formas para que la gestión en las entidades públicas sea óptima gracias al uso estratégico de la tecnología y garantizar la seguridad y la privacidad de la información." (Gobierno en línea, 2015).

Para la protección de datos personales, el gobierno de Colombia expidió la *Ley Estatutaria 1581 de 2012*, teniendo como aspecto importante la regulación del Habeas Data "como una garantía del derecho a la intimidad, de allí que se hablaba de la protección de los datos que pertenecen a la vida privada y familiar, entendida como la esfera individual impenetrable en la que cada cual puede realizar su proyecto de vida y en la que ni el Estado ni otros particulares pueden interferir, actualmente el hábeas data es un derecho autónomo, compuesto por la autodeterminación informática y la libertad (incluida la libertad económica). Este derecho como fundamental autónomo, requiere para su efectiva protección mecanismos que lo garanticen, los cuales no sólo han de depender pender de los jueces, sino de una institucionalidad administrativa que además del control y vigilancia tanto para los sujetos de derecho público como privado " (Certicamara, 2013).

La ley 1273 del 2009, "Por medio de la cual se modifica el Código Penal, se crea un nuevo bien jurídico tutelado - denominado "de la protección de la información y de los datos"- y se preservan integralmente los sistemas que las tecnologías de la información y las comunicaciones" (Congreso de Colombia, 2009), en la cual se relacionan los delitos informáticos y la protección de información y de los datos; contempla penas de prisión hasta 120 meses y multas hasta de 1.500 salarios mínimos mensuales. Para el trabajo realizado, aplica al capítulo 1 relacionado con:

Uso de software malicioso, interceptación de daños informáticos, violación de datos personales, suplantación de sitios web maliciosos para capturar datos personales y acceso abusivo a un sistema informático.

La ley 527 de 1999, "Por medio de la cual se define y reglamenta el acceso y uso de los mensajes de datos, del comercio electrónico y de las firmas digitales, y se establecen las entidades de certificación y se dictan otras disposiciones" (Congreso de la República, 1999), se pretender dar validez legal a la información electrónica y digital, por medio de firmas digitales y entidades de certificación.

Otra iniciativa del Gobierno *es Open Data* que "corresponde a una filosofía y práctica que persigue que determinados datos de los Gobiernos estén disponibles de forma libre a todo el mundo, sin restricciones de copyright, patentes u otros mecanismos de control, permitiendo el impulso del crecimiento económico, salvaguardar los derechos de ciudadanos y empresas, así como, delimitar las obligaciones de las administraciones" (Open Data, 2015). Al tener grandes cantidades de datos con esta iniciativa es importante tener una estrategia para la gestión de información en sus fases de captura, almacenamiento, procesas amiento y elaboración de reportes.

Existen diferentes entes de control con la función de generar varios mecanismos de protección y cumplimiento que las organizaciones deben de adoptar e implementar, con el fin de proteger los datos propios y de sus usuarios. Para el caso de Colombia, mediante la norma *NTC-ISO-IEC 27001* se han documentado los "requisitos para el establecimiento, implementación, mantenimiento y mejora continua de un sistema de gestión de la seguridad de la información. El sistema de gestión de la seguridad de la información preserva la confidencialidad, la integridad y la disponibilidad de la información, mediante la

aplicación de un proceso de gestión del riesgo, y brinda confianza a las partes interesadas acerca de que los riesgos son gestionados adecuadamente". (Icontec, 2013).

El segundo factor, corresponde a las *actividades de cumplimiento* que son realizadas en las organizaciones mediante la delegación de las funciones de protección de sus datos digitales, así como la plataforma computacional a las Áreas de Seguridad Informática, Áreas de Protección o Áreas de Cumplimiento Informático, quienes se encargan de gestionar aspectos de seguridad informáticas coordinando las Tecnologías de Información y Comunicaciones, los recursos humanos y los diferentes procesos, mediante un documento conocido como Políticas de Seguridad Informática.

De igual forma estas áreas han implementado varios controles informáticos mediante la instalación de diferentes aplicaciones de seguridad informática tradicionales como, por ejemplo: antivirus, firewalls, antispam, controles de navegación, control de acceso, instalación de parches de seguridad de software, sistemas de prevención de intrusos, etc. Todas las anteriores herramientas informáticas generan grandes cantidades de datos mediante el registro de eventos y acciones que se registran en archivos de texto conocidos como logs del sistema. El inconveniente se presenta al procesar cientos de Gigabytes de datos, que en muchos casos no disponen de un formato estándar y un punto único centralizado para el análisis de los mismos.

El tercer factor, corresponde a las *plataformas de la siguiente generación*, que se encuentran conformadas por nuevos conceptos y tecnologías de punta desarrolladas por algunas industrias, para procesar diferentes fuentes de datos estructurados y no estructurados, siendo un requerimiento importante el tiempo de respuesta para lograr obtener una respuesta adecuada a las preguntas realizadas.

4.1.2. Ataques informáticos evolucionan y son más efectivos

Según el resultado del estudio realizado por Verizon (Verizon, Data Breach Investigations Report, 2012), una de las causas más comunes que colocan en riesgo nuestros datos informáticos, son los producidos por el software maligno, término utilizado para hacer referencia cualquier programa informático diseñado para realizar acciones no autorizadas. Entre algunas de las categorías de malware podemos citar a los virus informáticos, caballos de Troya, gusanos, software espía, spam, phishing, secuestro de sesiones, etc. Términos a los que se hará referencia por "*malware"*.

Desde hace algunos años se ha venido utilizado la palabra "fraude informático" para identificar todas aquellas actividades ilegales que se realizan a través de un computador, que comprometen de forma negativa en la plataforma computacional de una empresa pequeña, mediana, grande o también para hacer referencia a la intromisión de un correo electrónico de una persona común y corriente, por un acceso indebido de un tercero sin autorización.

De acuerdo al estudio realizado por Verizon en el año 2016, con una cobertura de 64.199 incidentes de seguridad informática, se concluye que el 89% de los incidentes tienen un motivo de tipo financiero o para el espionaje. Los resultados del análisis concluyen que "las amenazas recientes tienen que ver con ataques de los siguientes tipos: hacking, activos físicos, malware, redes sociales y permisos no gestionados, que afectan los pilares de la seguridad informática de integridad, confidencialidad y disponibilidad." (Verizon, Regmedia, 2016) La estadística muestra que la mayor frecuencia de ataques está relacionada con el malware y que afecta en una gran proporción a la integridad de los datos.

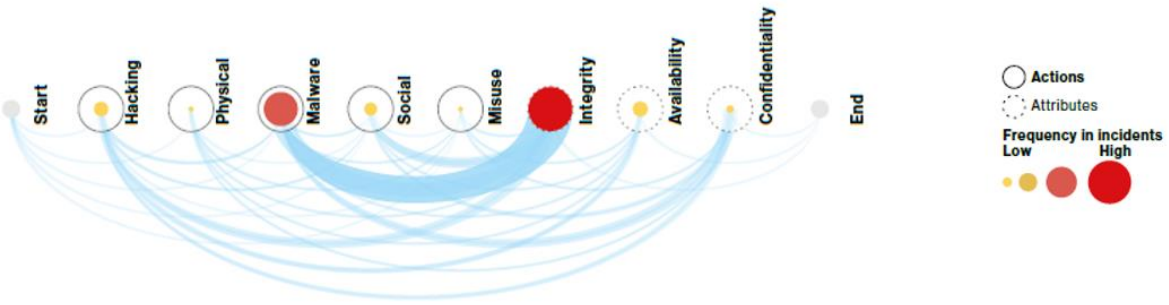


Figura 3 Ataque informáticos más efectivos

El resultado del estudio muestra “una alarmante falta de sistemas de monitoreo de seguridad adecuados en las que las organizaciones quedan vulnerables a delitos informáticos por varios días”. Algunos resultados de las cifras a resaltar son: “el 35% de los fabricantes de tecnologías de información declararon que podían *detectar una fuga de datos en cuestión de minutos* a partir del momento de su ocurrencia. El estudio también reveló que el 22% dijo que a menudo se necesita un día completo para identificar una violación, y el 5% indicó que puede tomar hasta una semana. Esto significa que, en promedio, se tarda 10 horas para una organización para reconocer una violación de la seguridad” (IntelSecurity, 2014).

4.1.3. Las herramientas tradicionales de Seguridad Informática no son efectivas

Las herramientas tradicionales informáticas, como antivirus, antispam, controles de navegación, etc., fueron efectivas en su momento, pero como estamos en un mundo que avanza continuamente, siguen apareciendo nuevas amenazas informáticas y los fabricantes de herramientas informáticas también avanzan. Es el caso de algunas soluciones SIEM (Security and Information Event Management) que permite recolectar eventos de diferentes fuentes como los logs de antivirus, firewall,

sistemas operativos, direcciones IP's. etc., con el fin de identificar eventos anómalos a través de reglas previamente definidas por el usuario. Por ejemplo, el SIEM puede determinar que hay muchos intentos fallidos de log para ingresar a un servidor y ejecuta una regla para bloquear el ingreso. Es probable, una limitación de los SIEM que depende de cómo se revisaron los datos. Normalmente dependen de las reglas que implementa el administrador de la herramienta, ya sea por su experiencia o por los eventos que aparecen en el día – día, pueden no ajustarse a la realidad.

En algunos escenarios, la solución SIEM tiene falencias, como se puede ver en el siguiente ejemplo: Una falla de seguridad se puede originar cuando un hacker ingresa simulando ser un usuario ordinario. Esto se puede conseguir a través de un ataque de tipo "phishing" para capturar el usuario y contraseña válida. A continuación, se puede intentar el ingreso a través de puertos públicos como el correo, portales, etc. y una vez adentro se pueden utilizar las herramientas propias del sistema operativo o utilizar malware con funcionalidades administrativas, así como actividades para cargar o descargar archivos, búsquedas, entre otras; para cometer un fraude, logrando que no sean detectados por las herramientas tradicionales de seguridad. Según lo mencionado en el artículo de Varonis, detectar el anterior comportamiento "se demora varios meses en detectar una violación de seguridad debido a que pasan de forma inadvertida por sus características avanzadas de penetración y ocultamiento" (Varonis, 2015).

Ante esta falencia aparecen las herramientas para el análisis de comportamiento del usuario (UBA: User Behavior Analytics), con la funcionalidad de realizar el análisis a través de los *comportamientos legítimos del usuario*, para identificar los patrones de uso de las aplicaciones que ejecuta, la actividad de la red, frecuencia de uso de los archivos, etc. "Su función es buscar actividades inusuales de los usuarios que se realizan en lugares equivocados" (Varonis, 2015). Según la

empresa Gartner, hay dos clases de UBA; las primeras son las estáticas, que pueden enviar mensajes de notificación en el evento que un archivo o proceso se ejecute fuera de un horario permitido. Las segundas tienen su análisis con base a modelos dinámicos o personalizados. Un factor importante radica en las capacidades de la herramienta para poder recuperar los archivos históricos de los usuarios. Una limitación de los UBA consiste en poder determinar el contexto del dato en donde se está aplicando. Por ejemplo, la palabra “Mercury” tiene diferentes significados como: un planeta, un elemento de la tabla periódica, una marca de carro, el apellido de un cantante, dios romano o una planta de Europa.

4.1.4. Motivación económica al cometer el fraude

Recientemente, en diversos medios escritos y tecnológicos, se mencionan de forma continua la materialización de los de fraudes informático, que traen como consecuencia el delito de robo de dinero. Se puede mencionar que es un fraude muy interesante y lucrativo debido a que no se requieren de muchos recursos económicos, se puede hacer por una sola persona, sin disparar una sola bala ni derramar una sola gota de sangre. A continuación, se menciona un par de ejemplos, ocurridos en el 2016.

El caso más reciente corresponde al pasado 3 de diciembre de 2016, según lo mencionado en CNN Tech, “los piratas informáticos ovaron en el 2016 una suma equivalente a 31 millones de dólares de bancos comerciales, mediante servidores inalados en los países bajos, para lanzar un ataque de denegación de servicio” (Pagliery, 2016)

El segundo caso, corresponde al intento de robo de US$101 millones que estuvo oculto por un mes y fue descubierto gracias a un error ortográfico según lo

informado por (BBC, 2016), “unos hackers, que todavía no han sido identificados y que efectivamente robaron en febrero de este año, la suma de US$101 millones de las reservas de divisas del Banco Central de Bangladesh depositados en una cuenta en el Banco de la Reserva Federal de Nueva York”. Este robo se caracteriza por una excelente planeación, mucha paciencia y sin necesidad de utilizar armas. El plan consistía en apoderarse de forma ilegal de las credenciales del administrador del sistema de transferencias electrónicas del Banco de Bangladesh, a continuación, estudiaron por más de un año las operaciones frecuentes sin despertar sospechas, para luego intentar robar 1.000 millones de dólares mediante transferencias. Los delincuentes suplantaron al Banco Bangladesh y empezaron a realizar operaciones con la Reserva Federal de Nueva York, para transferir los dineros a cuentas en Filipinas. El fraude informático no fue descubierto por los sistemas de seguridad, solo hasta el momento en que hicieron un giro a una cuenta con un error de ortografía ya que escribieron F**a**ndation en lugar del nombre: F**ou**ndation Shalika. Este evento se produjo luego de hacerse efectivas 35 solicitudes de transferencias por cerca de 100 millones de dólares”.

4.1.5. Nuevas tecnologías disponibles en el mercado

Debido a que nos encontramos en un mundo cambiante continúan apareciendo nuevas amenazas informáticas y es necesario adquirir nuevas tecnologías para poder soportar las nuevas demandas de seguridad informáticas de nuestro negocio y que se evidencian en las diferentes empresas de Colombia, así como también a nivel internacional, como se relacionan en los siguientes párrafos.

Gartner es una empresa a nivel mundial que realiza actividades de consultoría y de investigación de las TIC’s y en su reciente estudio menciona que

“las empresas en el mundo son cada vez más conscientes de la *necesidad de implementar tecnologías para analizar grandes volúmenes de datos*, ya que el 75% de éstas están *invirtiendo en Big Data* o planean hacerlo en los próximos dos años. El mismo informe señala que 3 de cada 4 empresas invertirán en Big Data en los próximos 2 años”. (Meulen, 2015).

Big Data agrega un valor a las empresas según “Un estudio realizado por investigadores del Massachusetts Institute of Technology y la Universidad de Pennsylvania con una muestra de 179 compañías, concluyó que aquellas que tomaban decisiones basadas en datos, mostraban entre un 5% y un 6% más de rendimiento.” (Tecnosfera, 2015).

Con las herramientas tradicionales hasta hoy, es muy dispendioso analizar un porcentaje mínimo de eventos de seguridad que pueden originarse entre la gran cantidad de datos estructurados y no estructurados, por lo que los fabricantes de tecnología han desarrollado sus estrategias en diferentes soluciones informáticas, con un énfasis en Big Data. Así mismo, la respuesta en el mercado es evidente según lo indica el estudio que “En 2014 el 73% de las organizaciones mundiales están invirtiendo o tienen planificado invertir en Big Data en los próximos dos años.” (OBS Business School, 2015).

En enero de 2013, la firma investigación de tecnología Vanson Bourne, del Reino Unido, utilizando la metodología de entrevista, realizó el estudio a 500 altos directivos responsables de la toma de decisiones del área de tecnología (200 en Estados Unidos, 100 en el Reino Unido, 100 en Alemania y 100 en Australia), con el “objetivo de investigar qué tan bien se encuentran las condiciones tecnológicas para hacer frente a los desafíos a la gestión de seguridad informática para el tratamiento de grandes volúmenes de información y variedad de datos”. El resultado

del estudio muestra “una alarmante falta de sistemas de monitoreo de seguridad adecuados en las que las organizaciones quedan vulnerables a delitos informáticos por varios días”. Algunos resultados de las cifras a resaltar son: “el 35% de los fabricantes de tecnologías de información declararon que podían detectar una fuga de datos en cuestión de minutos de su ocurrencia. El estudio también reveló que el 22% dijo que a menudo se necesita un día completo para identificar una violación, y el 5% indicó que puede tomar hasta una semana. Esto significa que, en promedio, se tarda 10 horas para una organización para reconocer una violación de la seguridad” (IntelSecurity, 2014).

Según lo mencionado en el paper por Trifu e Ivan , Big Data es un concepto que integra todo tipo de datos digitales estructurados como archivos de datos, de texto, bases de datos, etc., y de igual forma datos no estructurados como gráficos, los sonidos, de música, coordenadas satelitales, sensores, Internet de las cosas, etc. (Mircea Răducu TRIFU, 2014)

“El descubrimiento de Conocimiento a través de los datos tiene como objetivo *extraer información no obvia* mediante el uso de un análisis cuidadoso y detallado” (Usama Fayyad, 1996), a pesar que el artículo se publicó hace varios años, todavía continua vigente y con más fuerza, siendo apoyado por la gestión del conocimiento y los avances en herramientas tecnológicas que permiten el procesamiento de grandes cantidad de datos estructurados y no estructurados, para obtener conocimiento de acuerdo a las diferentes necesidades de cada organización.

Para entender el término de Big Data conviene mencionar la explicación de IBM: “en términos generales podríamos referirnos como 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 semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis. De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales" (Dagos, 2014) .

También podemos encontrar otra definición en medios de comunicación explicando: "Técnicamente, el 'Big Data' es un proceso que consiste en la incorporación de grandes cantidades de información provenientes de fuentes estructuradas (sistemas de información y bases de datos) y no estructuradas (redes sociales), para analizarlas mediante algoritmos y sistemas cognitivos que permiten determinar patrones de comportamiento para apoyar toma de decisiones." (Tecnosfera, 2015).

La importancia de Big Data radica en que la tecnología actual permite realizar muchas actividades en todos los aspectos de nuestra vida y se encuentra la posibilidad dejar un registro de todo lo que ha ocurrido; estas "trazas de información permiten realizar una mirada hacia atrás para saber lo que ocurrió, de igual forma permite saber cómo se encuentra la situación actual y lo más importante es que permite encontrar los hábitos y poder predecir un destino" (Mircea Răducu TRIFU, 2014).

Otra solución tecnológica desarrollada para analizar grandes cantidades de datos ha sido denominada *Business Intelligence*, que se define como "el conjunto de estrategias enfocadas a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en la organización o empresa" (Marqués, 2015). Este tipo de se fundamentan en realizar una inteligencia para analizar los datos a partir de diferentes fuentes de datos estructurados y que son

relacionados con reglas del negocio de cada empresa. A diferencia de Big Data, este tipo de herramientas requieren que los datos tengan una buena calidad de

La siguiente gráfica muestra de forma fácil la comparación entre Business Intelligence vs Analytics (Evaluando, 2013)

Figura 4 Diferencias entre Business Intelligence vs Big Data - Analytics

Otro concepto a tener muy en cuenta es el de *Data Mining* que consiste un "proceso analítico diseñado para explorar grandes cantidades de datos en busca de patrones consistentes y/o relaciones sistemáticas entre variables, con el fin de validar los resultados mediante la aplicación de los patrones detectados a nuevos subconjuntos de datos. El objetivo último de la minería de datos es la predicción" (Dell_Software, 2015).

Es posible "conocer las interrelaciones desconocidas entre un conjunto de datos mediante la aplicación de métodos de varias áreas como *Machine Learning*"

(I.H.Writen, 2011), que tiene como objetivo que los computadores aprendan de forma automática y de acuerdo a un contexto especifico, a través de ciertas técnicas de inteligencia artificial con el fin de identificar patrones a partir de millones de datos.

4.2. Características de las Nuevas Tecnologías

4.2.1. Gestión de los datos

En su comienzo, el concepto de Big Data incluía el modelo de las 3V's para la administración de datos, que ha venido avanzado con el tiempo; a continuación, indicaremos las características de las 5 V's. Según lo relacionado según lo mencionado en el libro (Aguilar, 2013):

"Volumen: Estamos pasando de la era del petabyte a la era del exabyte y para el año 2020 entraremos a la era del zettabyte.

Velocidad: Se requiere que el procesamiento de grandes volúmenes de datos y posterior análisis ha de hacerse casi en tiempo real para mejorar la toma de decisiones.

Variedad: Las fuentes de datos son de cualquier tipo, los datos pueden ser estructurados y no estructurados

Veracidad: Según un estudio de IBM, uno de cada tres líderes de negocio no se fía de las informaciones que utilizan para la toma de decisiones

Valor: Las organizaciones estudian obtener información de los grandes datos de una manera rentable y eficiente".

Para analizar eficientemente los datos con las características antes mencionadas es necesario disponer de métodos para almacenar, filtrar, transformar y recuperar los datos, se disponen de 3 estrategias para el análisis de datos: la primera es el *modelo privado*, consiste gestionar toda la infraestructura

computacional para el procesamiento y análisis en la institución financiera, lo que implica realizar un proceso de compra del hardware, software y servicios de ingeniería; la ventaja es que se dispone de un alto nivel de control de seguridad y privacidad de los datos. La segunda opción es el *modelo público*: que permite realizar el almacenamiento y procesamiento en la nube, con la disponibilidad de realizar buenos tipos de análisis y reducción de costos, evita los inconvenientes por obsolescencia rápida del hardware, permitiendo una fácil escalabilidad; es importante tener en cuenta la parte contractual relacionada con la disponibilidad, privacidad y seguridad. Este modelo puede tener las siguientes las siguientes variantes: en un solo sitio, como el caso de Google File System y la segunda opción que permite el almacenamiento de los archivos a través de múltiples sitios geográficos como el caso de Amazon Simple Storage Service, Nivanix Cloud Storage y Windows Azure Binary Large Objetct. En este modelo es importante tener en cuenta especificar en el contrato con el proveedor los servicios de analítica y los servicios de calidad. El tercer modelo es el *hibrido*, siendo una combinación de las estrategias 1 y 2, antes mencionadas, con la ventaja que desde la nube pública se proporciona lo necesario para la nube privada.

Otra ventaja del análisis en la nube consiste en que algunas empresas disponen en su modelo de negocio la venta de datos recopilados de diversas maneras, en diferentes segmentos de la industria, lo que da valor agregado para el análisis a realizar.

Según lo mencionado con anterioridad, es importante tener el conocimiento de las leyes relacionadas de protección de datos personales, para evitar que estos datos se alojen en servidores de otro país.

4.2.2. Tipo de plataforma computacional a implementar

Durante muchos años, todos los recursos computacionales fueron administrados directamente por las organizaciones, requiriendo de personas con roles y funciones específicas para gestionarlos, así como grandes recursos económicos para poderlos sostener y darles continuidad; este modelo es conocido con el término de “On-Premises”

Cuando se menciona el término de Computación en la Nube nos referimos a un concepto mediante el cual se ofrecen diferentes servicios de recursos computacionales a través de Internet, como muchos beneficios para las empresas, algunos de ellos son: la reducción de tiempos para adquirir y utilizar los recursos computacionales, reducir costos debido a que se paga dinero solo por los recursos utilizados, facilidad para escalar a mayores capacidades sin necesidad de comprar físicamente los recursos, facilidad para acceder a los recursos computacionales desde un navegador web, reducción de costos por licenciamiento de software debido a que está centralizado y los usuarios acceden a un solo punto sin importar donde se encuentren alojados los datos, facilidades para migrar las aplicaciones entre servidores, aumento de las capacidades de almacenamiento y procedimiento, mediciones exactas del recurso utilizado, la seguridad es optimizada debido a que la seguridad física es responsabilidad del proveedor y la seguridad de los datos y de las aplicaciones son responsabilidad del usuario.

A continuación, se explican de forma muy resumida las opciones en que se ofrecen los servicios computacionales a través de la nube:

IaaS (Infrastructure as a Service) es un modelo en donde: el usuario, compra un espacio para gestionar su propia infraestructura computacional, se encarga de proporcionar o definir el sistema operativo, aplicaciones y datos; el proveedor del

servicio, se encarga de administrar la instalación de los recursos computacionales como los servidores, capacidad de almacenamiento, servicios de red y virtualización. A manera de ejemplo se puede mencionar el modelo de Amazon Web Services, con el modelo de negocio de brindar máquinas virtuales en la nube.

PaaS (Platform as a Service) es otro modelo en que el usuario proporciona sus aplicaciones y/o los datos, es decir, las empresas desarrollan sus aplicaciones que se ejecutarán en la nube; por su parte el proveedor, proporciona los demás elementos como el sistema operativo, programas para desarrollar las aplicaciones y demás componentes. Algunos ejemplos son Windows Azure, Google App Engine, etc.

SaaS (Software as a Service) el tercer modelo consiste en que el usuario paga por el servicio de software en la nube, al cual accede a través de la Internet y pagar por el derecho a utilizar el software por un cierto período de tiempo en lugar de adquirir el licenciamiento de un software perpetua, para ser instalado en sitio. El proveedor tiene la responsabilidad del desarrollo, mantenimiento, actualizaciones, copias de seguridad, etc. Por ejemplo: Salesforce, Office365, servicios de Webmail, etc., en donde se paga de acuerdo al número de usuarios que utilicen el servicio.

Con los anteriores modelos de servicios en la nube es conveniente tener en cuenta que las funciones y responsabilidades de las personas que trabajan en áreas de tecnología tradicionales (modelo On-Premise) se han modificado, para la administración del software y hardware, de acuerdo a las nuevas características servicios ofrecidos por los diferentes modelo de Computación en la nube, como se puede ver en la siguiente gráfica (Technet, 2011)

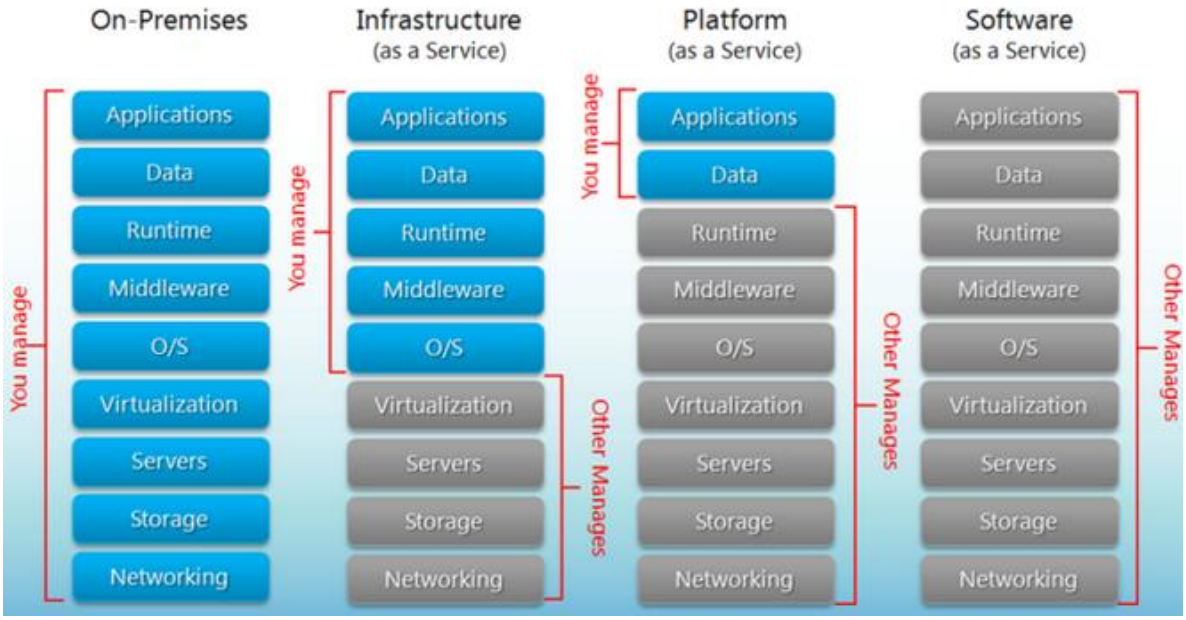


Figura 5 Servicios en la nube

http://robertgreiner.com/2014/03/windows-azure-iaas-paas-saas-overview/

4.2.3. Procesamiento de Datos

Para el procesamiento de grandes volúmenes de datos se mencionarán los siguientes conceptos y su principal finalidad.

Apache Hadoop es un proyecto para el desarrollo de software de código abierto para la computación escalable, fiable y distribuida. "La biblioteca de software Apache Hadoop es un marco que permite el procesamiento distribuido de grandes conjuntos de datos a través de grupos de ordenadores que utilizan modelos de programación simples. Está diseñado para pasar de los servidores individuales a miles de máquinas, cada una ofreciendo computación y almacenamiento local. En lugar de depender de hardware para ofrecer alta disponibilidad, la biblioteca en sí está diseñado para detectar y controlar los

errores en la capa de aplicación, por lo que la entrega de un servicio de alta disponibilidad en la parte superior de un conjunto de ordenadores, cada uno de los cuales puede ser propenso a fallos." (Hadoop, 2016).

Según lo mencionado en la página web de Hadoop, entre las características se encuentran los siguientes módulos:

- "Hadoop Común: Las utilidades comunes que soportan los otros módulos de Hadoop.
- Hadoop Distributed File System (HDFS ™): Un sistema de archivos distribuido que proporciona acceso de alto rendimiento para los datos de aplicación.
- Hadoop HILO: Un marco para la planificación de tareas y gestión de recursos de clúster.
- Hadoop MapReduce: Un sistema basado en hilo para el procesamiento paralelo de grandes conjuntos de datos"

Para el procesamiento de datos es importante tener en cuenta el siguiente par de conceptos. El primero, se refiere a *MapReduce* es "un modelo de programación distribuida y una técnica de procesamiento para grandes conjuntos de datos. El modelo contiene dos tareas: la función *Map* que toma un conjunto de datos y lo convierte en otro conjunto de datos, donde los elementos de dividen en tuplas (clave/valor). La función *Reduce* toma como entrada, la salida de la función Map, y combina las tuplas en otro conjunto más pequeño de tuplas" (Ghemawat, 2004). Los principales beneficios de MapReduce son la escalabilidad y la variedad de datos que puede procesar, como por ejemplo archivos, sitios web tablas de bases de datos, etc.

Para dar más claridad del modelo MapReduce, es posible entenderlo con el siguiente ejemplo que pretende responder a la pregunta: ¿Cuántas veces es nombrado Dios en la Biblia? (Biega, 2012) y la solución es de la siguiente forma "1) Ingresar al modelo un archivo de texto que contiene la Biblia en inglés. 2) Nuestra función Map() nos va a ir leyendo línea por línea el texto de la Biblia y por cada línea se va a quedar con los palabras que la componen. Esta función va a devolver un output de pares de las palabras separadas por un tab y el número 1. 3) La función Reduce() va a recibir el output de la función Map() y va a ir resumiendo por cada clave de los pares (palabras) para devolvernos su frecuencia. El output de la función Reduce() va a ser un listado de todas las palabras (sin repetirse) seguidas por un tab y el número de apariciones que tiene esa palabra (su frecuencia). Como vemos la palabra Dios (God en inglés) aparece 4.472 veces en la Biblia".

El segundo concepto a tener en cuenta son las *Bases de Datos NoSQL,* es uno de los métodos más utilizados para el almacenamiento y recuperación de datos; la característica principal consiste en "almacenar y organizar los datos como una colección de documento, en lugar de tablas estructuradas con campos de tamaño uniforme. La principal ventaja consiste en que el usuario puede adicionar un número de campos de cualquier longitud al documento" (Leavitt, 2010). El beneficio consiste en que los datos no tienen una configuración fija, es decir cada registro puede tener datos con formatos diferentes y en forma compleja. La velocidad de procesamiento se debe a que las operaciones se realizan en memoria y el volcamiento de datos se da cada tiempo determinado.

4.2.4. Construcción del modelo

Para la analítica de los datos es importante la construcción de un modelo que permita realizar los diferentes tipos de análisis a partir de los datos disponibles, con el fin de realizar diferentes pruebas y revisar los resultados obtenidos; luego se ingresaran nuevos datos para ir afinando el modelo e ir evaluando los comportamientos y predicciones a futuro.

Para los ambientes de Data Mining es muy importante el estándar: *Predictive Model Markup Languaje (PMML)* utilizado para importar y desplegar los modelos predictivos, debido a que ha alcanzado un nivel de madurez significativo y con un buen apoyo en la industria, permitiéndole a sus usuarios intercambiar modelos predictivos entre varias herramientas de software.

Se utiliza el motor estándar de la industria de puntuación para modelos predictivos estructurados, conocido como Adaptative Decision and Predictive Analytics (ADAPA), aprovechando la puntuación en la nube, con estándares abiertos ofreciendo una variedad de rutinas para la manipulación de datos y modelos predictivos. (Guazzelli, 2009). De igual manera menciona que los modelos pueden ser compartidos debido a que PMML “es un lenguaje basado en XML. PMML realiza una serie de transformaciones, que permiten la normalización, agregación y mapeo de valores; ofreciendo un conjunto de funciones aritméticas complejas y manipulación de caracteres; operadores boléanos y funciones de decisión. Algunos de los algoritmos predictivos son: Redes Neuronales, Maquinas de soporte Vectorial, Modelos de regresión, Árboles de decisión, Asociación de Reglas, Clustering de Secuencias, Native Bayes, etc.

Según lo mencionado en el paper “Big Data Computing and Clouds” (Marcos D. Assunção a, 2013), en la siguiente tabla se relacionan los trabajos realizados sobre computación en la nube:

Work	Goal	Service model	Deployment model
Guazzelli et al. [64]	Predictive analytics (scoring)	IaaS	Public
Zementis [138]	Data analysis and model building	SaaS	Public or private
Google Prediction API [59]	Model building	SaaS	Public
Apache Mahout [11]	Data analysis and model building	IaaS	Any
Hazy [88]	Model building	IaaS	Any

Figura 6 Trabajos realizados con servicios en la nube

A continuación, se mencionan algunos aspectos de los trabajos realizados por:

Guazzelli (Guazzelli, 2009): Realiza su trabajo sobre una plataforma genérica de infraestructura en la nube conocida cocida como Amazon Elastic Compute Cloud (*Amazon EC2*), con gran capacidad de cómputo en la nube, muchos usuarios activos y acuerdos de servicio (SLA's). El modelo de puntuación ofrecido por Adapa, tiene el objetivo de generar las puntuaciones y gran ventaja es poder ejecutar los modelos predictivos desde cualquier lugar del mundo, a través del sistema de pagos de Amazon. El proceso consiste en cargar el archivo de prueba, para ser analizado y brindar un resultado con las estadísticas sobre el total y porcentaje de archivos coincidentes y no coincidentes. Para los datos no coincidentes se genera una lista en donde se puede rastrear y localizar el origen del problema.

Zementis (Editor, Zementis, 2016) Provee soluciones de software para el análisis predictivo. Sus soluciones son basadas en motor de decisiones ADAPA y el adaptador UPPI (Universal PMML Plug In) que se empaqueta como una herramienta de plug-in para el análisis de líderes en la industria y programas para almacenar datos. Es posible importar y desplegar modelos predictivos, que se integra con aplicaciones de análisis en sitio o basadas en Hadoop y soporta computación masiva en paralelo. El modelo puede construirse para ser ejecutado "on premises" o puede ser configurado como SaaS usando el Servicio como

infraestructura (IaaS) a través de soluciones provistas como Amazon EC2 e IBM SmartCloud Enterprise.

Google Cloud Prediction API (Editor, Google Cloud Platform, 2016) provee a los usuarios un API para la construcción de modelos de Aprendizaje de Maquina. Las predicciones basadas en la nube proveen diferentes herramientas para ayudar a analizar los datos previamente ingresados, proporcionando la coincidencia de patrones y aprendizaje automático. Con el resultado es posible tener una idea de lo ocurrido, detectar tendencias y tomar medidas. La mayoría de las consultas de predicción tardan menos de 200 milisegundos y los datos se replican a través de múltiples centros de cómputo utilizando Google Cloud Store. El uso del servicio es limitado los seis primeros meses y luego el precio se da de acuerdo al acuerdo del servicio de disponibilidad del 99.9 % pagando por lo que únicamente se utiliza.

Apache Mahout (Editor, Mahout, 2016) es un conjunto de bibliotecas de aprendizaje automático probabilísticos y estadísticos, diseñado para que sus algoritmos sean escalables y provee librerías sobre Hadoop con funcionalidades de MapReduce. Sus principales características son: entorno de programación sencilla, amplia variedad de algoritmos prefabricados y un entorno de experimentación con sintaxis R.

Hazy (Editor, Hazy., 2016) es un proyecto que explora la integración de técnicas de procesamiento estadístico con los sistemas de procesamiento de datos, con el objetivo de lograr que estos sistemas sean más fáciles de construir, desplegar y mantener. El problema parte de la dificultad para usuarios al intentar obtener conocimientos específicos para el análisis estadístico y de aprendizaje automática, como Siri de Apple, Watson de IBM y sistemas de recomendación como Amazon y Netflix. Se pretende solucionarlo en la capacidad de combinar rápidamente

algoritmos individuales y lograr su despliegue. La investigación se ha centrado en la categoría de patrones comunes y la construcción de sistemas formados.

Azure Stream Analytics es un servicio de procesamiento en la nube de eventos de dispositivos, sensores, infraestructura en la nube y aplicaciones casi en tiempo real. Es posible cargar millones de eventos y hacer el análisis para comprender patrones de comportamiento, con un poderoso escritorio para visualizar datos, cuando ocurren casi en tiempo real. Su SLA garantiza la disponibilidad de 99.9%. Las principales ventajas radican en que no hay costo por adelantado, no hay pagos por cancelación, el cobro es solo por lo que se usa y la facturación es por horas.

Como se puede evidenciar, con los modelos antes mencionados, existe una gran cantidad documentación en Internet, relacionada con modelos y sus respectivas explicaciones. Para la elaboración de este trabajo y de acuerdo al alcance mencionado al inicio del documento, se busca realizar una predicción que puede ser muy cercana a la realidad o algo parecida con los datos previamente ingresados a un modelo. Puede sonar muy coloquial, pero la predicción no es decir o dar un resultado sin ningún sustento.

La comunidad científica ha realizado extensos estudios sobre modelos predictivos, quienes han desarrollado todas las etapas del *método científico.* Es importante mencionar que la gran mayoría los modelos predictivos tienen un sustento matemático. Los proveedores de tecnología han desarrollado procesos de innovación para poder adaptar los diferentes modelos a las soluciones informáticas que ofrecen a sus clientes en el mercado. A continuación, las soluciones informáticas son implementadas en ambientes de producción en las empresas, con el fin de obtener un resultado a sus necesidades. La buena noticia es que existe en

gran cantidad de casos de éxito y las organizaciones continúan instalando este tipo de herramientas.

De forma muy breve y sin profundizar en los conceptos y explicaciones matemáticas, a continuación, se mencionan las características de un modelo para series temporales, conocido como *ARIMA*.

Una *serie temporal* es una secuencia ordenada de observaciones, con la particularidad que cada observación se encuentra asociado a un momento de tiempo. La intención de estudiar un dato consiste en estudiar el comportamiento de una variable y su relación con otras variables a lo largo del tiempo.

Según lo expresado por Santiago (De la Fuente), “Box y Jenkin han desarrollado modelos estadísticos para series temporales que tienen en cuenta la dependencia existente entre datos, esto es, cada observación en un momento dado es modelada en función de los valores anteriores. Los modelos se conocen con el nombre genérico de **ARIMA** (AutoRegresive Integrated Moving Averange), que deriva de sus tres componentes: **AR** (Autoregresivo), **I** (Integrado) y **MA** (Medias Móviles ”

Así mismo, Santiago (De la Fuente) menciona que la metodología de Box y Jenkins se relaciona en cuatro fases:

> “La primera fase consististe en identificar el posible modelo ARIMA que sigue la serie, que lo requiere:
>
> - Decir qué transformaciones aplicar para convertir la serie observada en una serie estacionaria
> - Determinar un modelo ARMA para la serie estacionaria, es decir, los órdenes p y q de su estructura autoregresiva y de media móvil.

La segunda fase: Seleccionando provisionalmente un modelo para la serie estacionaria, se pasa a la segunda etapa de estimación, donde los parámetros AR y MA del modelo se estiman por máxima verosimilitud y se obtienen sus errores estándar y los residuos del modelo.

La tercera fase es el diagnóstico, donde se comprueba que los residuos no tienen estructura de dependencia y siguen un proceso de ruido blanco. Si los residuos muestran estructura se modifica el modelo para incorporarla y se repiten las etapas anteriores hasta obtener el modelo adecuado.

La cuarta fase es la predicción, una vez que se ha obtenido un modelo adecuado se realizan las predicciones del mismo.”

De forma resumida, como lo menciona (wikiversity, s.f.) “un modelo ARIMA es un modelo dinámico de series de tiempo, es decir las estimaciones futuras vienen explicadas por los datos del pasado y no por variables independientes. El modelo ARIMA (p,d,q) se puede representar como:

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^{p} \phi_i \Delta^d Y_{t-i} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

en donde *d* corresponde a las d diferencias que son necesarias para convertir la serie original en estacionaria, $\phi_1, \ldots, \phi_p$ son los ***parámetros*** pertenecientes a la parte "autorregresiva" del modelo, $\theta_1, \ldots, \theta_p$ los ***parámetros*** pertenecientes a la parte "medias móviles" del modelo, ϕ_0 es una constante, y ε_t es el término de error (llamados también ***innovaciones***).

Se debe tomar en cuenta que:

$$\Delta Y_t = Y_t - Y_{t-1}$$

“

Como se mencionó anteriormente, no se explicará en detalle el modelo ARIMA, sin embargo, si el lector desea más información puede consultar las siguientes páginas web:

http://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf

https://addi.ehu.es/bitstream/10810/12492/1/04-09gon.pdf

http://es.slideshare.net/gleandro/principios-de-econometria-arima

4.2.5. Análisis de datos

Las visualizaciones apoyan de forma considerable la **Analítica de Datos,** por medio de la cual se pretende dar un valor agregado al tipo de análisis realizado para la gestión de riesgos informáticos ocasionados por malware, debido a que las herramientas tradicionales se encargan de la detección, registro de la acción realizada por la herramienta y él envió de eventos de seguridad a una consola de administración, es decir se realiza un análisis de tipo descriptivo y de diagnóstico. Adicionalmente las recientes herramientas de análisis se encuentran en la capacidad de realizar otros tipos de análisis como los predictivos, prescriptivos y preventivos, según lo mencionado por Hewelett Packard (HP, Nur, & Manzano, 2015).

Las soluciones de analítica se pueden clasificar de la siguiente forma: *1)* ***análisis descriptivo****,* examina lo que esa pasando casi en tiempo real, utiliza datos históricos para identificar patrones, se enfoca con el modelamiento de comportamientos pasados para luego generar reportes gerenciales, con el fin de proporcionar unas probables tendencias. *2)* ***Análisis de Diagnóstico****:* ¿es una técnica utilizada para determinar por qué ha sucedido algo, tratan de entender lo

qué paso y por qué?.3*) **Análisis Predictivos**:* Intenta predecir el futuro mediante el análisis de datos presentes e históricos; sus funcionalidades están centralizadas principalmente en modelos matemáticos y estadísticos con el fin de predecir situaciones que podrían ocurrir en determinados escenarios. 4*) **Análisis prescriptivos**:* este análisis permite asistir en la toma de decisiones relacionadas con las acciones a realizar y la respectiva evaluación del impacto en los objetivos del negocio; ayudan al analista a determinar la mejor decisión a tomar entre una variedad de opciones para aprovechar una oportunidad o mitigar un riesgo a futuro. *5) **Análisis preventivo**:* ayuda a disminuir o evitar las fallas, mediante la realización de algunas actividades, para evitar que se produzcan eventos adversos.

Una manera fácil de entender los diferentes tipos de análisis es a través de presentaciones comerciales realizadas por empresas como Hewellet Packard, Microsoft y otras que se enfocan en la gestión de clientes. Para el caso de estudio se ha adaptado a los diferentes análisis para eventos relacionados con malware, con el fin de responder las siguientes preguntas:

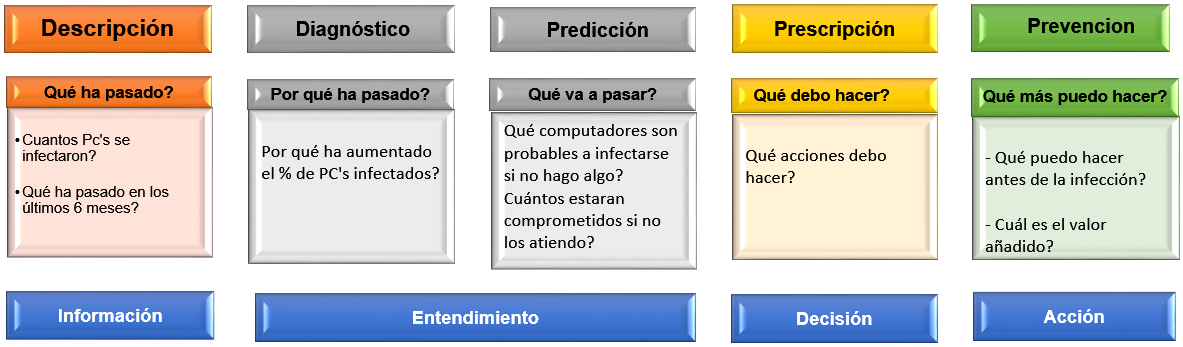


Figura 7 Tipos de Análisis

Fuente: Adaptación de presentaciones, tipos de Análisis HP, Microsoft, Intel Security, etc.

Los anteriores tipos de análisis tienen el fin de mejorar el entendimiento de un problema real y al cual se le puede dar una solución a través de diferentes puntos

de vista. Es un reto interesante para los “científicos de datos” poder transmitir los resultados de análisis de grandes volúmenes de datos, a personas directivas en las organizaciones, quienes conocen muy bien el negocio, pero no se encuentran interesados en términos técnicos relacionados con hardware y software.

Se han desarrollado estudios por diversas organizaciones, entre ellas Gartner, que intentan dar respuestas a las preguntas básicas de acuerdo al tipo de análisis realizado, como se puede apreciar en la siguiente figura (Gartner, Predicitve Analytics, 2015)

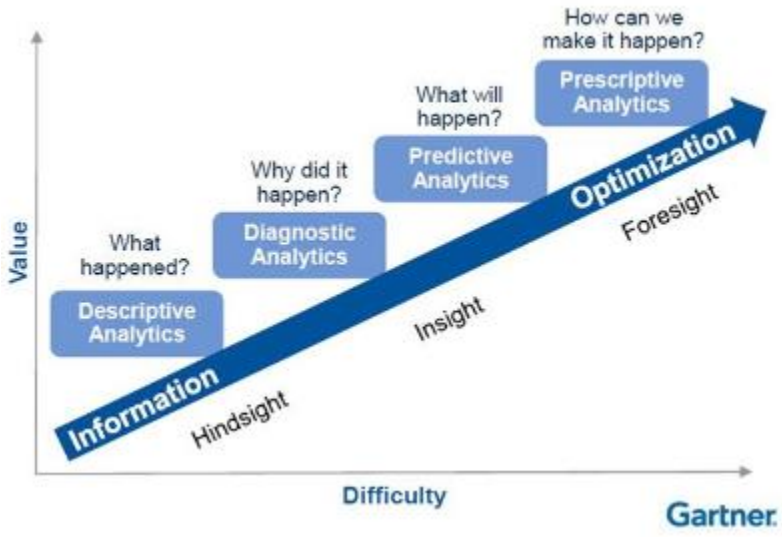


Figura 8 Tipos de Análisis para responder la pregunta

4.2.6. Visualización de los datos

Para poder analizar los datos con las diferentes características antes mencionadas, es necesario que las herramientas utilizadas dispongan de excelentes

funcionalidades para la presentación de los datos. Es importante que el usuario tenga la experiencia que al manipular los datos procesados nube, pareciera como si fueran procesados de forma local, con el fin de evitar los cuellos de botella por la red.

Durante el proceso de divulgación y con el fin de presentar una mejor comunicación de los resultados obtenidos en los diferentes procesos de análisis, es necesario conocer algunas técnicas efectivas que se han desarrollado el concepto de “Storytelling”. Este proceso es desarrollado luego de la exploración y análisis de grandes volúmenes de datos, es el medio utilizado por los analistas de la información, con el fin de explicar los resultados encontrados, a otras personas, que no disponen de conocimientos técnicos sobre el análisis de datos, explicando de manera fácil y visual las causas, consecuencias y acciones a tomar basados en el análisis de datos, para que sean recordados fácilmente.

Se han realizado varias investigaciones sobre técnicas para presentar datos a través de contar historias, siendo los aspectos más relevantes, los siguientes:” En las historias tradicionales el orden corresponde con la línea de tiempo determinando que los eventos tempranos tienen influencia para eventos futuros y no al contrario. Este proceso es similar al trabajo de los periodistas, recolectando información mediante investigaciones, entrevistas, para obtener datos clave y luego atarlos para crear una historia“ (Robert Kosara, 2013).

“En el modelo de clasificar patrones y aproximaciones, utilizados por los medios de comunicación para contar historias visuales, se identifican algunas aproximaciones para estructuras de historias semánticas, llamadas copa de Martini: iniciando con una introducción amplia, se estrecha en algo particular y luego abre una interacción con el espectador” (He, 2010).

El objetivo de una presentación es que el receptor pueda recordar la máxima cantidad de eventos, por lo que es muy importante la memoria visual, apoyado por “el estudio sobre los efectos en la memoria debida a los adornos gráficos (infografía) logran el efecto de facilitar su traerlo a la mente y recordarlo” (S. Bateman, 2010)

Existen varias audiencias para el StoryTelling y cada una requiere diferentes escenarios, como lo mencionado por (Robert Kosara, 2013): a*) presentación de uno mismo a una gran audiencia*, típico para los medios de prensa en donde la presentación se crea una vez, provee una vista estática, para el alcance de varias personas y no hay capacidad de actuar con el autor. *b) presentación en vivo en frente de una gran audiencia:* muy similar ha como se presenta la gran mayoría de negocios, el presentador puede responder algunas preguntas, el presentador puede pausar la historia y adaptarse a los cambios durante la presentación. *c) presentación en vivo en frente de una pequeña audiencia:* similar al punto c, pero existe una interrelación más cercana con los participantes y se pueden hacer seguimientos específicos a tareas o procesos.

Continuando con lo mencionado en el paper de (Robert Kosara, 2013), The New York Times, The Washington Post y otros no existen métricas claramente definidas para evaluar su nivel de eficacia. Las visualizaciones tienden a ser minimalistas y genéricas. Un punto muy importante es la *narración interacción* que permite poder cambiar las vistas de forma rápida, teniendo en cuenta más datos, logrando realizar análisis más rápidos y complejos. Las visualizaciones se pueden apoyar con texto, audio, video, enlaces a más información, para guiar al usuario a través de la historia mediante objetos destacados.

5. HIPOTESIS

Según lo mencionado por (Bernal, Metodología de la Investigación, 2016) a continuación, “se formula la hipótesis con el fin de probar una suposición y no solo mostrar los rasgos característicos de este trabajo”. De igual forma, “es importante tener claro que al aceptar una hipótesis como cierta no se puede concluir respecto a la veracidad de los resultados obtenidos, sino que solo se aporta evidencia a su favor”. De igual forma, menciona que la “Hipótesis de Trabajo (H_A) es la hipótesis inicial que plantea el investigador al dar una respuesta anticipada al problema objeto de la investigación”.

Para este trabajo, la hipótesis de trabajo es la siguiente:

> ***H_A**: Con los conceptos y las TIC’s disponibles en el mercado tecnológico para el año 2016, es posible predecir los computadores que posiblemente serán afectados por malware, mediante el análisis de grandes volúmenes de información logrando reducir los tiempos de respuesta.*

Para la identificación de variables, se tienen en cuenta las definiciones de César: “1) Variable dependiente: *resultado o efecto* producido por la acción de la variable independiente. 2) Variable independiente: Todo aquel aspecto, hecho, situación, rasgo, etc. que se considera como la *causa* de una relación de variables. 3) Variables intervinientes: son todos aquellos aspectos, hechos y situaciones del medio ambiente, las características del sujeto u objeto de investigación que están presentes o intervienen de manera positiva o negativa”.

Para este trabajo, se definen los siguientes tipos de variables:

Variable dependiente:

- Computadores infectados por malware.

Variables independientes: Las más representativas serían:

- Tipo y versión del Sistema Operativo
- Actualizaciones del sistema operativo y aplicaciones
- Software antivirus
    - Versión del producto, motor y definición de virus
    - Acciones del antivirus
- Permisos del usuario en el computador
- Control de lectura y escritura en dispositivos extraíbles
- Permisos de navegación
- Correos electrónicos entrantes y salientes
- Ataques de días cero

Variables Intervinientes: Mencionando unas pocas

- Ciberataques diseñados por Ciberdelincuentes y dirigidos a organizaciones específicas.
- Ubicación geográfica de la organización
- Situaciones coyunturales que captan la atención del usuario, por ejemplo: elecciones presidenciales, accidentes y tragedias, etc.
- Comentarios en las redes sociales
- Páginas especializadas de seguridad informática.

No se diligenciarán encuestas a Instituciones Financieras, debido a que Colombia es un importador tecnológico y de experiencias, por lo que se pretende

recomendar las características tecnológicas necesarias, así como las mejores prácticas a realizar, conservando la independencia de un producto y/o fabricante.

En la realización de la propuesta para la reducción de riesgos informáticos originados por malware, basada en correlación de eventos y casi en tiempo real se debe tener en cuenta las fases realizadas en un análisis tradicional para Big Data y Analytics, formadas por:

"el uso de diferentes tipos de datos provenientes de diversos orígenes para construir un modelo, previa realización de tareas de integración de datos, limpieza y respectivo filtro. Los datos tratados son usados para entrenar a un modelo y estimar los parámetros. Antes de implementarlo es necesario realizar las validaciones con los datos utilizados en ambiente controlado. Finalmente, el modelo es colocado en producción con los datos que se ingresaran casi en tiempo real. Los resultados son interpretados y evaluados, con el fin de afinar los existentes o crear nuevos modelos. Esta fase es denominada modelo de puntuación parar generar las predicciones, prescripciones y recomendaciones" (Marcos D. Assunção a, 2013).

6. PRUEBA DE CONCEPTO

Hasta el momento se ha relacionado en este documento el problema que se ha identificado luego de varios años de experiencia laboral, así como la revisión de la extensa documentación; luego los posibles conceptos y herramientas tecnológicas que pueden ayudar en la solución del inconveniente.

Para realizar esta Prueba de Concepto se realizaron algunas reuniones con los directivos del Departamento de Seguridad Informática, con el objetivo de mostrar los beneficios de utilizar nuevas tecnologías que permiten encontrar una solución a necesidades no cubiertas por las herramientas tradicionales de seguridad informáticas instaladas y en funcionamiento en la Institución Financiera.

Adicionalmente se dispone la inquietud que cada día crecen incrementalmente los datos correspondientes a registro de eventos de aplicaciones, a manera de ejemplo, durante una semana, en promedio **se generan más de mil millones** de eventos de seguridad informática, para aquellas herramientas más representativas dentro de la Institución.

Figura 9 Eventos semanales y representativos de seguridad informática

Debido a los diversos tipos de riesgos informáticos, se tienen instaladas diversas aplicaciones de seguridad y que en varias oportunidades se disponen de diferentes consolas de administración que dificultan la atención oportuna de incidente, debido a la falta de correlación de eventos.

La motivación de este trabajo radica en poder realizar la integración de diferentes tipos de datos estructurados (datos corporativos de la Institución financiera) y no estructurados (datos externos correspondientes a redes sociales y

páginas Web), con el fin de poder realizar predicciones relacionadas con malware que se acerquen y se puedan presentar en el entorno corporativo, para así poder anticiparnos a los eventos de seguridad, para tomar acciones preventivas para evitarlas o reducir su impacto.

En esta sección se mencionarán las diferentes actividades que se realizaron en una Institución Financiera, en Bogotá Colombia, con el objetivo de realizar una "Prueba de Concepto" - POC para intentar realizar: la predicción de los posibles computadores que pueden infectarse por malware.

A continuación, se relacionan, las actividades que se programaron y ejecutaron en la Prueba de Concepto.

6.1. Identificación del Problema y Alcance

Luego de una experiencia laboral cercana a los 20 años, en actividades relacionadas con la administración de soluciones de seguridad informáticas para proteger los activos digitales de la organización, en especial el software relacionado para la gestión de riesgos ocasionados por software maligno, existe la preocupación al saber que, a pesar de los grandes avances de la tecnología, en todas las empresas continúan originándose incidentes de seguridad relacionados por malware. Los ataques exitosos por malware originan pérdidas millonarias a las organizaciones, de forma directa e indirecta.

Para la elaboración de la Prueba de Concepto, decidí elegir los datos correspondientes a las acciones generadas por el software antivirus. Estos eventos corresponden a todas las acciones realizadas por el software, correspondientes al funcionamiento normal como: los registros de inicio al sistema, iniciar una tarea de revisión de virus, las actualizaciones del producto, reinicios, etc.; así como eventos

exitosos relacionados para la limpieza de malware y los posibles mensajes de error, por ejemplo, que no se pudo realizar la tarea debido a que el archivo estaba en uso, no existían los permisos, etc.

6.2. Revisar documentación

La documentación revisada y depurada se mencionó en el capítulo 4, referente al Marco Teórico, relacionando que hay grandes cantidades de información publicada en Internet, a través estudios documentados. Así mismo los proveedores colocan en sus páginas web las bondades de las herramientas informáticas que comercializan.

Se evidenció que hay poca documentación relacionada con el tema de procesos analíticos para la gestión de riesgos informáticos causados por malware, en especial para el sector bancario y financiero; situación que presumiblemente puede ocurrir debido a que existen políticas de seguridad organizacional con el fin de evitar un pánico generalizado, la pérdida de credibilidad y la respectiva pérdida de los clientes, entre otros.

6.3. Selección del proveedor

Partiendo de la hipótesis: *Con los conceptos y las TIC's disponibles en el mercado tecnológico para el año 2016, es posible realizar predicciones para la gestión de malware, mediante el análisis de grandes volúmenes de información logrando reducir los tiempos de respuesta,* se realizaron varias consultas en Internet para tener diferentes opciones de cuáles eran las herramientas disponibles y para sorpresa se encontró el concepto del *Ecosistema de Big Data y Analíticas***.**

Así como se ha mencionado el gran volumen de datos que son generados por diferentes aplicaciones y sensores, también existe en el mercado, más de 350

aplicaciones, herramientas y desarrollos para Big Data, según se muestra en el más reciente panorama de Big Data (Matt Turck, 2016). Se puede apreciar diferentes vistas agrupada por las siguientes categorías: Infrastructure, Analytics, Applications, Cross-Infrastructure, Open Source y los Data Sources Datos & API's.

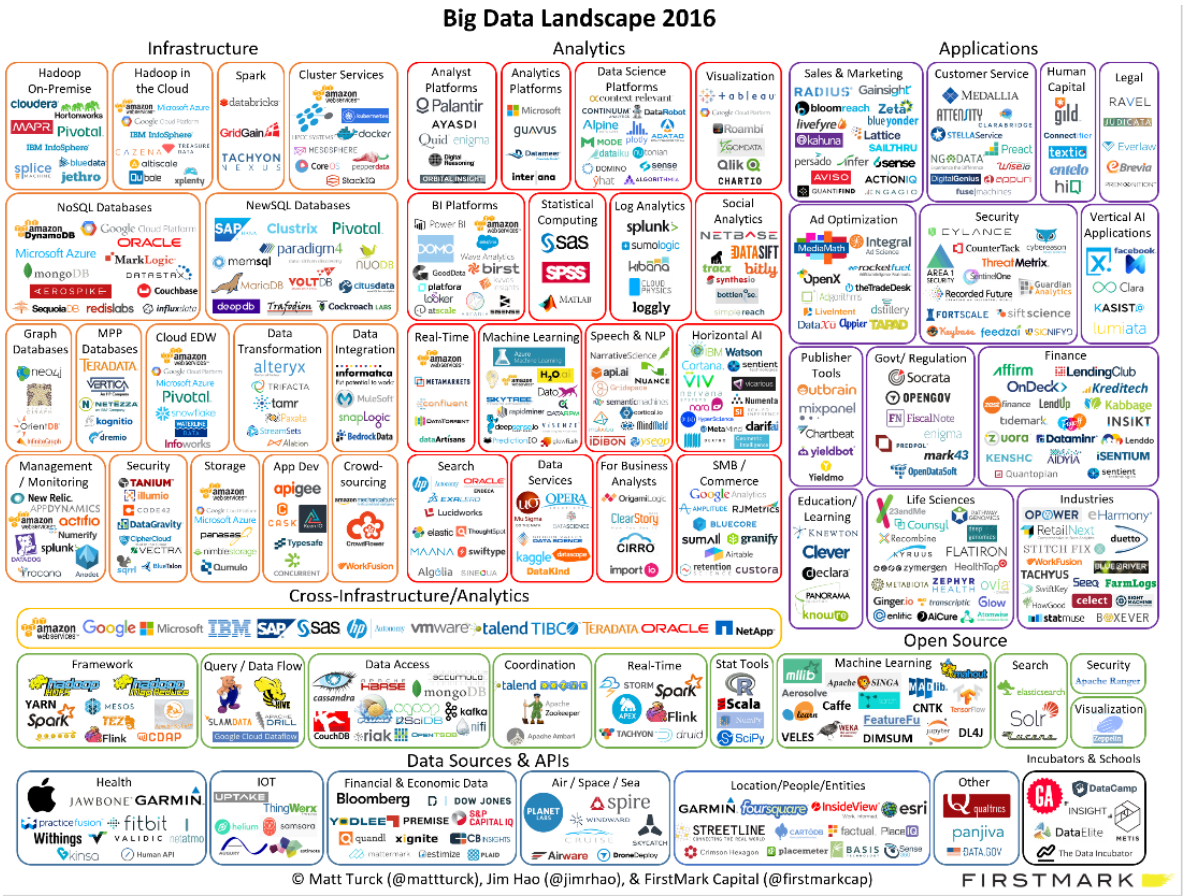


Figura 10 Ecosistema de Big Data

Dentro de este trabajo no se encuentra definido entrar a dar detalles ni profundizarlas, solo se muestran para dar una idea al lector de la gran cantidad de

herramientas desarrolladas para poder gestionar de una buena forma los grandes volúmenes de datos.

Luego de elaborar el marco teórico, del capítulo 4, es importante mencionar que se realizó una labor que involucro varios días, para poder identificar las principales características de las herramientas de “Analytics” y Big Data ofrecidas en el mercado comercial de soluciones de TIC’s.

A continuación, se relacionan algunas características.

Primera: Parece obvio, pero es la más importante. El producto debe de contar con representación en Colombia y debe existir soporte en Bogotá, con el fin de poder realizar las correspondientes configuraciones. En el evento de formalizar el proyecto este requerimiento es primordial para iniciar el proceso de contratación.

Segunda, la solución informática debe poderse instalar en los equipos de cómputo disponibles en la Institución Financiera, así como los respectivos sistemas operativos licenciados. Esto con el fin de rapidez en el proceso de contratación y aprovechar los conocimientos de ingeniería. Es poco viable que aprueben un proyecto si no existe la tecnología ni el recurso humano en la Institución, para dar el soporte. Los servidores y sistemas operativos son proporcionados por la entidad financiera. El software de los aplicativos es proporcionado por el proveedor, según las alternativas disponibles, por ejemplo: software de prueba por un tiempo limitado.

Tercera, para seleccionar al fabricante es deseable que la empresa que provee y da soporte a la nueva tecnología, disponga de tener una trayectoria y experiencia comprobada en Colombia; esto con el fin de disponer de una seriedad y garantías exigidas por la Institución Financiera.

Cuarta, la Prueba de Concepto no obliga a establecer ningún compromiso comercial ni contractual entre la Institución Financiera y el Proveedor de la solución informática.

Quinto, el proveedor se compromete a respetar las cláusulas de privacidad y protección de datos personales establecidos en las políticas de seguridad establecidas por el Área de Seguridad Informática. Algunos de los datos procesados, se pueden correlacionar y brindar información confidencial como identificación de cuentas de usuarios, de computadores, configuraciones de red, etc.

Debidos a factores externos, como la gran diversidad de herramientas y grandes avances de la tecnología, así como los factores internos dentro de la Institución Financiera, que no dispone de las herramientas de analíticas al igual que no se para eventos de seguridad informática tienen recurso humano capacitado en este tema, se considera una buena alternativa para revisar los estudios practicados por empresas especializadas en investigar a las fortalezas y debilidades de los productos de fabricantes soluciones y herramientas de seguridad informa, así como las tendencias del mercado. Por lo anterior, se mencionan los comentarios de dos empresas, como:

Primera, “Gartner es una empresa de consultoría dedicada de manera exclusiva a investigar la industria de las TI, analizar las tendencias del mercado y elaborar el ranking de soluciones tecnológicas para facilitar la selección de soluciones y productos, basados en una metodología de trabajo propia y un equipo de trabajo con una vasta experiencia y distribuido en todo el planeta. Gartner, nos presenta los rankings de fabricantes de tecnologías en algo que denominó los cuadrantes mágicos”. (Revista HelpDesk, 2015).

La interpretación de los cuadrantes de Gartner es la siguiente: La primera categoría, corresponde a los *Líderes* y es la mejor posición a donde quieren estar los proveedores más representativos del mercado, debido a que ofrecen una solución amplia de productos maduros, que evolucionan de acuerdo a las

condiciones y requerimientos del mercado, de igual forma, la salud empresarial y financiera se encuentran en óptimas condiciones. El segundo corresponde a los *Visionarios*, en donde se ubican las empresas que tienen una acertada visión del mercado actual y tienen buenas ideas, pero por el momento no tienen la capacidad de lleva a cabo la implementación de las mismas. En el tercer cuadrante, se ubican los *Retadores* que corresponden a proveedores bien posicionados en el mercado, pero ofrecen poca variedad de productos para cubrir las necesidades del mercado, deficiencias en los canales de ventas o limitaciones en la presencia geográfica para distribuir sus productos. En el cuarto cuadrante, se posicionan los *Jugadores de Nicho*, corresponde a aquellos proveedores que no alcanzan a puntuar en las categorías anteriores, sin precisar que no son buenos productos., pero existen desarrollos en sus planes de negocio.

A manera de ejemplo se muestra el resultado de la publicación para las herramientas de Big Data y Analytics Platforms (Gartner, Magic Quadrant for Business Intelligence and Analytics Platforms, 2016)

Figura 11 Análisis de Gartner

Segunda, Forrester Research, “una empresa independiente de investigación de mercados que brinda asesoramiento sobre el impacto existente y potencial de la tecnología a sus clientes y al público en general” (Forrester, 2016). Esta empresa también publico en mayo del 2016, el siguiente resultado relacionado con plataformas de Big Data y Analítica.

Forrester no nombra a un fabricante como ganador. La intención es asignar a la empresa dentro de una de las cuatro “olas” o segmentos, identificados con los siguientes nombres: desafíos, contendientes, actores fuertes y líderes. Así mismo,

mediante el tamaño de un círculo se representa la presencia que tiene la empresa dentro del mercado.

Figura 12 Análisis de Forrester

Para entender un poco más del funcionamiento y las características de las nuevas herramientas, se identificaron los fabricantes que se encuentran mejor ubicados en los reportes de Gartner y Forrester; luego se solicitó mayor información y se tuvo reuniones informales, a los siguientes proveedores:

- Microsoft
- Hewlett-Packard
- IBM

A continuación, se programó para el mes de noviembre de 2016, realizar la prueba de concepto con 1 fabricante de tecnologías de Big Data y Analytics,

a. alterix
b. Hewlett-Packard, con la solución VERTICA
c. tableau

6.4. Implementación de la herramienta

Para la implementación de la herramienta, se tuvieron algunas reuniones con Hewlett-Packard, quien explicó el modelo de negocio para Big Data y Analytics, mediante un resumen y más de cien casos de uso.

A continuación, el fabricante Hewlett-Packard realizó el contacto con el proveedor itPerform, con el objetivo de comentar los detalles de la prueba de concepto y el caso de uso; el cual se mencionará en un par de páginas adelante.

Las características tecnológicas para realizar los análisis predictivos, se encuentran dados por las diferentes herramientas, es importante mencionar que para el alcance de esta prueba de concepto en resumen se requieren 2 tipos de servidores. El primero es un servidor en sistema operativo Windows Server, con 6 GB en RAM y 500 GB en almacenamiento, para poder instalar la herramienta de analíticas y visualización de resultados. El segundo es un servidor con sistema operativo CentOS, con 10 GB en RAM y 1 TB en disco duro.

Luego el proveedor itPerform preciso que, para la implementación, se necesitan dos servidores, para instalar los siguientes componentes de software:

1) Servidor Windows 2008 R2

- alterix; Es una herramienta que ofrece las capacidades para realizar flujos, de una forma intuitiva y varias funcionalidades para ejecutar tareas de preparación de datos. Para la elaboración de los flujos de trabajo, dispone de la propiedad de “Arrastrar y soltar”

- tableau: Es un software orientado a la Inteligencia de Negocios que permite ingresar diferentes tipos de datos, con la funcionalidad que el usuario final puede realizar su propio análisis, ayudando a mejorar y acelerar la toma de decisiones

2) Servidor Linux CentOs 7.0

- HP VERTICA: Es la plataforma de última generación, de alto rendimiento para el análisis de datos en tiempo real. Permite a las organizaciones gestionar y analizar volúmenes masivos de datos estructurados y semiestructurados de forma rápida y fiable.

Los requerimientos tecnológicos pueden variar de dependiendo del alcance, pero para esta prueba de concepto, se dispone de lo siguiente

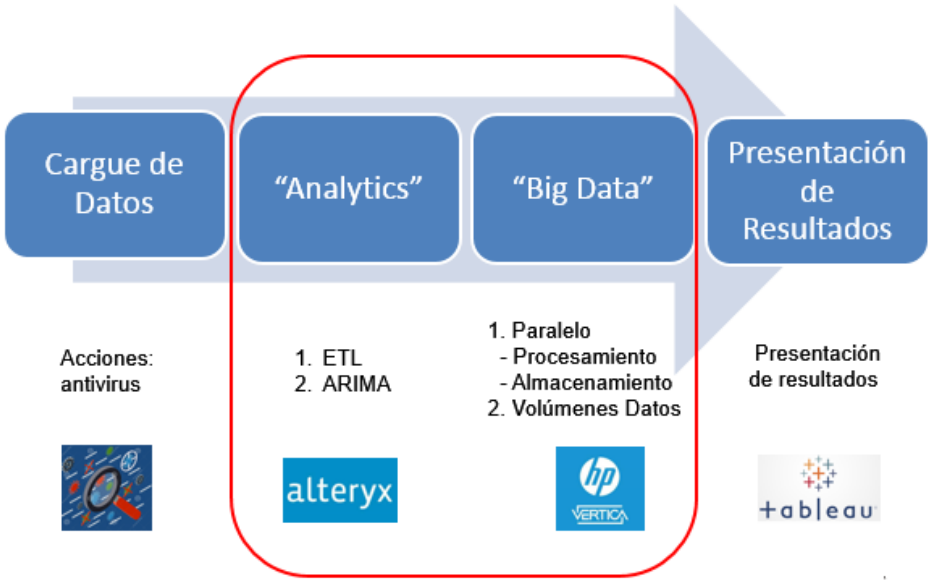


Figura 13 Componentes de Analytics y Big Data

A continuación, se muestran una serie de graficas correspondientes a las configuraciones, análisis e interpretación de los resultados, obtenidos durante la Prueba de Concepto.

En términos generales la arquitectura a implementar para realizar procesos de analítica, a partir de grandes volúmenes, para obtener respuestas a consultas casi en tiempo real, se puede observar en la siguiente gráfica:

Figura 14 Componentes estándar de la arquitectura de la POC

Para la Prueba de Concepto, se realizó la siguiente Arquitectura para la implementación de la herramienta en la plataforma computacional de la Institución Financiera

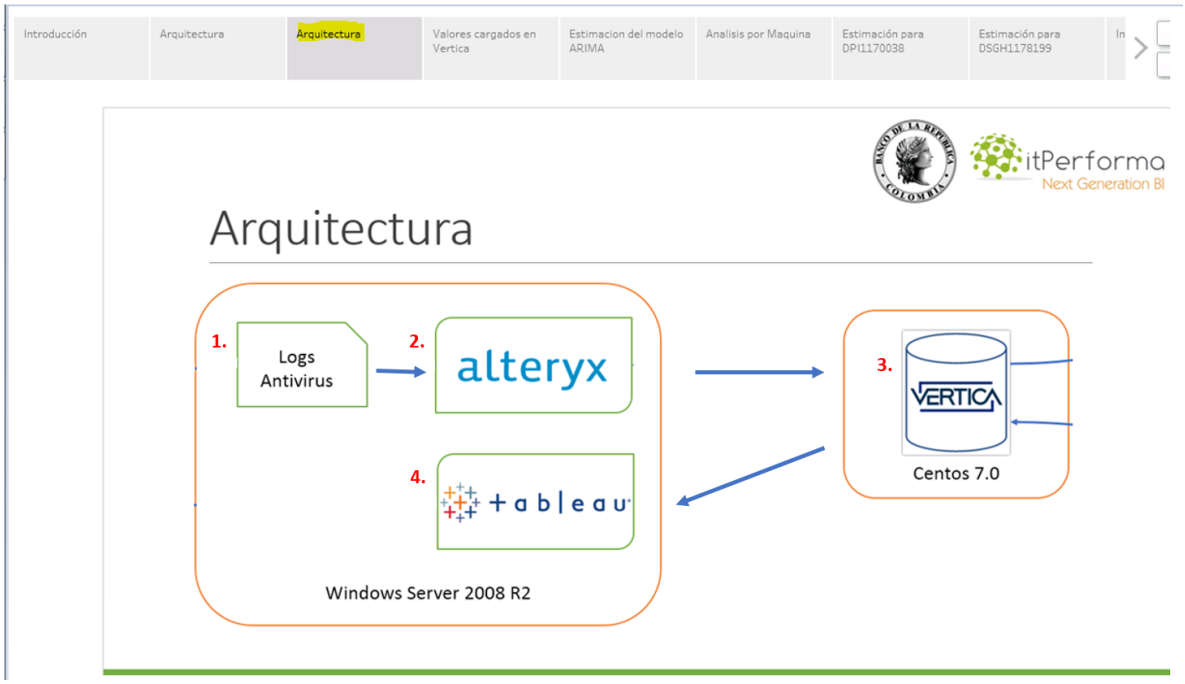


Figura 15 Arquitectura de la POC – Institución Financiera

1. Obtención de datos del log del antivirus, correspondiente a las acciones del antivirus, en los últimos 2 meses.
2. Se cargan los datos del log a **alterix**, para realizar los procesos de Extracción, Transformación y Cargue.
   Se hace la construcción del modelo analítico
3. Para aprovechar las funcionalidades de Big Data, se carga el modelo en **VERTICA**
4. El resultado del modelo se lleva a **tableau** para visualizar datos.

Conviene mencionar que por las características de las herramientas, es posible ingresar los datos de forma automática a través de “jobs”, la ejecución del modelo es muy rápida y la visualización de los datos es en tiempo real.

6.5. Ingreso de datos

Vale la pena mencionar que los datos para Prueba de Concepto se obtuvieron de un ambiente en producción y corresponden a las acciones realizadas por el antivirus instalado en cerca de 4.000 computadores, de la Institución Financiera.

Los datos se encuentran alojados en la herramienta de seguridad SIEM (Security Information and Event Management). Debido a configuraciones recientes se disponen de eventos del antivirus para un período de tiempo, cercano a 2 meses.

El proceso de “ingestión de datos” corresponde a los datos de log de las acciones del antivirus, con cargados a la herramienta **alterix.**

La estructura de los datos a cargar a la herramienta, no corresponde a una base de datos tradicional o formato csv, sino que en cada línea se coloca el nombre del evento y su respectivo valor

Fue necesario realizar una transformación de datos debido a que el log arrojado por el SIEM no conserva un formato único, debido a que para eventos normales puede colocar 39 columnas, mientras que en los eventos relacionados con malware se adicionan 17 columnas, correspondientes a datos adicionales que son relacionados con el malware, a manera de ejemplo

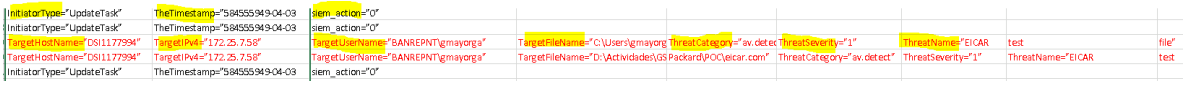


Figura 16 Fuente de datos del antivirus

Así mismo, la estructura no corresponde a una base de datos tradicional o formato csv, sino que en cada línea se coloca el nombre del evento y su respectivo valor.

La anterior situación, origino una oportunidad interesante para probar las funcionalidades para Extracción, Transformación y Cargue de Datos (ETL - Extract, Transform and Load, por sus siglas en inglés), de la herramienta **alterix.**

El archivo con todos los registros de log del antivirus, correspondiente para 68 días, está conformado por 476.591 registros que corresponden a 984 MB, a lo largo del proceso de ETL, se pueden ver las diferentes iteraciones como por ejemplo: eliminar espacios en blanco, cambiar formatos de fecha, cambiar filas por columnas, etc., y los respectivos tamaños de los archivos procesados, alcanzando a tener un tamaño de 1.6 BG, hasta obtener un archivo final solamente con 184.697 registros y un tamaño de reducido de 82 MB.

El tiempo de procesamiento para la ejecución del modelo, en la herramienta alterix fue de 6.13 minutos.

A continuación, se detallan las operaciones realizadas en **alterix**, para las funcionalidades ETL (Extract, Transform and Load), que es utilizado para realizar la calidad de datos.

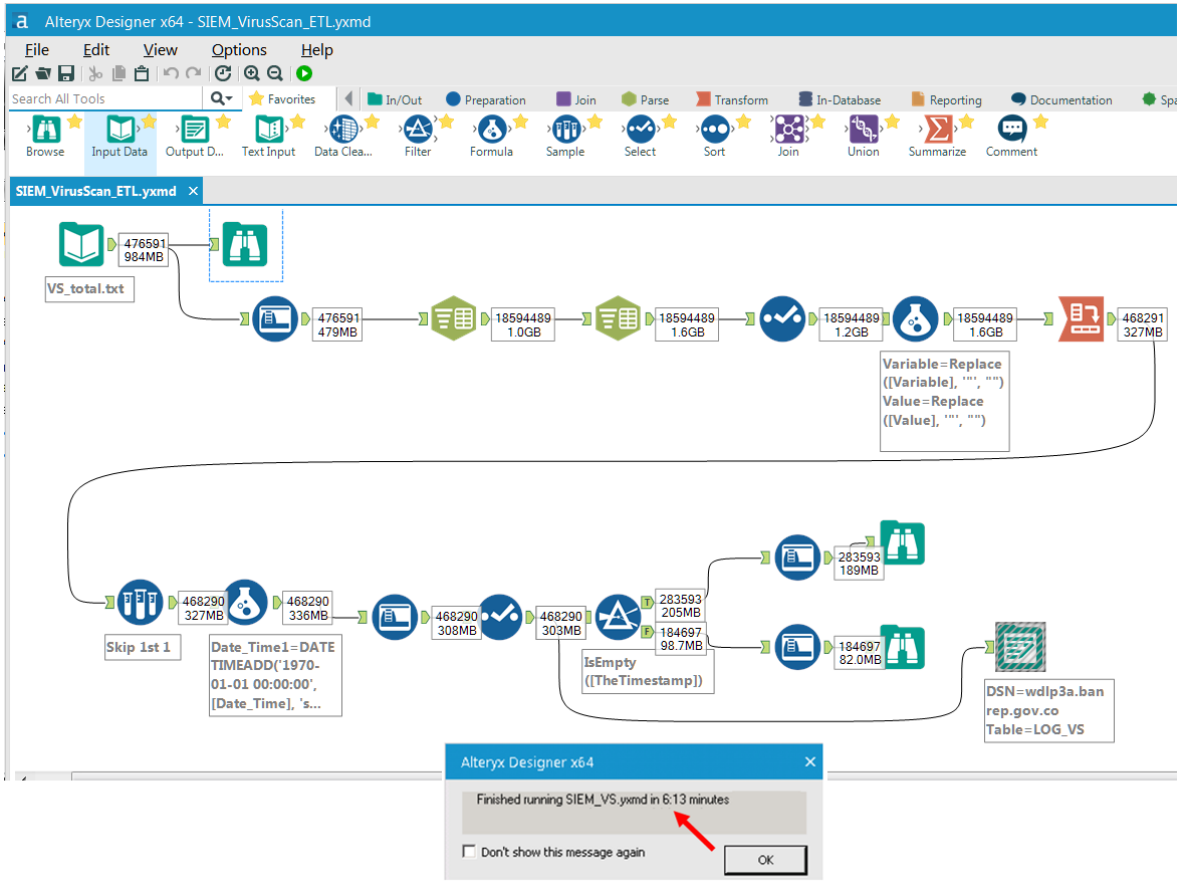


Figura 17 Calidad de datos – Modelo ETL

Para esta Prueba de Concepto, se ingresa un archivo de texto que fue generado por el SIEM, correspondiente a un periodo de 2 meses y solo para eventos del antivirus.

La ventaja del modelo consiste en poder ingresar grandes cantidades de datos, para su análisis dependiendo de los históricos, así como diferentes fuentes de datos. Con el fin de lograr una mayor exactitud en la predicción, se adicionarán

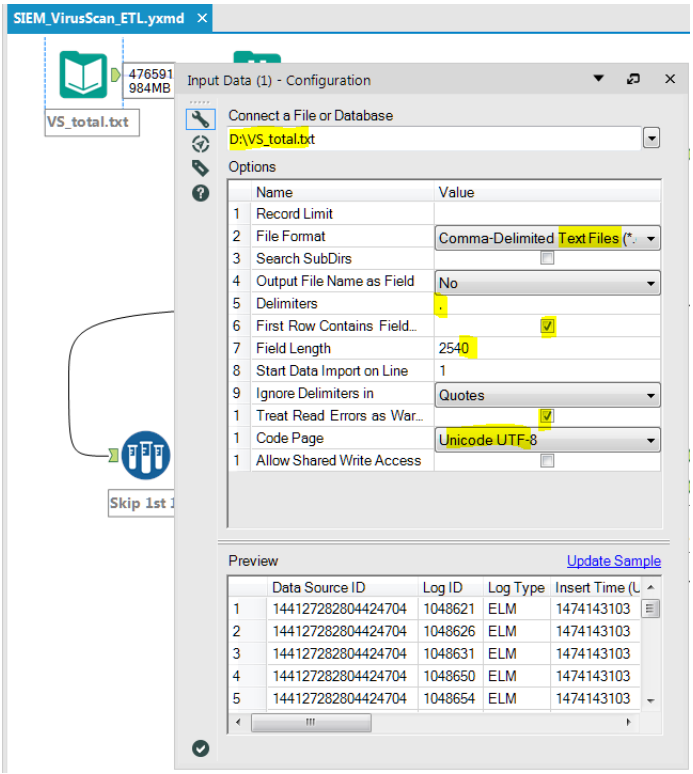
datos estructurados, correspondientes a los eventos de log de las acciones reportadas por las herramientas de controles de navegación, antispam, FireEye y el servicio 7*24; de igual forma se adicionarán al modelo datos no estructurados correspondientes a las redes sociales y a datos recolectados por diferentes extracciones de páginas Web de fabricantes de soluciones de seguridad.

El estimado con las nuevas fuentes de datos es de 20 gigas al mes en archivos de tipo texto, siendo necesario mantener los históricos a lo largo del tiempo. Dependiendo de las fuentes de datos no estructurados es necesario afinar el modelo con más operaciones de ETL, para garantizar la calidad de los datos.

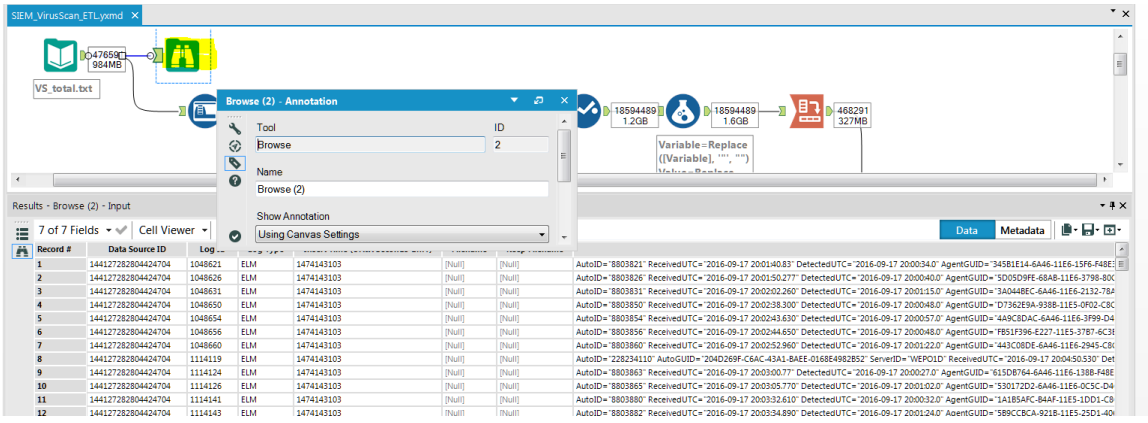
En la herramienta de **alteryx** es posible seleccionar cada elemento del modelo para poder visualizar la respectiva configuración. A continuación se relaciona lo siguiente:

Proceso de cargue de datos: Se ingresa el archivo con los datos correspondientes a las acciones realizadas por el software antivurs, en este caso el Virus Scan del fabricante Intel (antes McAfee). Es importante mencionar que toda actividad es registrada, por ejemplo: acciones de limpieza de archivos infectados con virus, errores en la limpieza, acciones de cuarentena, acicones fallidas, etc. Para conocer más detalles, es posible consultar el artículo de McAfee con la lista completa de eventos del virusScan, donde se relacionan 230 eventos.

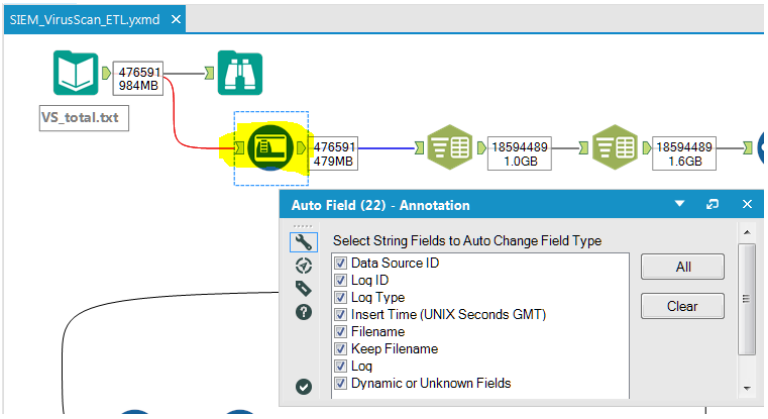
https://kc.mcafee.com/corporate/index?page=content&id=KB52417



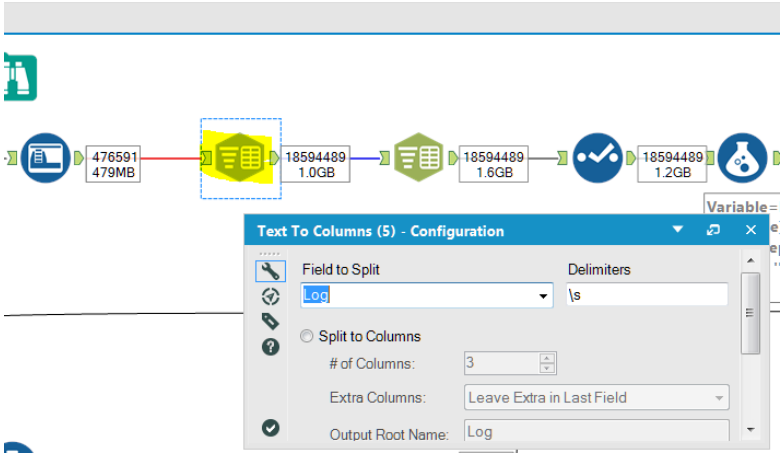
Proceso para visualizar los datos cargados



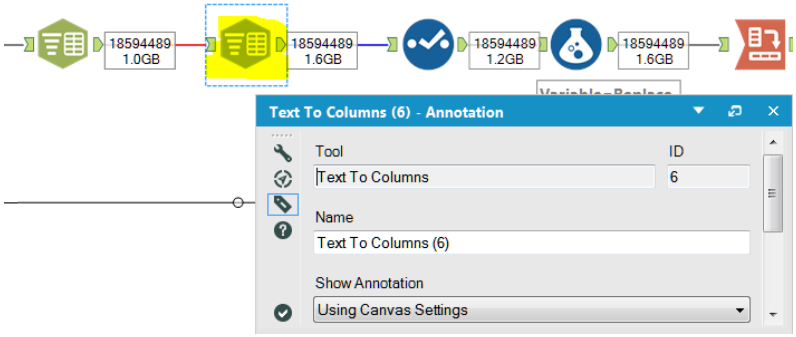
Proceso automático para asignar formato a las columnas, de acuerdo al tipo de datos.



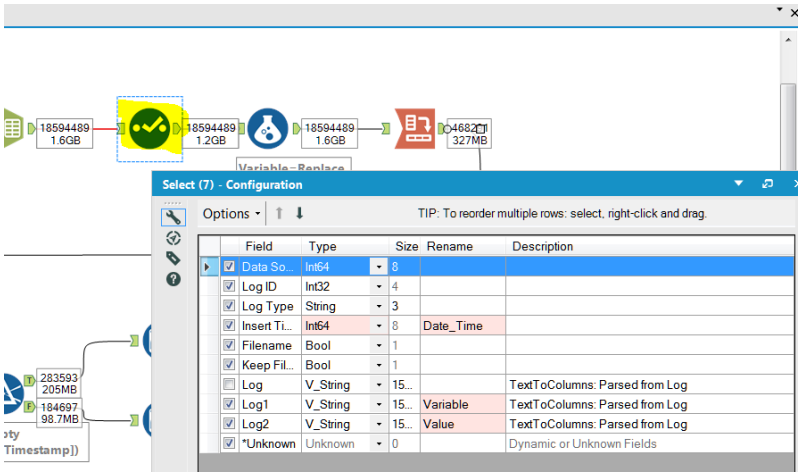
Proceso para eliminar espacios en blanco



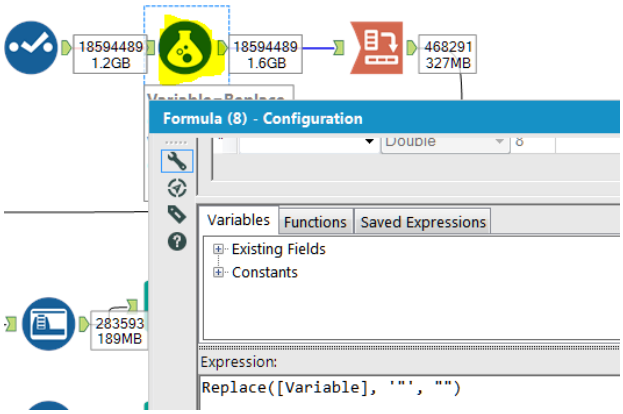
Proceso para convertir texto en columnas



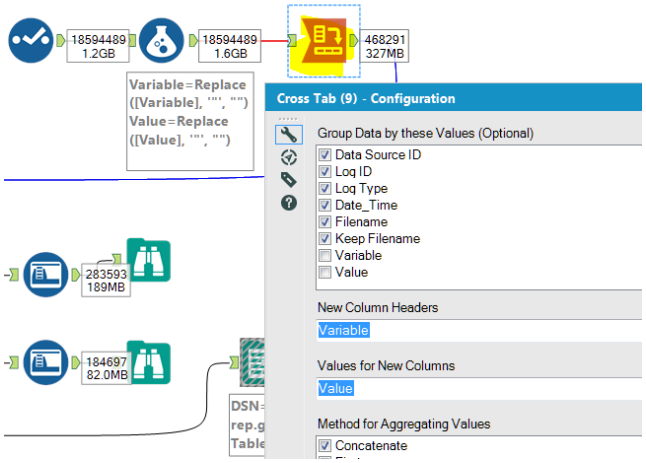
Proceso para establecer el formato a la fecha



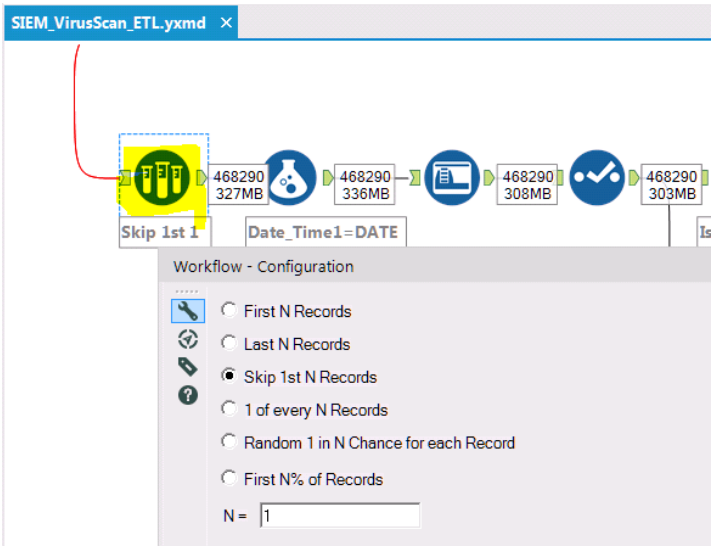
Proceso para retirar las comillas



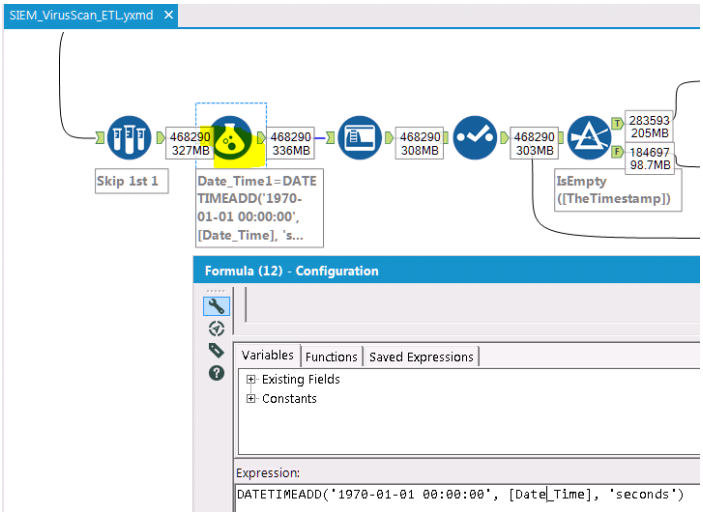
Proceso para agrupar datos



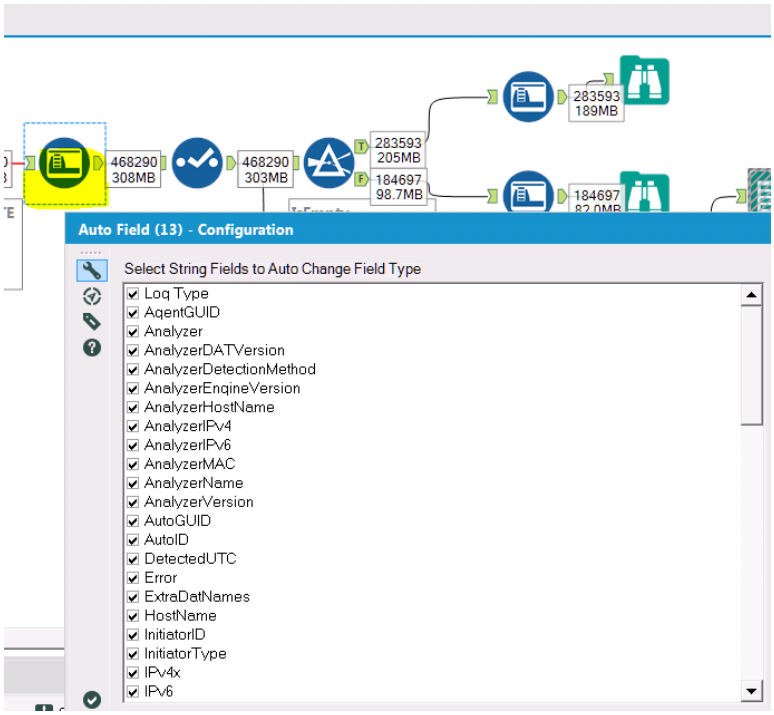
Proceso para recorrer el archivo y totalizar



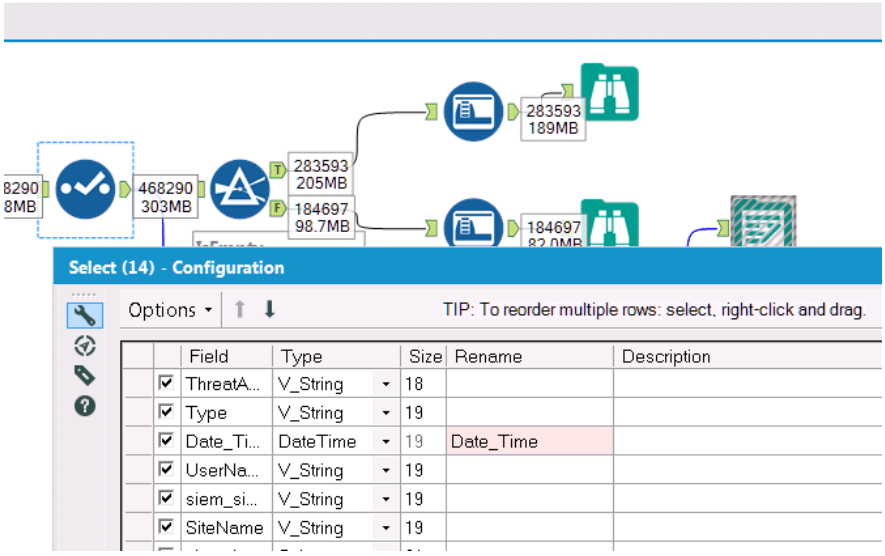
Proceso para ajustar el formato de la fecha



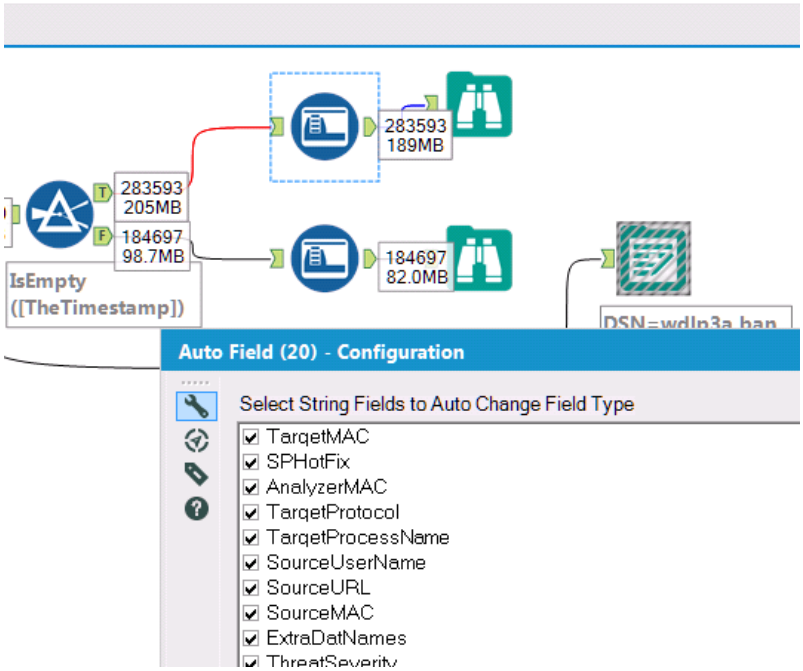
Proceso para seleccionar las columnas.



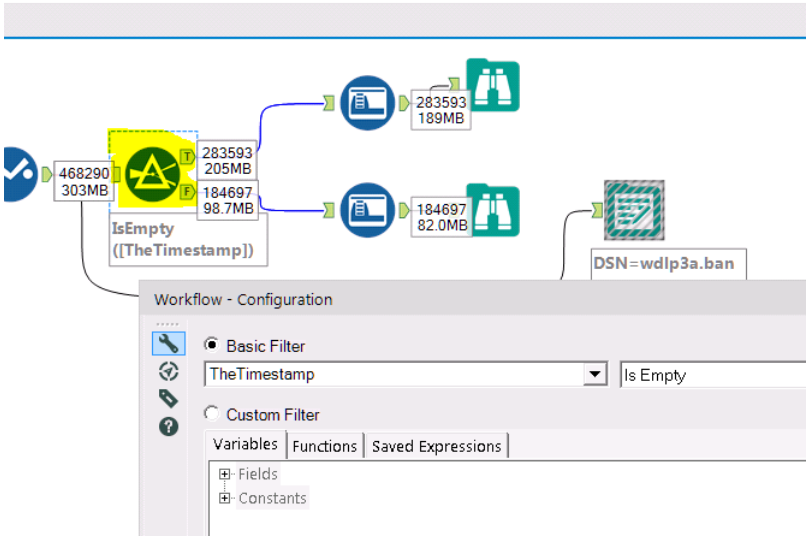
Proceso para cambiar el formato a la fecha



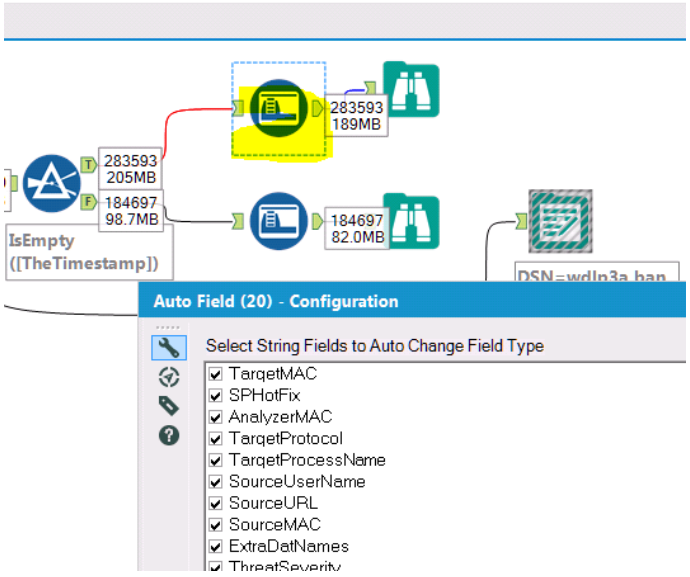
Proceso para visualizar datos



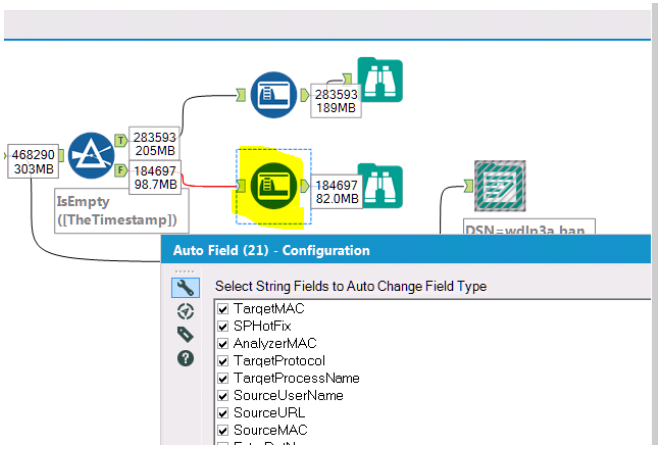
Proceso para seleccionar los campos vacíos, es decir para aquellos registros que no tiene una acción por parte del software antivirus.



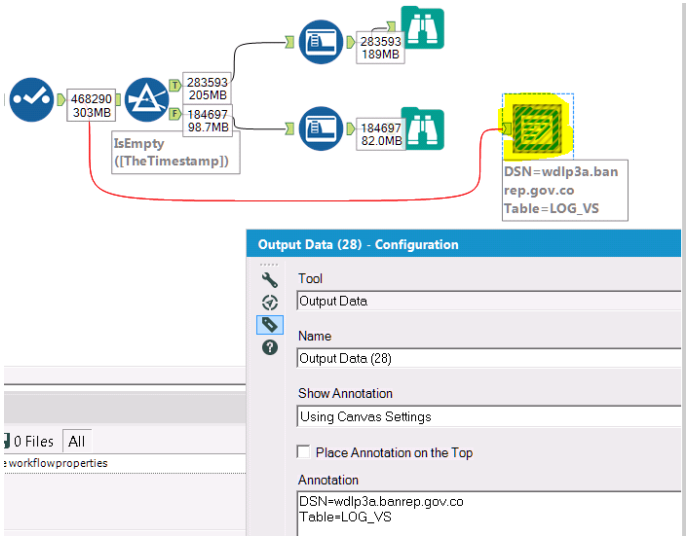
Proceso para visualizar datos



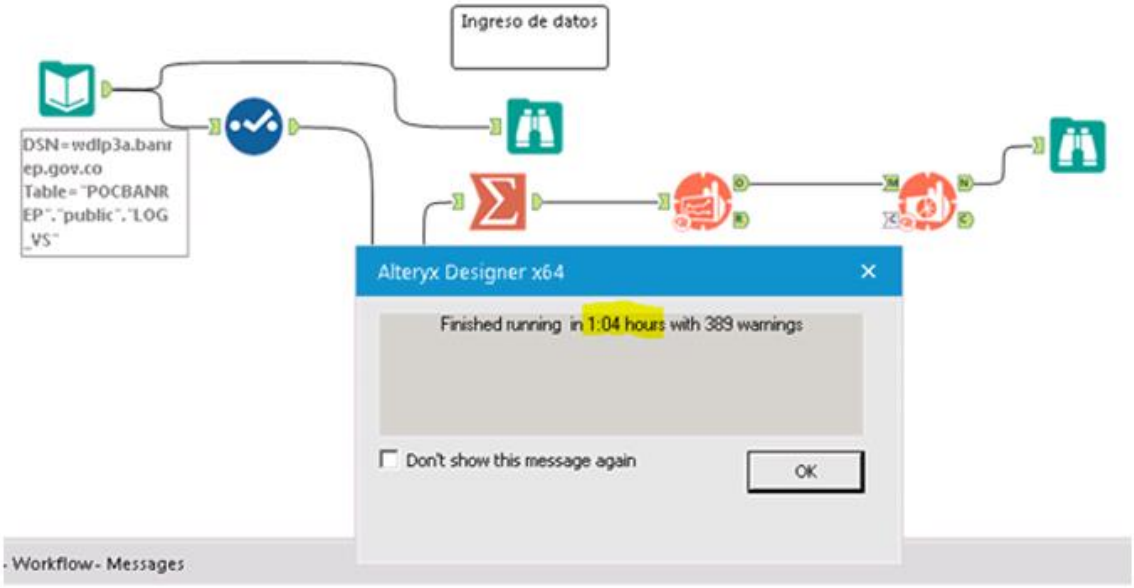
Proceso para visualizar los campos



Resultado del modelo, en un archivo de texto LOG_VS



Al ejecutar el modelo ARIMA, en un servidor dedicado con sistema operativo Windows, el tiempo de procesamiento es de más de 1 hora.



Luego, se ejecuta el mismo proceso en el servidor VERTICA y el tiempo de respuesta corresponde a 6 minutos. Esta disminución se debe a las funcionalidad básicas mencionadas en el concepto de Big Data, como por ejemplo las actividades en paralelo que se realiza para el procesamiento y almacenamiento, funciones de map-reduce, etc.

Como resultado del anterior modelo, los datos son cargados en el sistema, se pueden visualizar y manipular

6.6. Análisis de datos

El análisis que actualmente se realiza en la Institución Financiera corresponde a un “análisis tradicional” que es realizado por dos procesos: El primero corresponde a la herramienta SIEM y el segundo es ejecutado por medio de la generación de consultas de la consola de Administración del antivirus.

Para continuar con el lineamiento de este trabajo, a continuación se mencionan los modelos de **análisis predictivos**. El primero, corresponde a los *modelos predictivos* que tienen la finalidad de analziar los resultados anteriores para evaluar que probabilidad tiene un elemento para mostrar un comportamiento especifico en el futuro, de igual forma encuentra patrones para responder a las preguntas de cómo se va a comportar?. Los cáculos son realizados en tiempo real. El segundo, tiene que ver con los *modelos descritpivos*, con la finalidad de describir las relaciones en los datos para clasificarlos en grupos. Este modelo intenta realizar la predicción solo sobre un elemento y normalmente el procesamiento de datos es realziao de forma offline.

Por ultimo, los *modelos de decisión* describen la realción de todos los elementos que intervienen en una decisión.

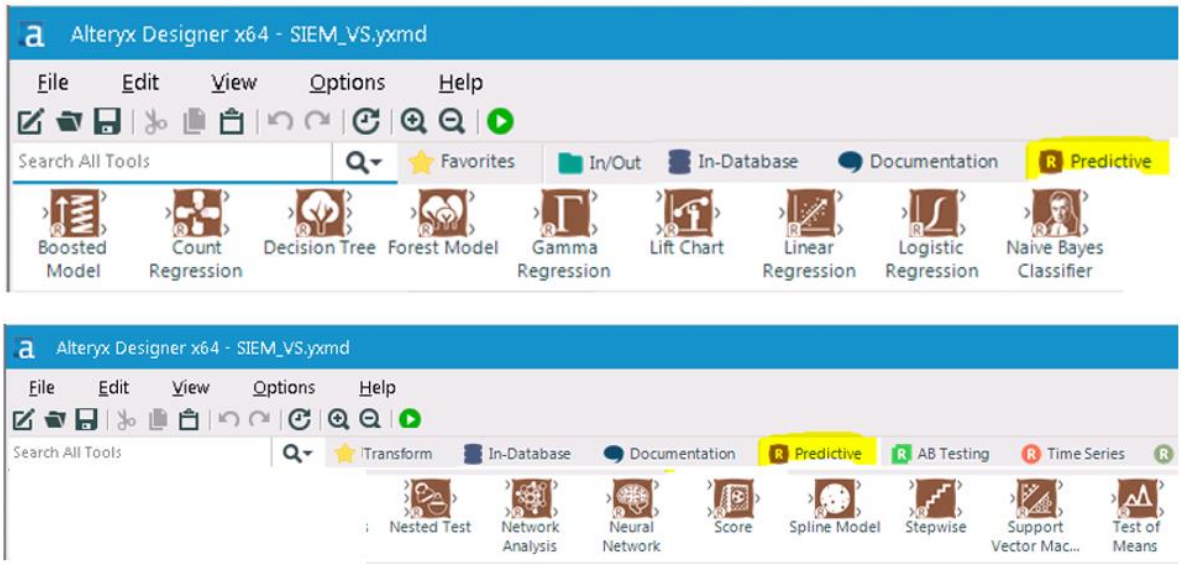
De acuerdo a la Prueba de Concepto, para el uso de analíticas se disponen de la siguiente configuración de las herramientas de analíticas:

El análisis es realizado por la herramienta **Vertica** a través de varios modelos predefinidos**,** de tipo Predictivos y Prescriptivos como se puede ver en las siguientes gráficas.

Las herramientas informáticas para realizar análisis de predicciones, traen incorporados varios modelos estadísticos o matemáticos, para poder realizar el

análisis de datos. A manera de ejemplo, de forma similar el software Excel dispone de fórmulas para realizar operaciones con números.

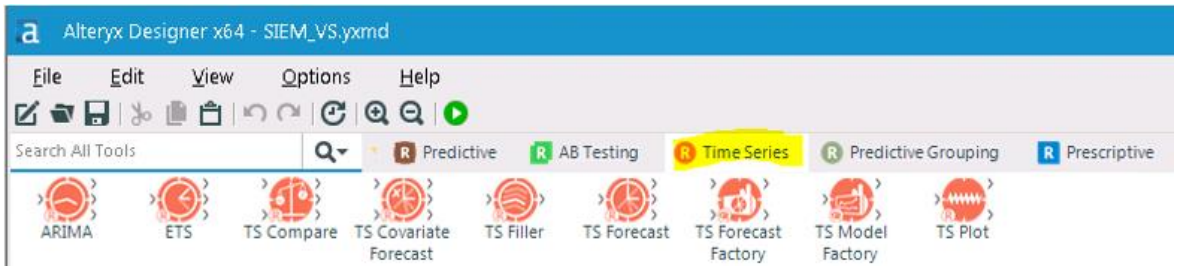
Entre *los modelos predictivos*, que la herramienta dispone de los siguientes: Boosted Model, Count Regression, Decision Tree, Forest Model, Gamma Regression, Lift Chart, Linear Regression, Logistic Regression, Naive Bayes Classified, Nested Test, Network Analysis, Neuronal Network, Score, Spline Model, Stepwise, Support Vector Machine y Test Of Means.



En las pruebas de configuración, se exploró el modelo predictivo de "*Decision Tree*", se ingresaron los datos y el resultado, solo trae cual el computador con más cantidades de eventos que posiblemente se informara el antivirus. Como lo indica su nombre, este modelo apoya a la toma de decisiones para tomar la mejor opción.

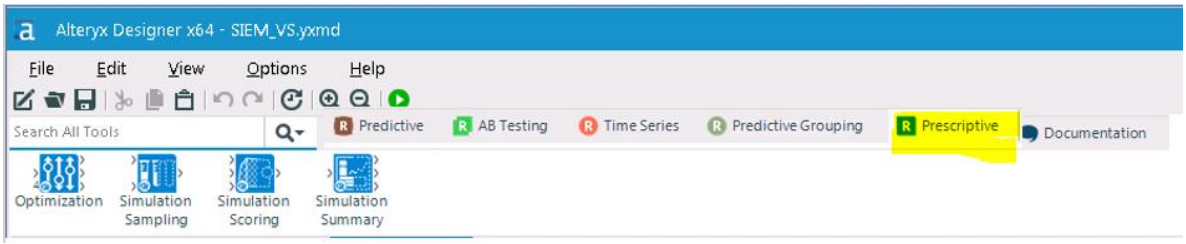
De igual forma, se realizaron pruebas con el modelo de "*Naive Bayes*" que dispone de la funcionalidad de predecir mediante una clasificación. En este caso, el modelo nos predijo los computadores agrupados por un tipo de malware.

Para las *Series de Tiempo*, se cuenta con modelos: Arima, ETS, TS Compare, TS Covariate Forest, TS Filter, TS Forecast, TS Forecast Factory, TS Model Factory y TS Plot.



Debido a la documentación consultada, existen varias recomendaciones para el modelo ARIMA que utiliza el modelado por series de tiempo. Este modelo permite describir y pronosticar la serie del comportamiento correspondiente a las acciones realizadas por el software antivirus en cada uno de los computadores y en donde la variable tiempo juega un papel muy importante, debido a que se disponen de registros por día. El modelo estadístico utiliza fu500ncionalidades de variaciones y regresiones de datos estadísticos con la finalidad de encontrar un patrón de comportamiento en los computadores.

Para el *análisis prescriptivo*, existen los modelos de: Optimización, Simulation Sampling, Simulation Scoring y Simulation Summary



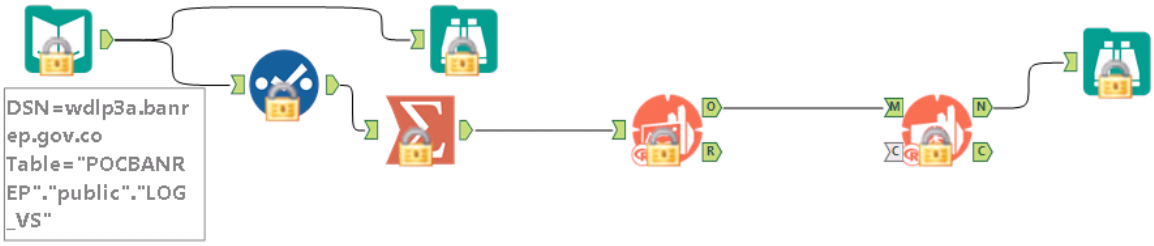
El modelo predictivo configurado, es el siguiente:

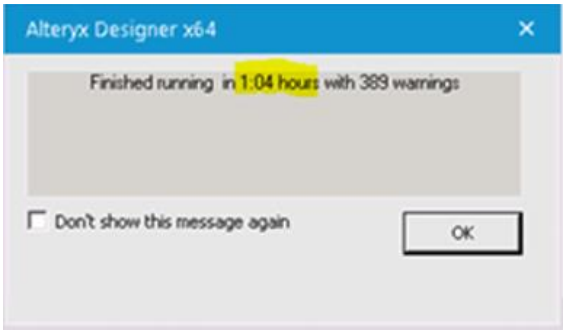
A los datos obtenidos desde la solución de seguridad “Siem” y que corresponden a las acciones realizadas por el software antivirus que protege a los computadores de la institución financiera. Cuando se carga el archivo a alterix, se más de 3 minutos.

A través de alterix, se le aplican funciones básicas de ETL, con el fin de preparar los datos que serán utilizados en el modelo, disminuyendo el tiempo de cargue a 10 segundos en la primera consulta; para las consultas posteriores tarda menos de 2 segundos.

Los datos procesados se ingresan al modelo predictivo ARIMA con el fin de obtener la predicción de computadores infectados.

La ejecución del modelo en alterix ejecutado en un servidor con sistema operativo Windows, tardo más de una hora en mostrar el resultado, teniendo en cuenta que solo se ingresó una fuente de datos (la del Siem). Este tiempo se verá incrementado cuando se disponen de varios insumos al modelo.





Vale la pena mencionar que alterix se ejecutó en un servidor con sistema operativos CentOS y es interesante ver el tiempo de procesamiento al ejecutar el modelo en VERTICA se demora en promedio 5 minutos. Este resultado apoya y da valor a la hipótesis para ver que existen herramientas en el mercado que permiten procesar grandes volúmenes de información en tiempos muy reducidos.

A continuación, se muestran las gráficas correspondientes al **modelo ARIMA**

Proceso de ingreso del archivo con el resultado del proceso de ETL, con la respectiva calidad de datos.

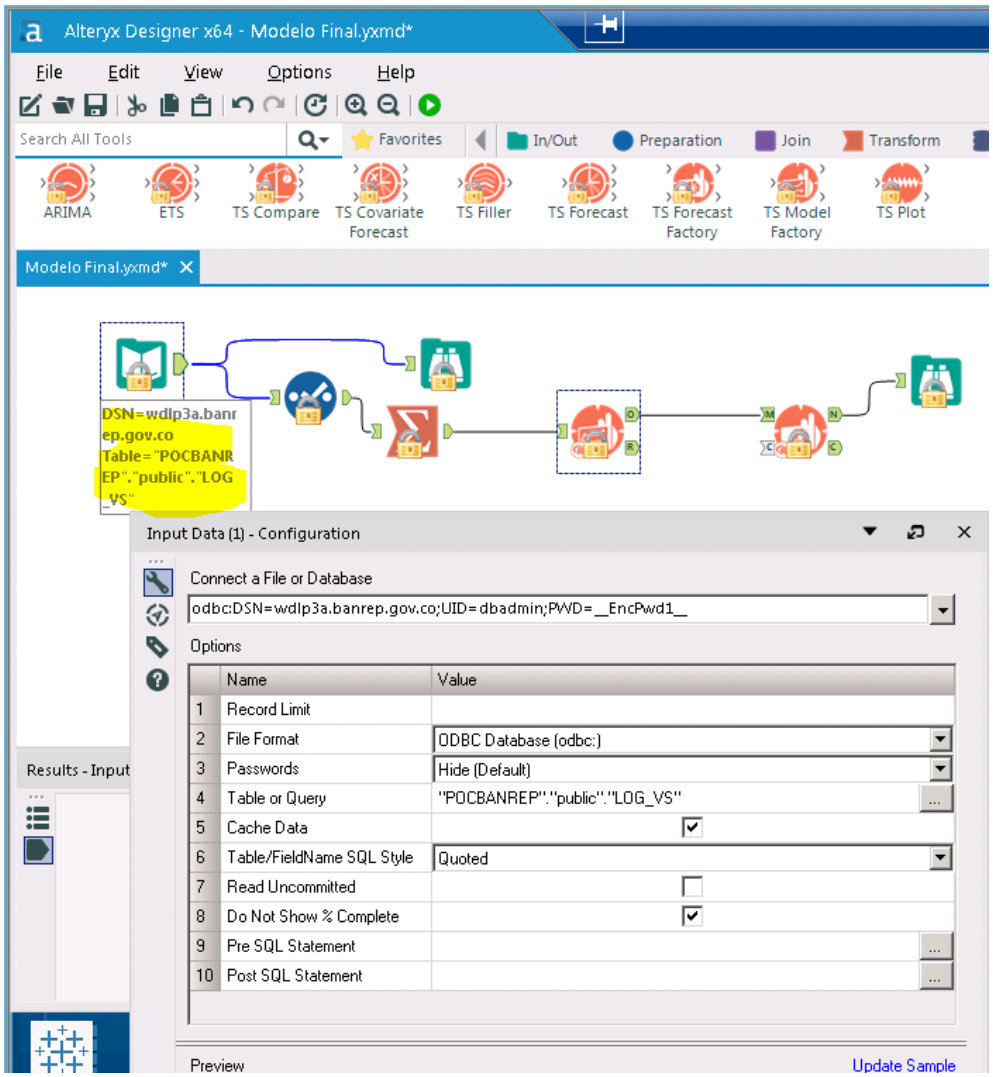
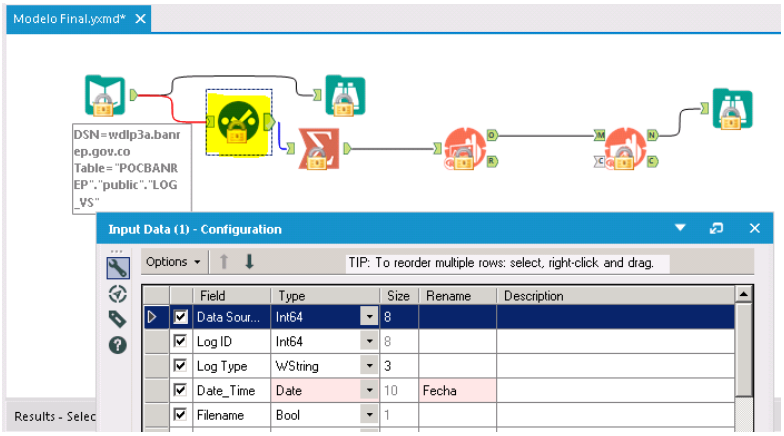
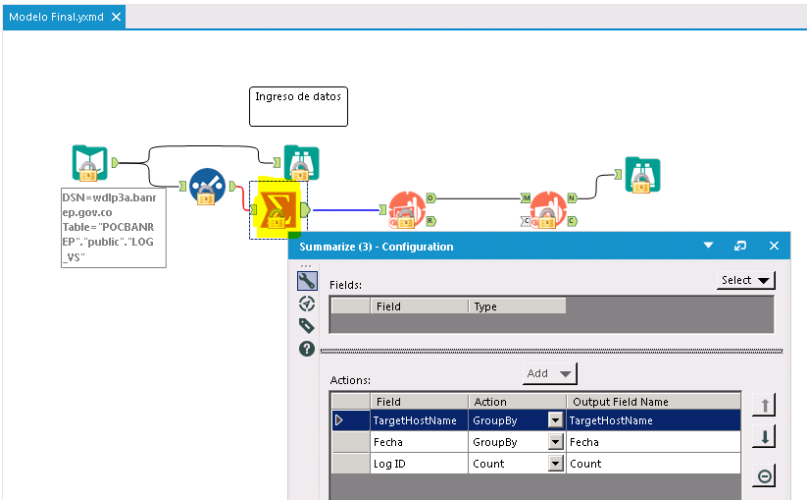


Figura 18 Modelo de Analytics - ARIMA

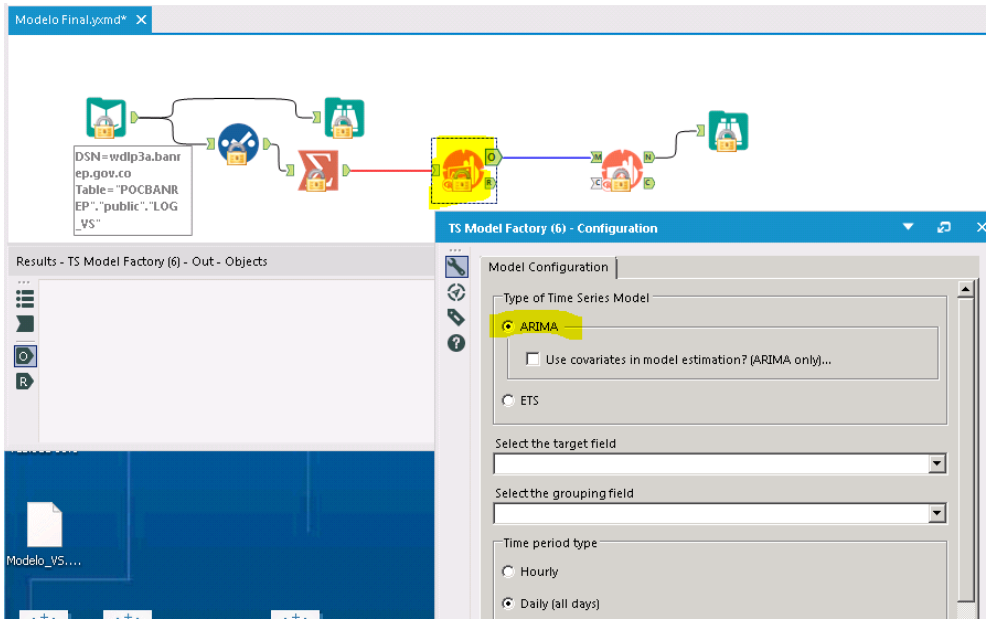
Proceso para configurar el campo de la fecha



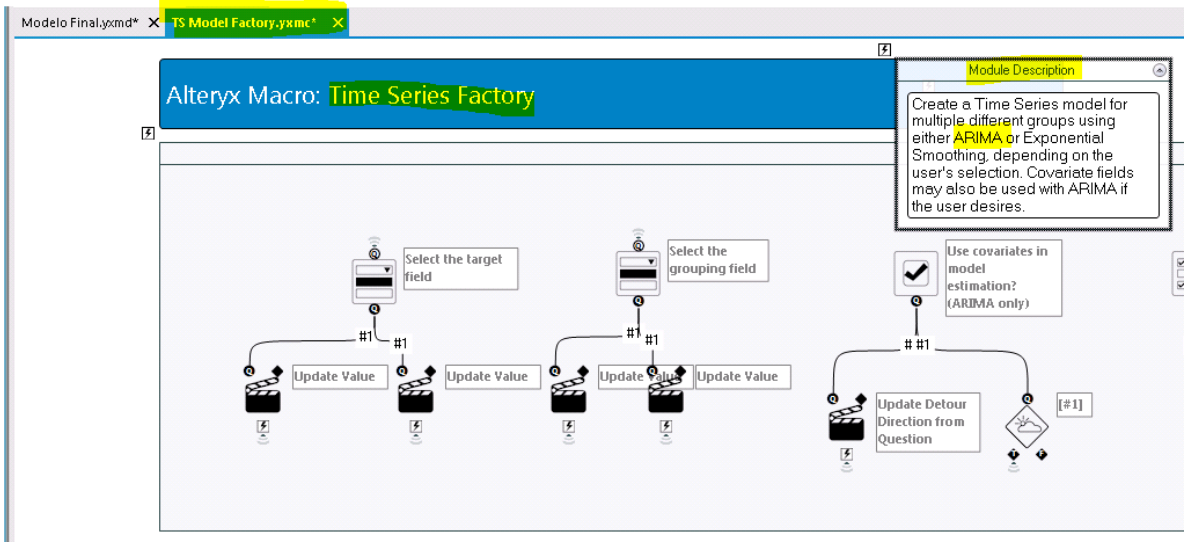
Proceso para agrupar por la fecha. La base del modelo son las series de tiempo

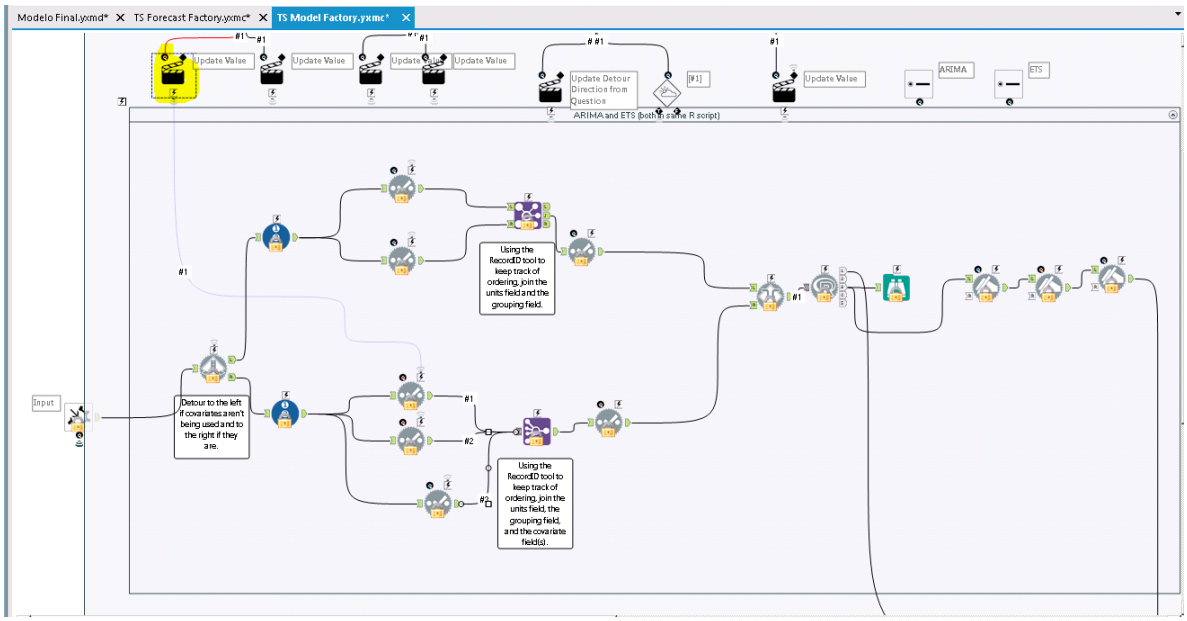


Proceso para definir el modelo de series de tiempo ARIMA

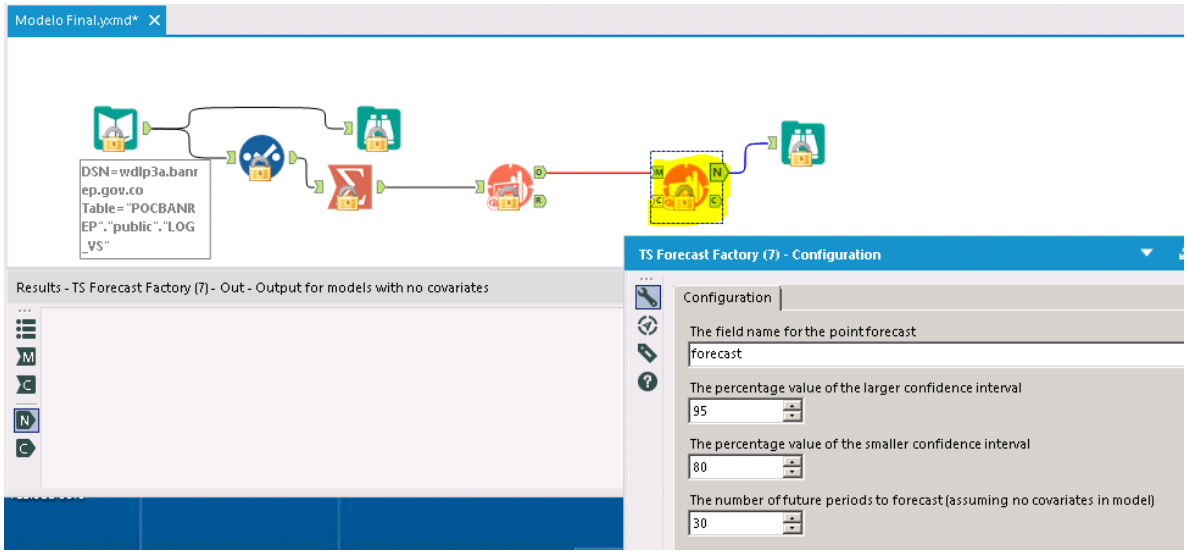


Al revisar las propiedades del modelo ARIMA, al interior se puede apreciar lo siguiente


Modelo Final.yxmd*
TS Model Factory.yxmc*
Alteryx Macro: Time Series Factory
Module Description
Create a Time Series model for multiple different groups using either ARIMA or Exponential Smoothing, depending on the user's selection. Covariate fields may also be used with ARIMA if the user desires.
Select the target field
Select the grouping field
Use covariates in model estimation? (ARIMA only)
Update Value
Update Value
Update Value
Update Value
Update Detour Direction from Question
[#1]

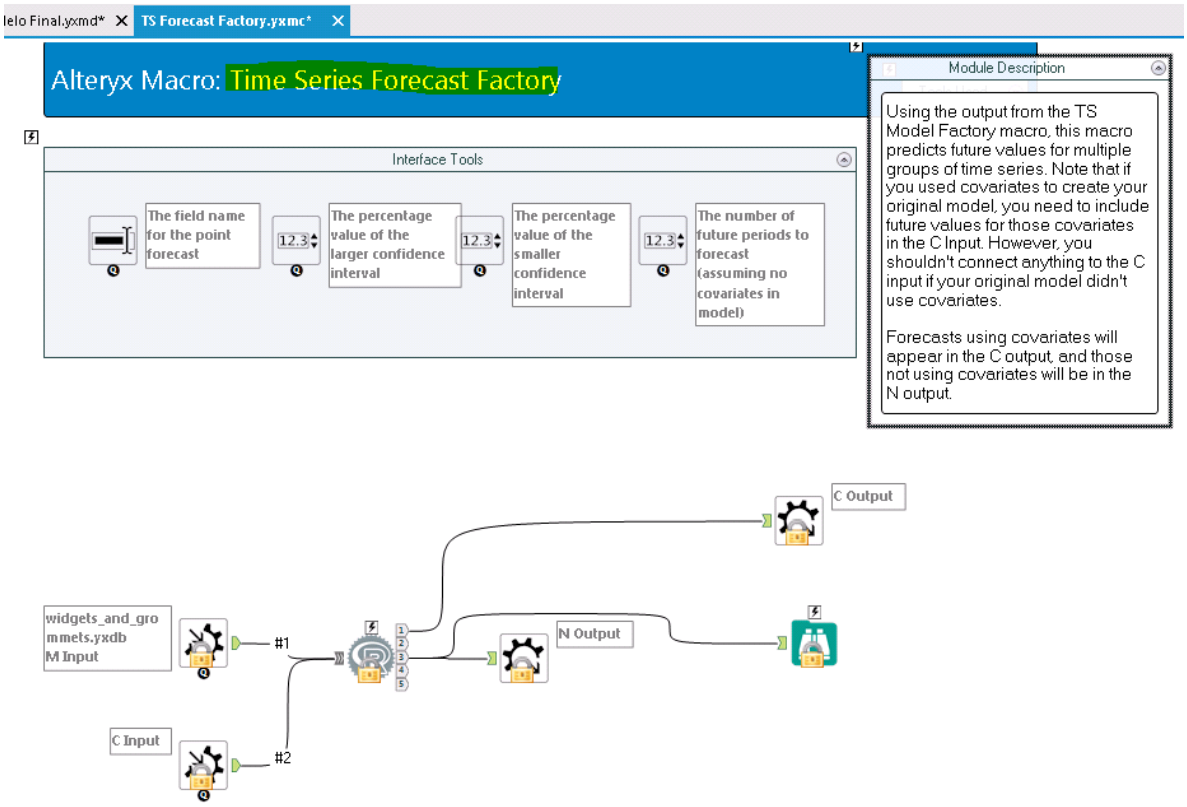

Modelo Final.yxmd*
TS Forecast Factory.yxmc*
TS Model Factory.yxmc*
Update Value
Update Value
Update Value
Update Value
Update Detour Direction from Question
[#1]
Update Value
ARIMA
ETS
ARIMA and ETS (both in native R script)
Input
Detour to the left if covariates aren't being used and to the right if they are.
Using the RecordID tool to keep track of ordering, join the units field and the grouping field.
Using the RecordID tool to keep track of ordering, join the units field, the grouping field, and the covariate field(s).

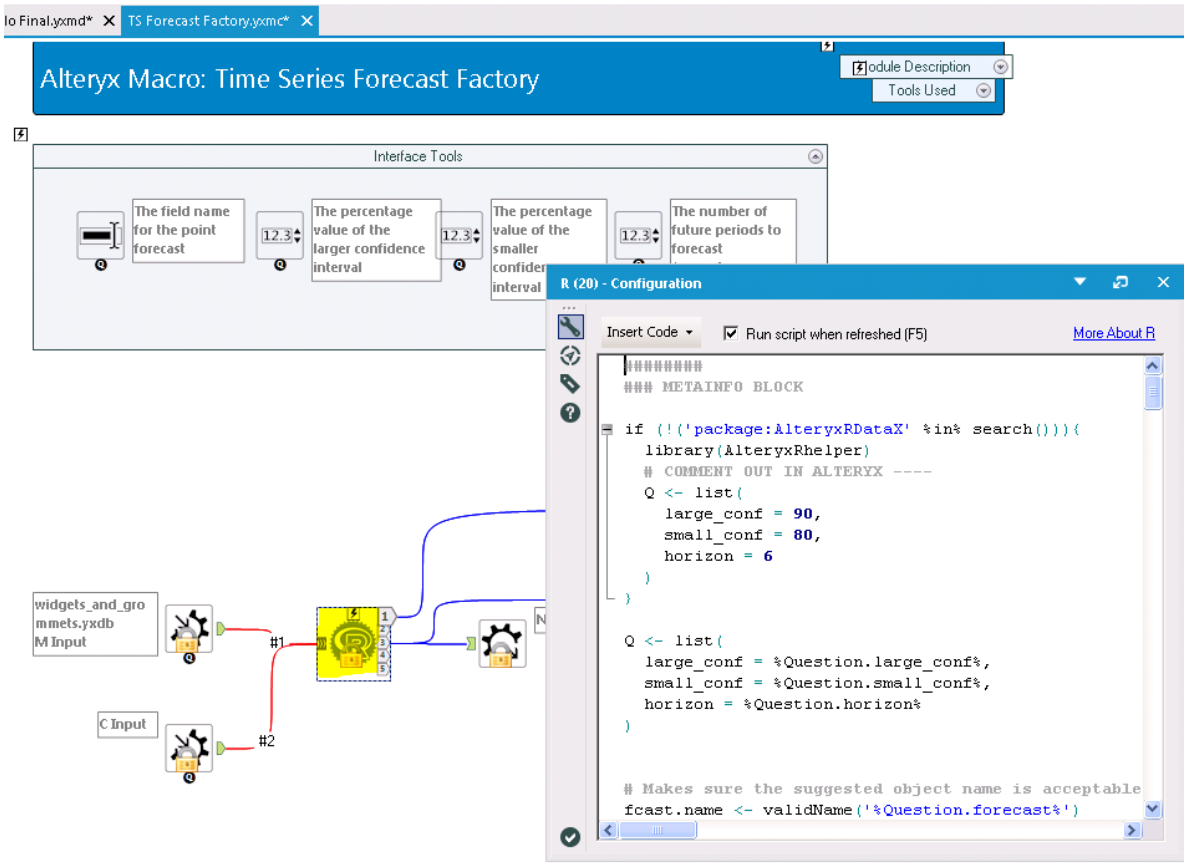
Proceso para definir los intervalos de predicción del modelo



Al revisar las propiedades del modelo predictivo, se aprecia la macro que predice los valores futuros.

En el centro se puede apreciar el modelo “R”


lelo Final.yxmd*
TS Forecast Factory.yxmc*
Alteryx Macro: Time Series Forecast Factory
Module Description
Using the output from the TS Model Factory macro, this macro predicts future values for multiple groups of time series. Note that if you used covariates to create your original model, you need to include future values for those covariates in the C Input. However, you shouldn't connect anything to the C input if your original model didn't use covariates.
Forecasts using covariates will appear in the C output, and those not using covariates will be in the N output.
Interface Tools
The field name for the point forecast
The percentage value of the larger confidence interval
The percentage value of the smaller confidence interval
The number of future periods to forecast (assuming no covariates in model)
C Output
widgets_and_grommets.yxdb
M Input
#1
N Output
C Input
#2



Resultado del modelo

El resultado del modelo ARIMA es un archivo texto, con las diferentes probabilidades.

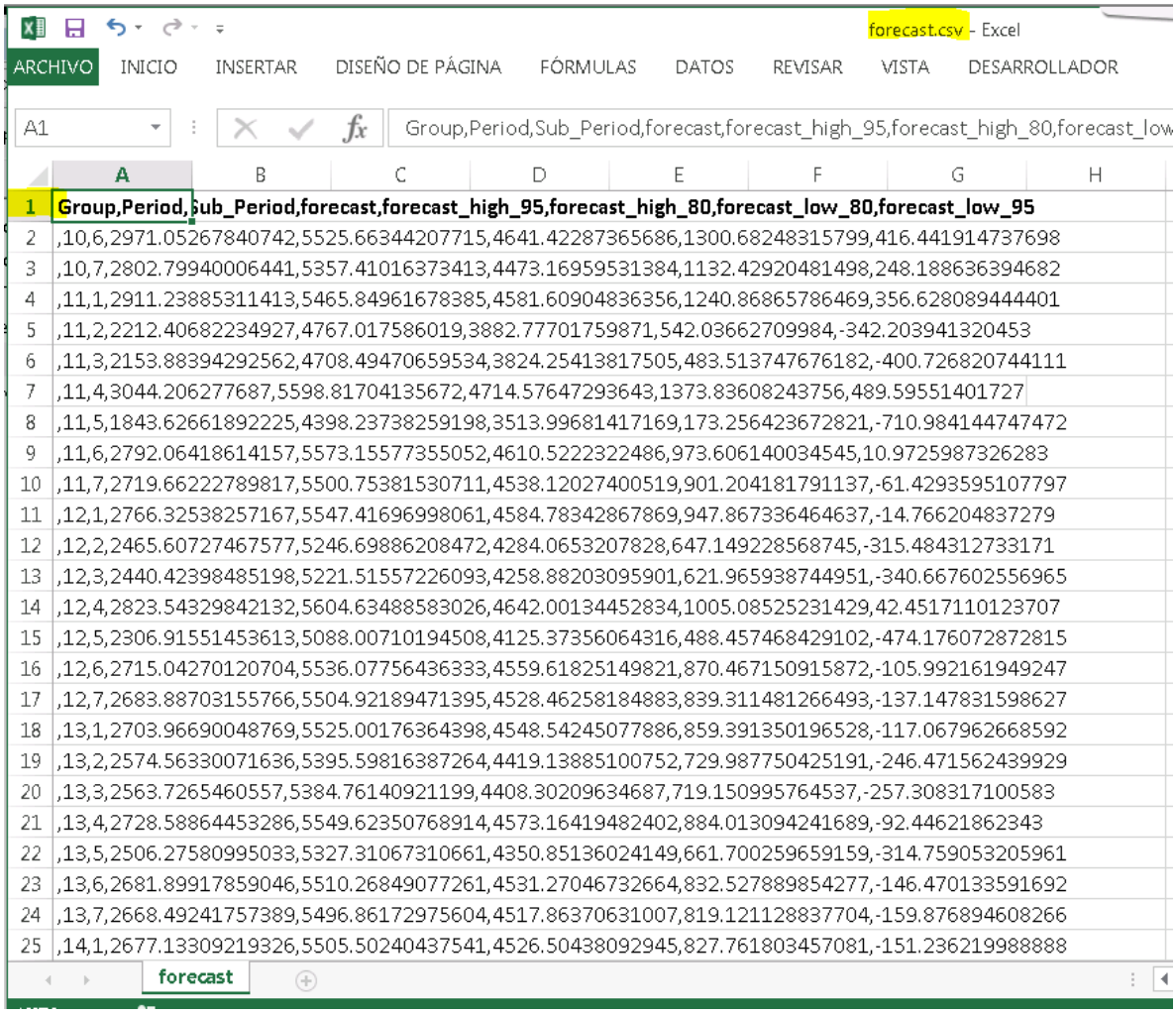
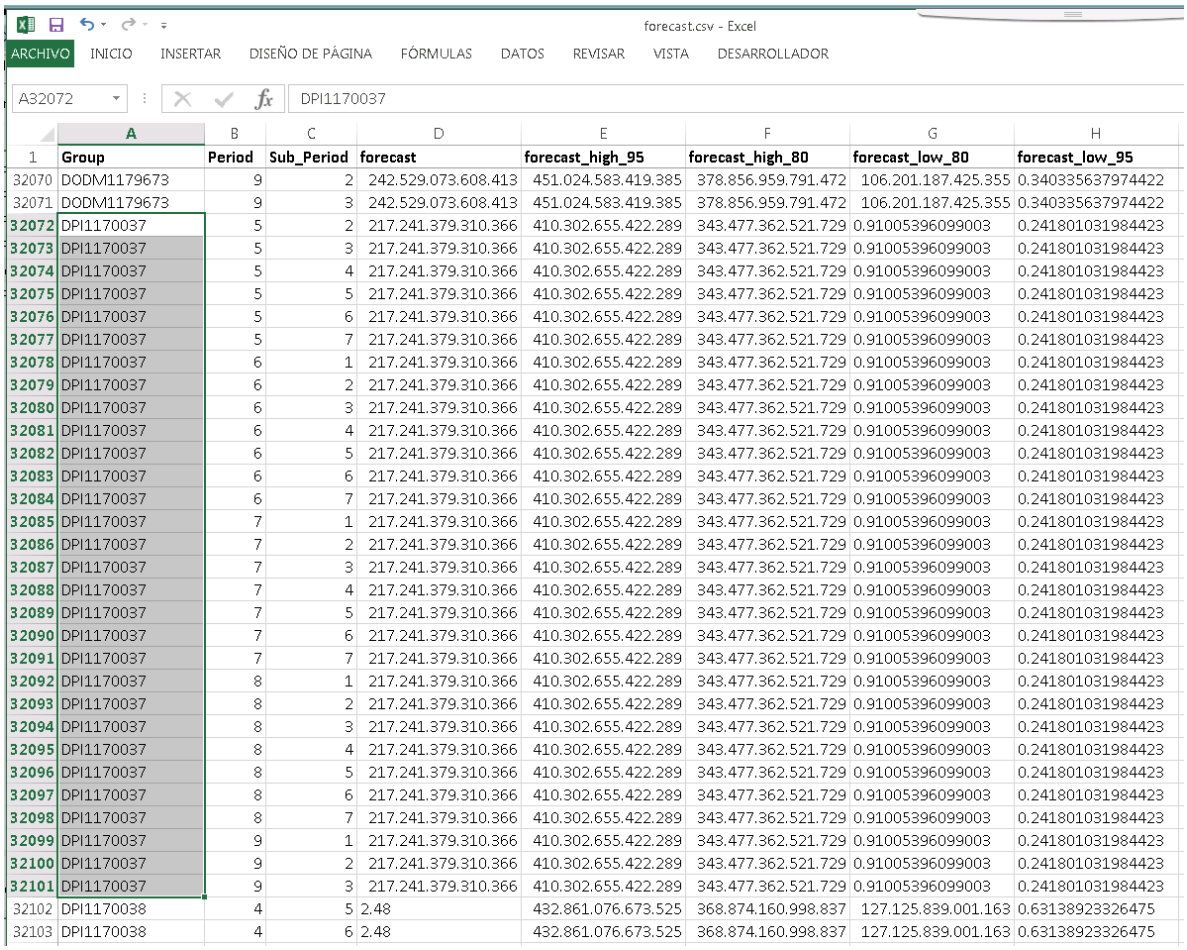
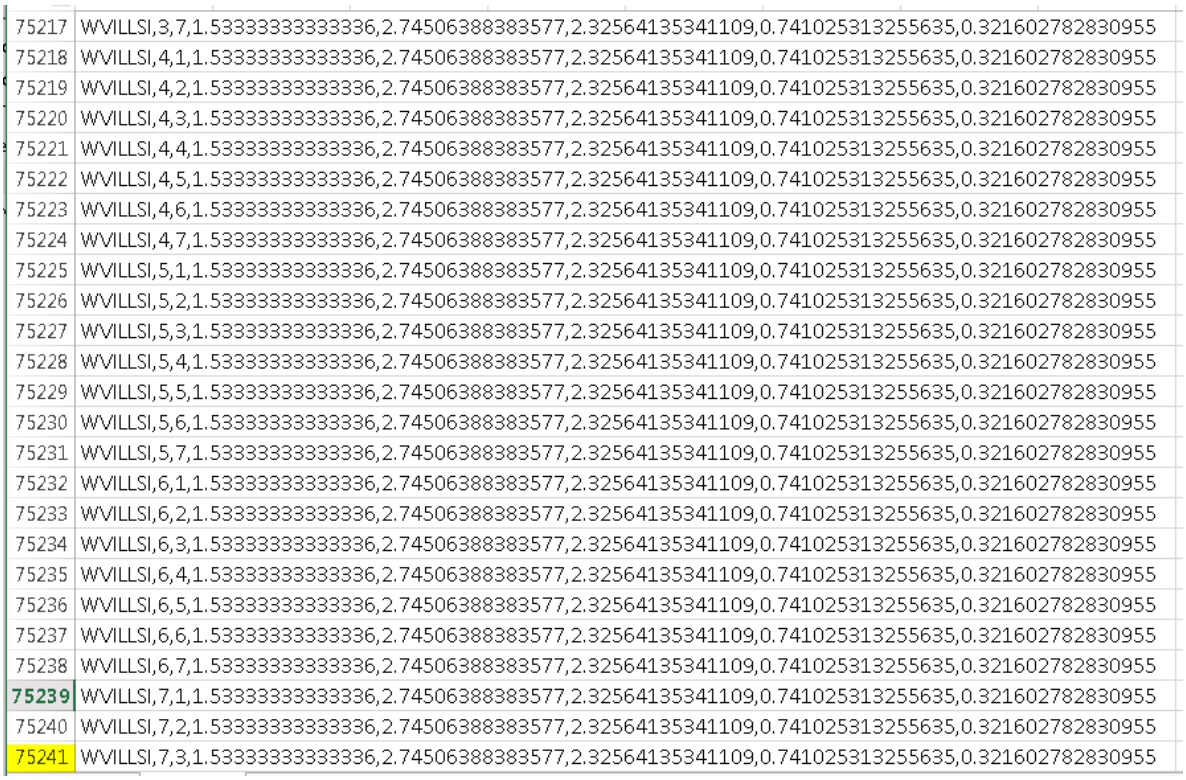


Figura 19 Resultado del modelo ARIMA

A continuación, se muestran los resultados de predicción, para el computador DPI1170037. Luego de 2 semanas, al revisar los eventos del software antivirus, efectivamente se puede apreciar que existieron acciones del antivirus.



	A	B	C	D	E	F	G	H
1	Group	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
32070	DODM1179673	9	2	242.529.073.608.413	451.024.583.419.385	378.856.959.791.472	106.201.187.425.355	0.340335637974422
32071	DODM1179673	9	3	242.529.073.608.413	451.024.583.419.385	378.856.959.791.472	106.201.187.425.355	0.340335637974422
32072	DPI1170037	5	2	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32073	DPI1170037	5	3	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32074	DPI1170037	5	4	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32075	DPI1170037	5	5	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32076	DPI1170037	5	6	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32077	DPI1170037	5	7	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32078	DPI1170037	6	1	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32079	DPI1170037	6	2	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32080	DPI1170037	6	3	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32081	DPI1170037	6	4	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32082	DPI1170037	6	5	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32083	DPI1170037	6	6	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32084	DPI1170037	6	7	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32085	DPI1170037	7	1	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32086	DPI1170037	7	2	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32087	DPI1170037	7	3	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32088	DPI1170037	7	4	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32089	DPI1170037	7	5	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32090	DPI1170037	7	6	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32091	DPI1170037	7	7	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32092	DPI1170037	8	1	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32093	DPI1170037	8	2	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32094	DPI1170037	8	3	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32095	DPI1170037	8	4	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32096	DPI1170037	8	5	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32097	DPI1170037	8	6	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32098	DPI1170037	8	7	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32099	DPI1170037	9	1	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32100	DPI1170037	9	2	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32101	DPI1170037	9	3	217.241.379.310.366	410.302.655.422.289	343.477.362.521.729	0.91005396099003	0.241801031984423
32102	DPI1170038	4	5	2.48	432.861.076.673.525	368.874.160.998.837	127.125.839.001.163	0.63138923326475
32103	DPI1170038	4	6	2.48	432.861.076.673.525	368.874.160.998.837	127.125.839.001.163	0.63138923326475



75217	WVILLSI,3,7,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75218	WVILLSI,4,1,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75219	WVILLSI,4,2,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75220	WVILLSI,4,3,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75221	WVILLSI,4,4,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75222	WVILLSI,4,5,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75223	WVILLSI,4,6,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75224	WVILLSI,4,7,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75225	WVILLSI,5,1,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75226	WVILLSI,5,2,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75227	WVILLSI,5,3,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75228	WVILLSI,5,4,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75229	WVILLSI,5,5,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75230	WVILLSI,5,6,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75231	WVILLSI,5,7,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75232	WVILLSI,6,1,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75233	WVILLSI,6,2,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75234	WVILLSI,6,3,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75235	WVILLSI,6,4,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75236	WVILLSI,6,5,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75237	WVILLSI,6,6,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75238	WVILLSI,6,7,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75239	WVILLSI,7,1,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75240	WVILLSI,7,2,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955
75241	WVILLSI,7,3,1.53333333333336,2.74506388383577,2.32564135341109,0.741025313255635,0.321602782830955

6.7. Presentación de datos

Como se puede apreciar en el anterior punto, se utilizaron dos herramientas: la primera es **alterix**, en donde se puede visualizar los resultados correspondientes a los datos procesados y que es mejor comprendido por personas con conocimiento en matemáticas y estadísticas.

En las Instituciones Financieras, para los roles directivos con necesidades de reportes específicos y de fácil comprensión, es conveniente el uso del software comercial **tableau**, que dispone de grandes beneficios para la presentación de datos.

A continuación, se relacionan las gráficas obtenidas con su respectiva interpretación:

A continuación, se ilustran las posibles predicciones con un intervalo de confianza del 80%, para malware correspondiente a virus y a troyanos.

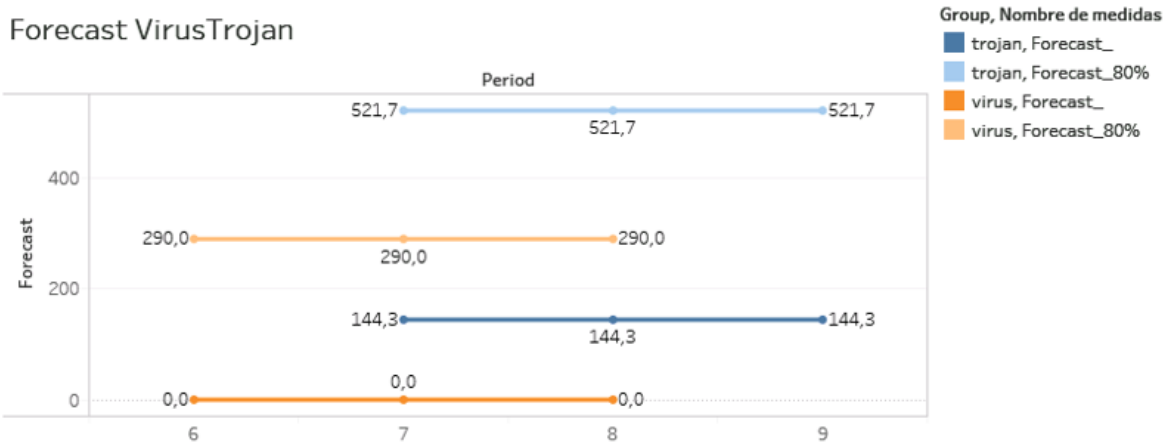
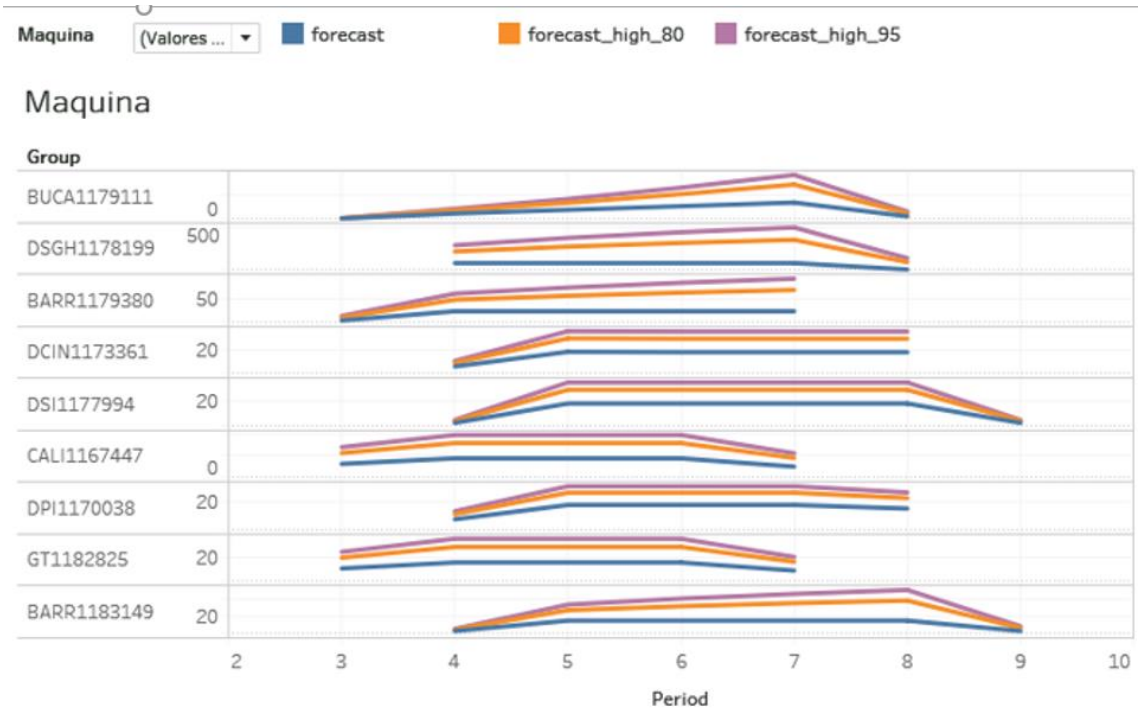


Figura 20 Predicción de Virus y Troyanos

Luego de la ejecución del modelo predictivo ARIMA se obtiene un listado con los computadores en los que posiblemente existirán acciones del software antivirus en los computadores de la Institución Financiera. El modelo predice el nombre del computador con su correspondiente cantidad de acciones. La ubicación física del equipo se puede generar por la deducción del estándar de nombres utilizado para nombrar a los computadores. Para este caso se tomaron los más representativos para la ciudad principal de Bogotá, así como para sucursales de Bucaramanga, Barranquilla, Cali, etc.



Luego de la ejecución del modelo predictivo y al cabo de 2 semanas de tiempo trascurrido, se generó un listado desde la consola de administración del antivirus, con las acciones realizadas sobre el malware.

Nombre del sistema->Hora de generación del ev	Número de Eventos de amenazas
BARR1183149	76
DPI1170038	19
DSGH1178199	14
DCIN1173361	10
BARR1179380	7
CALI1167447	5
GT1182825	5
BUCA1179111	4
DRBL1179329	3
BUCA1173532	2
CALI1156354	2
DCIN1179289	3

Figura 21 Comparación del modelo predictivo vs la realidad

Un aspecto importante de este trabajo *es predecir y obtener un reporte con nombres de computadores en lo que se presentará una infección por malware.* Se realiza la verificación de la predicción versus la realidad y existe un alto grado de correspondencia para los nombres de los computadores.

Los datos son contrastados por la predicción del modelo, donde se aprecia que los eventos por malware van a incrementarse y luego disminuirán hasta un valor de cero. Este comportamiento es normal debido a los controles realizados por el software antivirus debido a que realiza correctamente la funcionalidad de detectar el malware hasta eliminarlo.

El gráfico predictivo es muy importante debido a que con una simple visualización podemos saber que ocurrirá, sobre algunos computadores y su comportamiento. Es de gran ayuda en el caso que el comportamiento no descienda, sino que tenga una tendencia constante o incremental y en este caso se dispone de información importante para poder tomar decisiones con el fin de evitar la infección sobre esos computadores, como por ejemplo: incrementar el número de revisiones baja demanda del antivirus, adicionar controles sobre algunos tipos de archivos y carpetas, actualizar parches sobre los computadores posiblemente afectados, monitorear exclusivamente los equipos relacionados, informar al área del negocio (usuario final) que debe estar pendiente y avisar de cualquier síntoma o acciones inusuales en su computador, etc.

Luego de varias actividades documentadas en este trabajo, es de gran importancia la gráfica anterior debido a que se dispone de un entregable real, con un tiempo de ejecución casi en tiempo real y consiste en un reporte de 1

sola hoja en la que se relaciona la siguiente predicción: ***los computadores que posiblemente se van a infectar con malware***.

Al llegar a este punto, se ha trabajado en lograr obtener un entregable que sirva para apoyar la hipótesis de este documento, que intenta probar una suposición sin la restricción que sea verdadera o falsa.

Hasta el momento la predicción se ha comprobado y corresponde con la realidad, es decir: se conoció con 2 semanas de anterioridad lo que posiblemente iba a ocurrir.

El siguiente tipo de interpretación y que corresponde a una interpretación complementaria, se relaciona con el número de eventos. Para este caso de estudio, el número de acciones tiene un efecto secundario debido a que pueden:

1) Acercarse a la realidad

2) Encontrarse no adecuadas a la realidad.

Para el primer caso, relacionado con las predicciones del número de eventos cercanos a la realidad tenemos el ejemplo del computador identificado con el nombre **DPI1170038**, donde el número de amenazas que se predijo fue de 14.9, con un intervalo de confianza del 80%, llegando a 22 acciones realizadas por el antivirus.

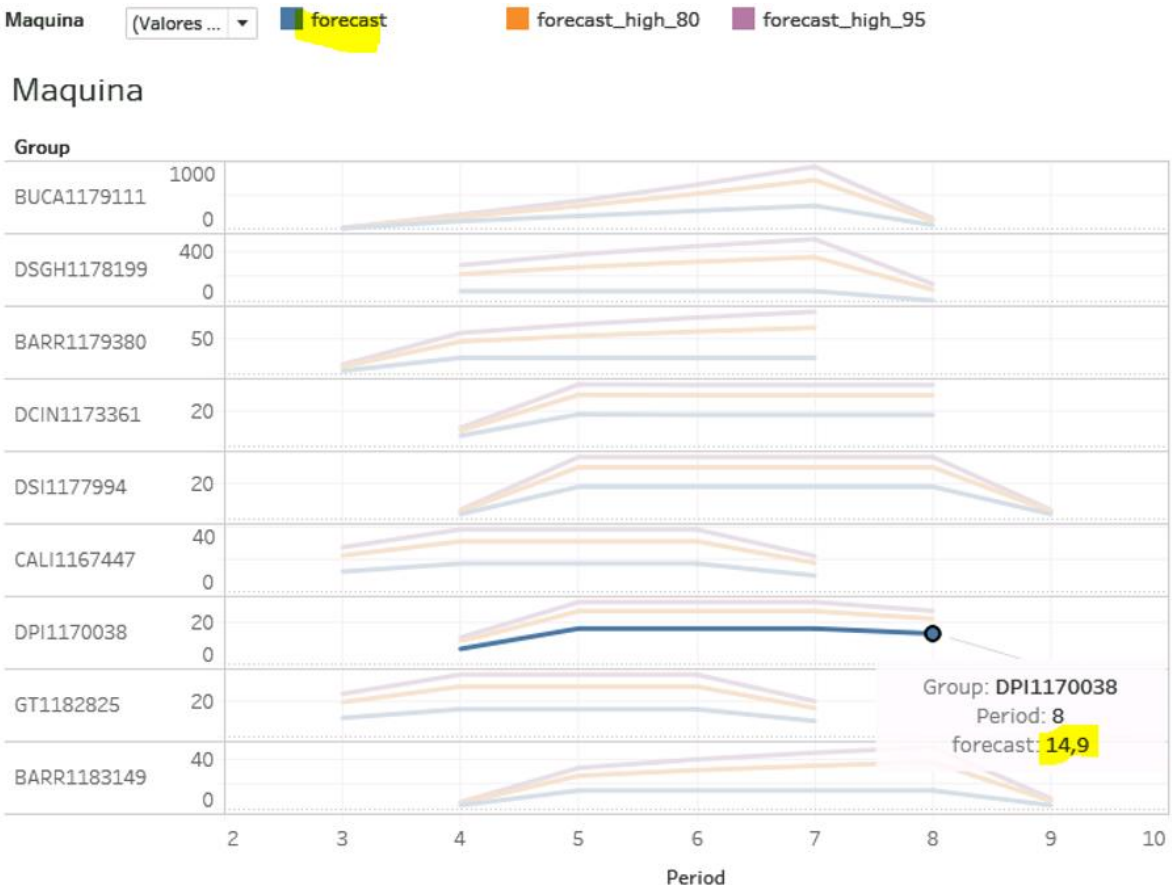


Figura 22 Predicción Acertada, para número de eventos

Lo reportado por el sistema antivirus (consola de administración del ePolicy) fue de 19 acciones.

Nombre del sistema->Hora de generación del ev	Número de Eventos de amenazas
BARR1183149	76
DPI1170038	19
DSGH1178199	14
DCIN1173361	10
BARR1179380	7
CALI1167447	5
GT1182825	5
BUCA1179111	4
DRBL1179329	3
BUCA1173532	2
CALI1156354	2
DCIN1179289	3

Figura 23 Ocurrencia Acertada, para número de eventos

Es importante mencionar en este punto, que la hipótesis de este trabajo, se valida con el anterior par de gráficas donde se muestra un modelo que predice sobre que computadores se presentaran acciones del antivirus y se puede comprobar un par de semanas que así sucedió.

Para el segundo caso, se refiere a un número de acciones en donde el pronóstico no se acerca a la realidad, como por ejemplo a la predicción de malware para un computador ubicado en la ciudad de Bucaramanga, identificado con el nombre **BUCA1179111**, en donde el pronóstico fue de 351 eventos, pero en realidad solo se presentaron 4 eventos.

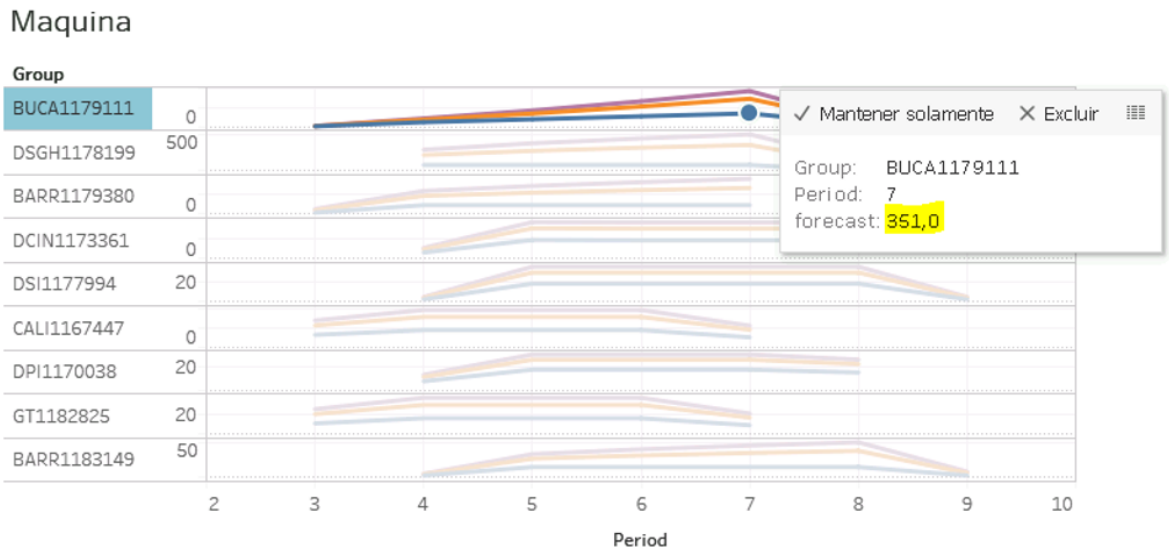


Figura 24 Predicción Desfasada, para el número de eventos

Nombre del sistema->Hora de generación del ev	Número de Eventos de amenazas
BARR1183149	76
DPI1170038	19
DSGH1178199	14
DCIN1173361	10
BARR1179380	7
CALI1167447	5
GT1182825	5
BUCA1179111	4
DRBL1179329	3
BUCA1173532	2
CALI1156354	2
DCIN1179289	3

Figura 25 Ocurrencia Desfasada, para el número de eventos

Para este caso los valores de predicción relacionados con acciones del antivirus no corresponden. Esto puede deberse a que no tenemos suficientes fuentes de datos ni correlación de eventos. Es importante resaltar que el modelo efectivamente “si” predijo que iba a ocurrir algo con este computador.

Como se evidencia en las gráficas obtenidas como el proceso de analíticas, para los computadores que fueron infectados por malware, se puede apreciar que todas tiene un comportamiento de descenso y la tendencia es de cero. Para verificar esta predicción se genera un reporte en la consola de Administración del Antivirus, para mediados del mes de diciembre de 2016, que muestre las acciones del antivirus para las máquinas antes infectadas y el sistema informa que para todos los computadores no hay eventos de malware.

El proceso de gestión de gestión de malware, es apoyado por los análisis predictivos, debido a que el modelo informa las posibles infecciones de computadores de forma anticipada y esto nos permite realizar acciones encaminadas monitorear de forma más puntual los computadores específicos que se verán comprometidos. En caso de tratarse de computadores de funciones críticas para la institución Financiera, se pueden reforzar las medidas de seguridad mediante controles en la red interna o en las propias aplicaciones. A manera de ejemplo, es muy riesgoso la presencia d posibles virus en computadores de transacciones operativas a nivel nacional o internacionales, en especial cuando se acercan las fechas de cumplimiento de obligaciones.

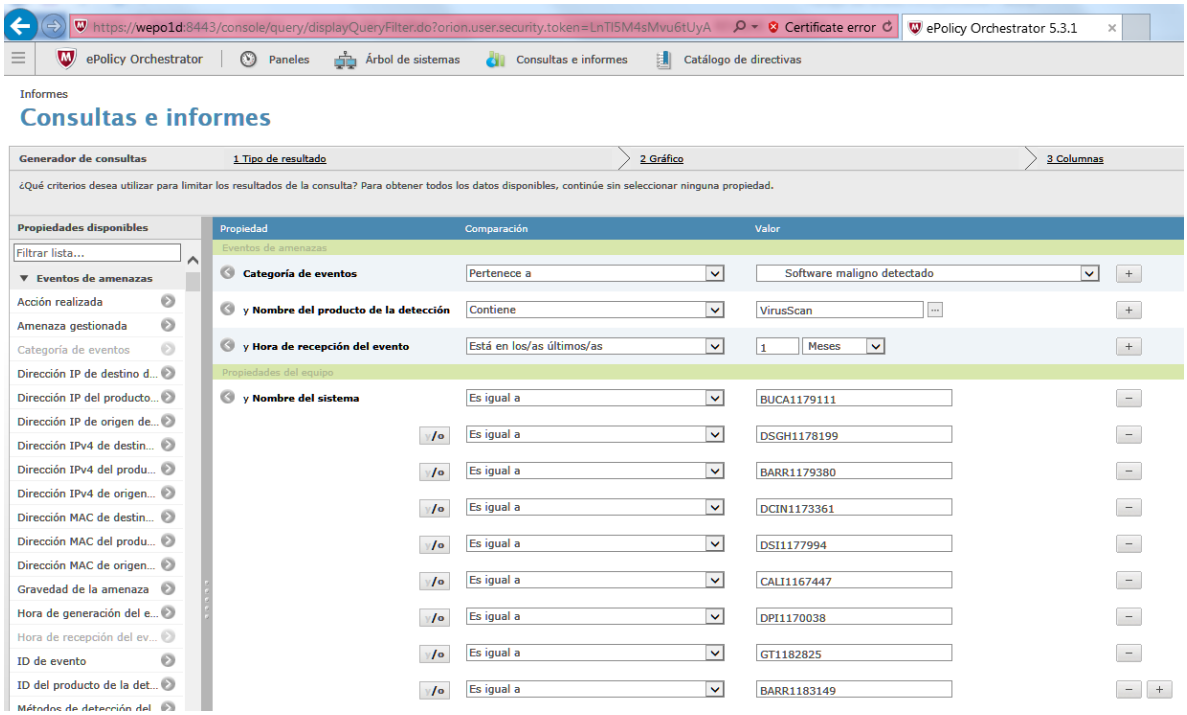


Figura 26 Configuración del reporte para verificar la no presencia de virus

El resultado apoya la predicción del modelo Arima, en donde las gráficas muestran una disminución de las acciones del antivirus, hasta llegar a cero (0).

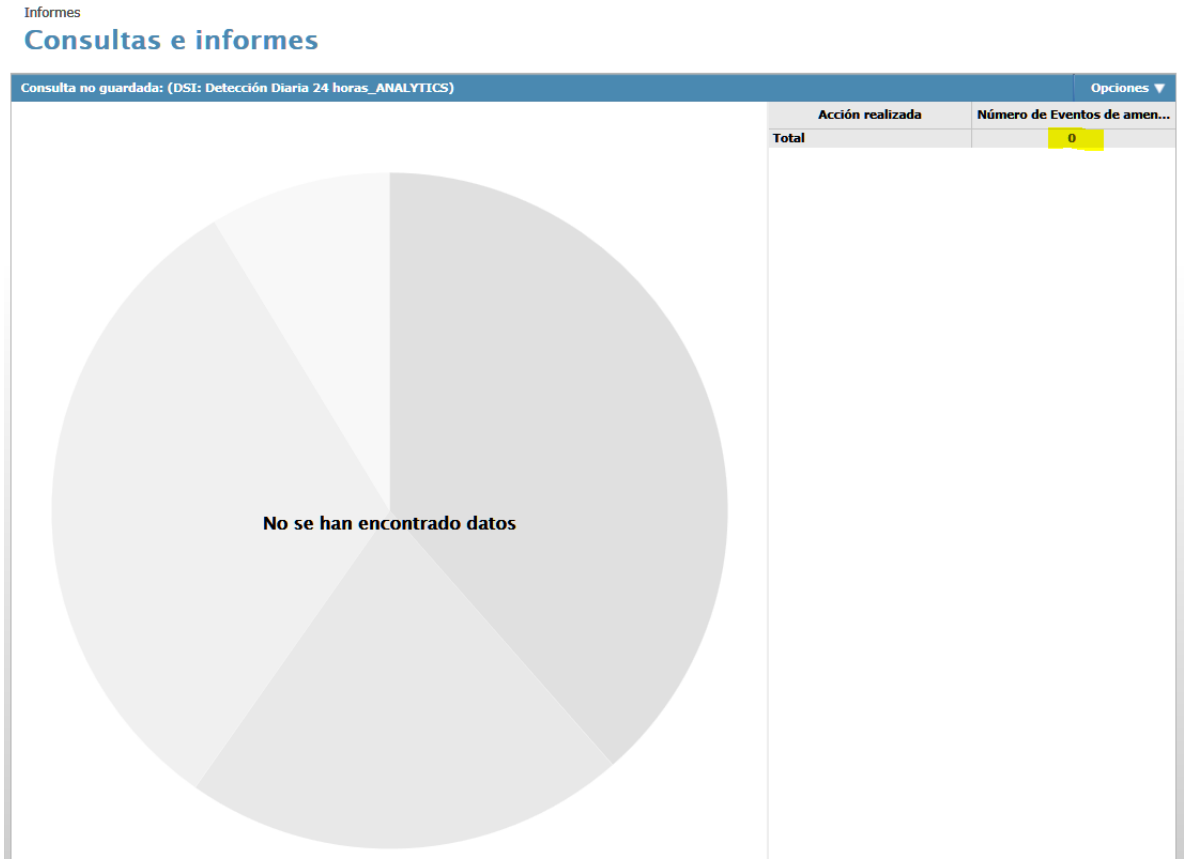


Figura 27 Verificación que no continuaron eventos de infección

Luego de realizar las actividades mencionadas en este capítulo 6 para la Prueba de Concepto y al obtener resultados exitosos, se coordinó una reunión conjunta entre el proveedor de la solución informática - itPerfom y el Área de Seguridad Informática, con el objetivo de presentar el trabajo desarrollado y los resultados a la Jefatura de Protección de la Información, del Departamento de Seguridad Informática. Se muestran los procesos de cargue de datos, proceso de Analytics y resultados.

Presentación de la propuesta a Área de Seguridad Informática

Como consecuencia de la reunión se logra la aprobación para continuar con las actividades de Analytics, para el primer semestre del año 2017, a través de nuevos proyectos mediante las siguientes alternativas:

L*a primera*, es retomar el modelo de análisis de datos actual correspondiente a los eventos de malware, con el fin de adicionar más fuentes de datos internos y externos. Para los datos internos se pueden agregar al modelo eventos generados por los sistemas de control de navegación por Internet; sistemas antispam, que controlan el correo basura; la herramienta FireEye, controla eventos no controlados por los sistemas tradicionales de seguridad informática; y los eventos reportados por un proveedor externo que realiza monitoreo 7* 24. Para el análisis de los datos externos, las herramientas actuales disponen funcionalidades de obtener datos de las diferentes redes sociales, así como consultar páginas en Internet y obtener los datos deseados, para el caso de los fabricantes de anti-malware. Lo anterior, para logra una mayor precisión en la predicción del modelo.

La segunda alternativa, corresponde a iniciar nuevos proyectos de “Analytics”, con un par de casos de uso más estructurados, con datos más estructurados. La primera opción es determinar la prioridad para aplicar parches o actualizaciones de software correspondientes a sistemas operativos y/o aplicaciones, en los computadores de escritorio y servidores. La segunda opción es predecir el uso de la red de datos, a partir de flujos de red.

Los dos nuevos proyectos, se originan por el interés despertado por ingenieros del Área de Seguridad Informática, al enterarse del desarrollo este Trabajo de Grado y algunas explicaciones puntuales que se les ofreció, ante unas necesidades para poder resolver un problema que actualmente tienen y no ha sido posible resolverlo.

6.8. Restricciones para la Prueba de Concepto

Para realizar esta Prueba de Concepto, se presentaron las siguientes restricciones en las siguientes áreas:

Para los *Datos*: En la Institución Financiera se disponen de varias herramientas informáticas que generan eventos de log, los cuales con enviados y centralizados en un recolector de eventos. El resultado de esta operación es que en promedio llegan más de 1.200 millones de eventos a la semana. Debido alcance limitado de la Prueba de Concepto, solo se trabajarán con datos relacionados con las actividades realizadas por el software antivirus.

El lapso de tiempo de eventos al SIEM para el antivirus es relativamente corto, debido a que se implementó hace poco la replicación de datos hacia el SIEM, con una disposición de solo 68 días de historia y representa cerca de 500.000 registros.

Cuando se habla de grandes volúmenes de información, 500.000 registros parecen algo reducido, pero es importante mencionar que es necesario afinar el modelo y se lograra un mejor porcentaje en la predicción, al tener datos de varios años y de igual forma datos externos a la Institución financiera, principalmente proveniente de redes sociales y diferentes páginas de los fabricantes y consultores de seguridad.

Para coordinar los *trabajos con el proveedor*, fue necesario tener charlas informales con representantes de los fabricantes: Microsoft, Hewlett Packard e IBM. Con dos fabricantes no fue posible coordinar la prueba de concepto, debido a que: primero, un fabricante enviaba hacia la nube los datos a ser analizados, el segundo no tenía disponibilidad de tiempo porque sus ingenieros estaban en otros proyectos

y se acercaba la temporada de fin de año. Adicionalmente, no existía ningún tipo de relación comercial, a pesar de las expectativas.

En relación con el *análisis de datos,* la Institución Financiera especifico que los análisis se realizaran en la plataforma computacional instalada en sitio. Por lo anterior fue necesario que el proveedor (itPerform) representante de Hewlett Packard, asignara un par de ingenieros para trabajar en las oficinas de la Institución financiera durante algunos días.

Para realizar la prueba de concepto, solo de dispuso de licencias de prueba para el software de alterix, tableau y Vertica, para un sistema operativo tipo Server y que fue prestada por el proveedor. Esta licencia solo se encontraba activa por 14 días.

El *recurso humano* disponible en la Institución Financiera no disponía de experiencia en los conceptos de Big Data y Analytics; así mismo, las actividades se realizaron en horario laboral y sin la dedicación exclusiva para la Prueba de Concepto, por lo que la curva de aprendizaje fue algo lenta.

7. CRONOGRAMA DE ACTIVIDADES Y PRESUPUESTO

Fecha de inicio: 1 de Junio de 2016

ACTIVIDAD	Marzo Mayo	Jun 1	Jun 15	Jul 1	Jul 15	Ago 1	Ago 15	Sept 1	Sept 15	Oct 1	Oct 15	Nov 1	Nov 15	Nov 30	Dic 15
Identificar problemática e investigar posibles soluciones. Documentar propuesta															
Presentar propuesta a jurado y realizar modificaciones según las recomendaciones															
Investigar, analizar y documentar el caso específico de la entidad financiera															
Explicar el proyecto al área de seguridad informática de la institución financiera															
Desarrollar del marco teórico del estudio															
Recolectar resultados de estudios realizados por fabricantes y empresas															
Prueba de Concepto con Fabricante															
Procesamiento de datos															
Análisis de resultados y elaboración del informe final															
Redacción de Informe final para revisión.															

Figura 28 Cronograma

7. PRESUPUESTO DE INVERSION

El siguiente cuadro ilustra el presupuesto de inversión requerido para la realización de este proyecto de investigación.

Rubros	**Fuentes de financiamiento**		
	Recursos Institución Financiera	**Recursos Propios**	**Total**
1. Personal (1/4 tiempo x 6 meses	$ 6.000.000	$ 0	**$ 6.000.000**
2. Libros		$ 300.000	**$ 300.000**
3. Otros	$ 500.000	$ 200.000	**$ 700.000**
Total	**6.500.000**	**$ 500.000**	**$ 7.500.000**

Figura 29 Presupuesto

8. PRODUCTOS GENERADOS

A continuación, se relacionan los productos que se generan a partir del trabajo realizado:

- Documento de Trabajo de Grado, titulado: *"Uso de Analíticas para predecir los computadores afectados por malware, en una institución financiera en Colombia, 2017",* en el cual se relaciona y justifica las actividades desarrolladas para la obtención del título de la Maestría de Gestión de la Información, en la Escuela Colombiana de Ingeniería Julio Garavito.

- Diseño e implementación de la Prueba de Concepto para predecir que computadores se van a encontrar afectados por intentos de infección de malware, a partir de las acciones realizadas en el pasado por el software antivirus.

- Reporte con resultados de la Prueba de Concepto, que contiene la lista de posibles computadores en los que existirá actividad maliciosa y cantidad aproximada de acciones del software antivirus, para ser presentada al Área de Seguridad Informática de la Institución Financiera.

- Autorización para continuar con el desarrollo de un "caso de uso" más estructurado relacionado con Analytics, para el año 2017.

9. APORTES DEL AUTOR

Los aportes del autor de este trabajo se relacionan a continuación:

Primero, realizar un Estado del Arte, con el fin de realizar un acercamiento al tema de Big Data y Analytics, a través de consultas de diferentes publicaciones en Internet, así como en medios físicos, con la grata sorpresa que existe gran cantidad de información al respecto, en diferentes sectores de la economía documentados con cifras y soportados con casos de uso que representan una gran utilidad para

las empresas. En el sector de seguridad informática, algunos proveedores de soluciones de seguridad están desarrollando algunos aplicativos y se encuentran en el proceso de pruebas y comercialización.

No se encontró información documentada para el caso específico de poder predecir los computadores que posiblemente se van a infectar en una Institución Financiera. Esto se puede deber a, la suposición, que son actividades internas de cumplimiento y de gestión de incidentes de seguridad que son tratados internamente. Adicionalmente, en Colombia y en otros lugares, no existe la cultura de publicar los incidentes de seguridad informática debido a las consecuencias negativas que esto conllevaría a las organizaciones, como las pérdidas directas e indirectas, en especial el aspecto económico y la pérdida de credibilidad y confianza.

El segundo aporte consistió en poder identificar el tipo de amenaza informática que más afecta a las organizaciones y delimitar el trabajo para encontrar mecanismos informáticos para solucionar el problema, que continua activo con el trascurrir del tiempo, de tal forma que se “infectan los computadores” con malware, a pesar de tantos años de investigación y desarrollos en este campo.

Como tercer aporte, conviene mencionar que entre los diferentes tipos de análisis de datos (diagnostico, reactivo, preventivo, predictivo y prescriptivo), se identificó que el tipo de análisis “predictivo” por series de tiempo ARIMA, ayuda a identificar eventos futuros a partir de eventos pasados. Los análisis predictivos con fundamentación matemática y estadística, son incorporados a soluciones informáticas.

El cuarto aporte consistió en identificar los componentes tecnológicos requeridos para realizar un análisis predictivo. Un aspecto importante a mencionar es que se verificó el modelo diseñado en papel mediante una Prueba de Concepto,

mediante el uso de datos relacionados con las acciones del antivirus, obtenidos de un ambiente de producción. Como producto de la Prueba de Concepto se obtuvo un reporte donde “predice” los computadores en los que posiblemente existirán eventos relacionados con malware. Se realiza una verificación del reporte obtenido, permitiendo que transcurran un par de semanas y efectivamente, sobre los computadores informados anteriormente, se registran eventos por malware.

Al obtener un reporte de computadores que posiblemente se verán afectados por malware, brinda un valor agregado a las diferentes áreas de una Institución Financiera, por ejemplo: para el Área de Tecnología, apoya a los diferentes ingenieros para poder tomar acciones preventivas y diseñar diferentes mecanismos para monitorear un posible incidente de seguridad informático. En un modelo tradicional hay que esperar a que suceda el evento y luego actuar para encontrar una solución al problema que en muchas ocasiones puede demorar varios días. Para el Área de Negocio, es posible distribuir las funcionalidades operativas de las aplicaciones, en otros computadores evitando las consecuencias directas e indirectas mencionadas al comienzo de este trabajo. A nivel general de la organización, se aumenta considerablemente la percepción de seguridad informática de todos los empleados, debido a que ellos consideran al Área de Seguridad de Información, como un aliado estratégico para la protección de sus activos de la información, que utiliza herramientas de vanguardia y con capacidades de responder a nuevas preguntas que antes no era posible.

El quinto aporte es lograr la sinergia de varios aspectos independientes como: la experiencia y conocimientos de varios años de actividades laborales relacionadas con en Malware, logrando incorporar nuevos conceptos y herramientas de Big Data y Analytics, con el fin de obtener un reporte de computadores en lo que sucederán eventos.

Considero como más importante, el aporte número seis, consistente en que este trabajo sirve como "*semilla*", para futuros casos de uso relacionados con Analytics y Big Data, en una Institución Financiera. Luego de mostrar los resultados de la POC a los directivos del Área de Seguridad Informática, es muy reconfortante, escuchar buenos comentarios y en especial obtener el aval y visto bueno, para continuar con esta labor para otros casos de uso, cada vez más elaborados, durante el año 2017.

10. TRABAJOS FUTUROS

Debido a que se logró el interés y el apoyo del Área de Seguridad de Información en la Institución Financiera, se disponen de 2 alternativas:

La primera alternativa es estructurar el caso inicial para la detección de malware, incorporando elementos adicionales al modelo predictivo, como los siguientes:

- Fuentes de datos internos, como, por ejemplo: registros de navegación, anti spam, FireEye, servicio de monitoreo 7*24.
  Adicionalmente, más registros de datos correspondientes a aplicaciones diferentes al antivirus, como: la verificación de autenticidad de páginas libres de malware, a través de la herramienta de SiteAdvisor; módulos para prevenir la fuga de información confidencial para los datos en movimiento, en uso y en reposo; sistemas de detección de intrusos (HIPS), etc.
- Fuentes de datos Externos: Datos de redes sociales en donde se hagan comentarios relacionados con malware, páginas de los fabricantes de soluciones de seguridad informática, en las cuales realicen comentarios sobre las características y efectos de malware, etc.
  En la siguiente ilustración se identifican los nuevos componentes del modelo.

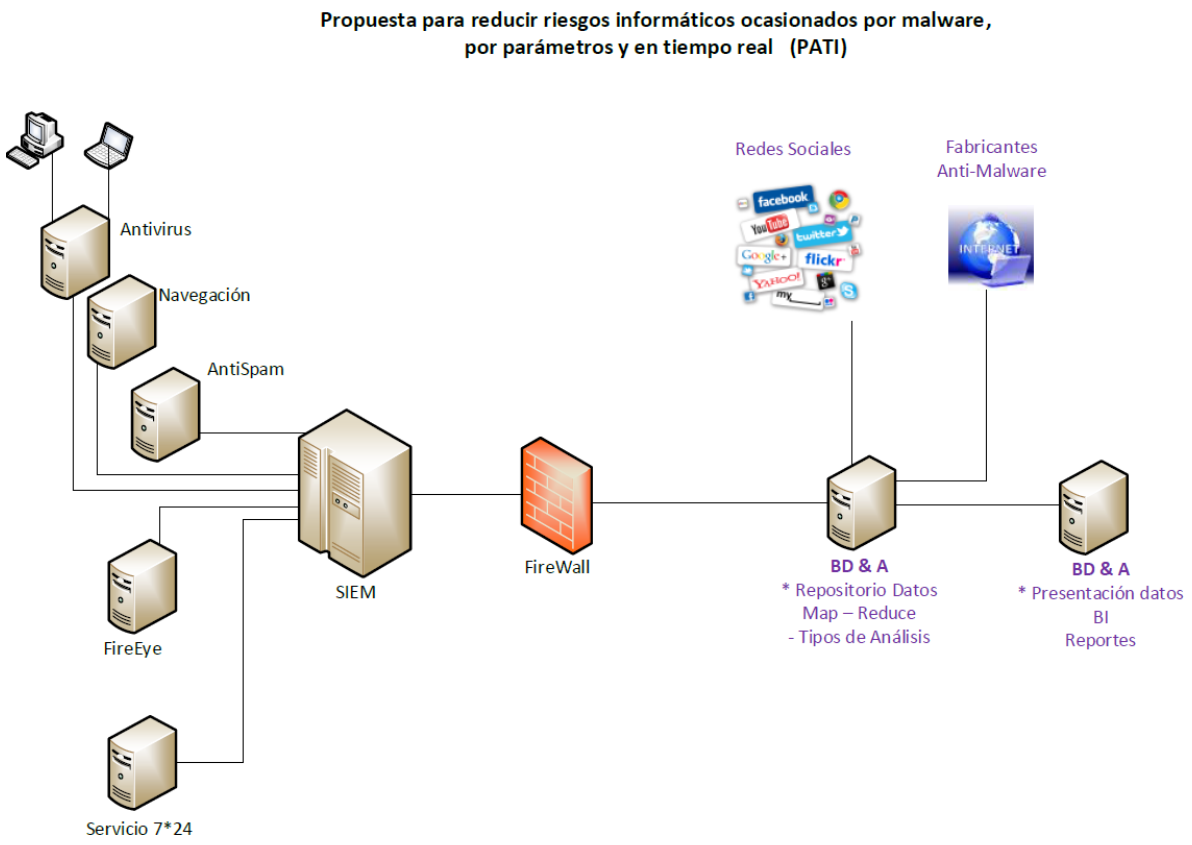


Figura 30 Propuesta: Adicionar fuentes de datos al modelo

La segunda alternativa es diseñar un “caso de uso” diferente. Luego de mencionar algunas funcionalidades relacionadas con Analytics, un par de ingenieros del área de Seguridad informática de la Institución Financiera, se interesaron en este tema y consideran que puede ser un apoyo importante para solucionar algunos inconvenientes para los siguientes procesos:

1) Determinar la prioridad para aplicar parches en los computadores de escritorio.

Como se mencionó al comienzo de este trabajo, una causa por la cual se infectan de malware los computadores es debido a que personas inescrupulosas, aprovechan las vulnerabilidades de “día cero”. Este ataque consiste descubrir en los sistemas operativos y aplicaciones algún fallo en la programación con el fin de causar efectos no esperados en el sistema, a pesar de disponer de los mecanismos de seguridad actualizados. La solución es poder analizar diferentes variables de un entorno corporativo y priorizar los computadores más delicados del negocio para poder aplicar los parches de una forma segura.

2) Análisis de flujos de red.

   Actualmente se dispone de una herramienta que mide la utilización del canal de la red de datos que fue utilizado, es decir se realiza un análisis de tipo de diagnóstico. Se desea conocer una predicción de cuál será el ancho del canal utilizado en días futuros, ¿para coordinar una excelente disponibilidad de los diferentes servicios que son prestados por la Institución Financiera?

11. CONCLUSIONES

En el año 2016, existen conceptos y tecnologías de la información que permiten realizar predicciones con el fin de apoyar y brindar nuevas soluciones a las necesidades de las organizaciones de una forma más eficiente. Esto es apoyado por varias publicaciones que muestran la obtención de resultados exitosos, en una gran variedad de casos de uso en diferentes sectores de la economía a nivel mundial.

Las herramientas de Big Data y Analíticas se enfocan en los datos, cobrando una importancia relevante en el conocimiento de datos de cada negocio y su respectiva interpretación.

Las soluciones tradicionales de seguridad no ofrecen una protección completa contra los nuevos desarrollos de malware especializado. Por tal motivo es necesario el uso de conceptos y la implementación de herramientas de Big Data y Analytics, que pretenden realizar diferentes tipos de análisis a través de modelos estadísticos y con grandes cantidades de datos.

Para la elección del modelo de análisis predictivo, se tuvo en cuenta la documentación relacionada en este trabajo, incluyendo los modelos como Arima, Árboles de Decisión y Naive Bayes. Adicionalmente, la recomendación realizada por el proveedor con experiencia en software de Analytics y Big Data. Para la Prueba de Concepto se decidió utilizar el modelo Arima, obteniendo buenos resultados. Conviene mencionar que el alcance de este trabajo no es realizar la comparación de resultados entre modelos predictivos.

La Prueba de Concepto realizada en la Institución Financiera, se configura e implementa en la plataforma computacional disponible, con los datos de producción del software antivirus y con el software de prueba facilitado por el proveedor itPerform.

Se logra alcanzar el objetivo general enunciado en este trabajo: **Predecir** mediante el uso de herramientas analíticas tecnológicas y en un corto plazo de tiempo, los posibles computadores que pueden ser "infectados" por software maligno. La verificación es realizada mediante un producto consistente en un listado con nombres de computadores que posiblemente se infectaran.

En el reporte existe la variable secundaria *Cantidad de Acciones,* y que puede ser más acorde con la realidad mediante el ingreso de orígenes de datos internos y externos al modelo, para obtener una mayor precisión, en cuanto al número de acciones. Esta situación se puede contemplar como un caso de uso diferente al enunciado en este trabajo, para ser realizado como trabajo futuro.

El término importante en este trabajo es "predecir" y de acuerdo a lo expresado por (Definición.de, s.f.) "En el ámbito de la ciencia, una predicción es un anticipo de lo que ocurrirá de acuerdo al análisis de las condiciones existentes. Es frecuente que las predicciones surjan tras experimentos o investigaciones que permiten conocer las condiciones y estimar que, si se repiten, el resultado será el mismo. Las predicciones científicas, sin embargo, no siempre se cumplen ya que suelen existir variables desconocidas u otras cuya dinámica no se puede anticipar con precisión". Por lo anterior vale la pena mencionar que en este trabajo la predicción es un acercamiento a la realidad y que puede ser más exacta de acuerdo al conocimiento utilizado para identificar los datos que se ingresan a un modelo estadístico.

Es necesario dar un manejo serio y prudente en la interpretación y divulgación de este tipo de trabajos, debido a que pueden causar falsas expectativas y pánico generalizado en el sector financiero; afectando el buen nombre de una Institución Financiera.

EQUIPO INVESTIGADOR

Gerardo Mayorga García. Candidato a Master en Gestión de información, Escuela Colombiana de Ingeniería, Bogotá, Colombia. Ingeniero de sistemas de la Universidad Central. Ingeniero especializado del Departamento de Seguridad informática del Banco de la República de Colombia.

REFERENCIAS BIBLIOGRAFICAS

Aguilar, L. J. (2013). *Big Data Análisis de Grandes Volumenes de Datos en Organizaicones.* México: Alfaomega.

BBC. (15 de Marzo de 2016). *BBC Mundo*. Obtenido de http://www.bbc.com/mundo/noticias/2016/03/160315_economia_robo_banco_central_bangladesh_ac

Bernal, C. (2016). *Metodología de la Investigación.* Bogotá: Pearson.

Biega, R. L. (24 de Octubre de 2012). *Mi visión de la tecnología.* Obtenido de http://relopezbriega.com.ar/2012/tecnologia/introduccion-a-hadoop-mapreduce/

BlueCoat. (Mayo de 2013). *SlideShare*. Obtenido de Big Data Security Intelligence and Analytics for Advanced Threat Protection: http://es.slideshare.net/BlueCoat/solera-networks-big-data-security-intelligence-and-analytics-for-advanced-threat-protection?next_slideshow=1

Catalyst, I. (s.f.). *Information Catalyst*. Obtenido de http://informationcatalyst.com/

Certicamara. (23 de Agosto de 2013). *Colombia Digital.* Obtenido de http://colombiadigital.net/actualidad/articulos-informativos/item/5543-abc-para-proteger-los-datos-personales-ley-1581-de-2012-decreto-1377-de-2013.html

Colombia, C. d. (17 de 10 de 2012). *Alcadía de Bogotá*. Obtenido de http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=49981

Colombia, G. d. (s.f.). *Catalogo de Datos Abiertos*. Obtenido de http://www.datos.gov.co/frm/Acerca/frmAcercaDe.aspx

Congreso de Colombia. (enero de 2009). *Alcaldia de Bogotá.* Obtenido de http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=34492

Congreso de la República, d. C. (21 de Agosto de 1999). *Alcadia de Bogotá.* Obtenido de http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=4276

Dagos, M. (13 de Noviembre de 2014). *IBM Developers Works*. Obtenido de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

De la Fuente, S. F. (s.f.). *Series Temporales: Modelo Arima.* Obtenido de Estadistica.net: http://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf

Dell_Software. (5 de Mayo de 2015). *Data Mining Techniques*. Obtenido de http://documents.software.dell.com/statistics/textbook/data-mining-techniques

Design, M. A. (s.f.). *Welcome to the age of Big Security Data*. Obtenido de McAfee: http://www.mcafee.com/img/infographics/mcafee-big-security-data.jpg

Drucker, P. F. (1994). *The Age of Social Transformation*. Obtenido de The Atlantic Monthly: https://docs.google.com/file/d/0B-5-JeCa2Z7hd0pSYkV6S3FwOTA/edit

Editor. (9 de Abril de 2015). *Mintic.* Obtenido de http://estrategia.gobiernoenlinea.gov.co/623/articles-7941_manualGEL.pdf

Editor. (Junio de 2015). *Prompt Coud*. Obtenido de https://www.promptcloud.com/blog/big-data-to-fight-cyber-crime/

Editor. (2016). *Google Cloud Platform*. Obtenido de Google Cloud Prediction API Documentation: https://cloud.google.com/prediction/docs/

Editor. (2016). *Hazy.* Obtenido de http://i.stanford.edu/hazy/home

Editor. (2016). *Mahout*. Obtenido de What is Apache Mahout: http://mahout.apache.org/

Editor. (2016). *Zementis*. Obtenido de http://zementis.com/

Evaluando. (Enero15 de 2013). Obtenido de Qué es el análisis predictivo?: http://www.evaluandosoftware.com/que-es-el-analisis-predictivo-2/

Forrester. (10 de Mayo de 2016). *The Forrester Wave: Big Data Text and Analytics Platforms, Q2 2016.* Obtenido de http://reprints.forrester.com/#/assets/2/202/'RES122667'/reports

Gartner. (2015). *Predicitve Analytics*. Obtenido de http://www.gartner.com/it-glossary/predictive-analytics/

Gartner. (04 de 02 de 2016). *Magic Quadrant for Business Intelligence and Analytics Platforms*. Obtenido de https://www.gartner.com/doc/reprints?id=1-2XXET8P&ct=160204

Ghemawat, J. D. (2004). *MapReduce: Simplied Data Processing on Large Clusters.* Obtenido de https://static.googleusercontent.com/media/research.google.com/es//archive/mapreduce-osdi04.pdf

Gobierno en línea. (9 de Abril de 2015). Obtenido de http://estrategia.gobiernoenlinea.gov.co/623/articles-7941_manualGEL.pdf

Guazzelli, A. S. (2009). *Efficient deployment of predictive analytics through open standards and cloud computing.* Obtenido de ACM SIGKDD Explorations Newsletter, 11(1), 32-38.: http://www.kdd.org/exploration_files/p5V11n1.pdf

Hadoop. (2016). *Hadoop*. Obtenido de What Is Apache Hadoop?: http://hadoop.apache.org/

He, E. S. (2010). Narrative Visualization: Telling Stories with Data. En T. o. Graphics.

HP, Nur, & Manzano, A. M. (2015). *HPE Big Data Descripcioón.*

I.H.Writen. (2011). *DAta Mining: Practical Machine Learning tool a Techniques.* Morgan Kaufmann.

Icontec, I. (20 de 12 de 2013). *Icontec Internacional.* Obtenido de http://tienda.icontec.org/brief/NTC-ISO-IEC27001.pdf

IntelSecurity. (2014). *McAfee.* Obtenido de Needle in a Datastack Report: http://www.mcafee.com/us/resources/reports/rp-needle-in-a-datastack.pdf

Josh Parenteau, R. L. (04 de February de 2016). *Magic Quadrant for Business Intelligence and Analytics Platforms.* Obtenido de Gartner: https://www.gartner.com/doc/reprints?id=1-2XXET8P&ct=160204

Leavitt, N. (2010). *Technology News.* Obtenido de Will NoSQL Databases Live Up to Their Promise?: http://leavcom.com/pdf/NoSQL.pdf

Mangelsdorf, J. (Noviembre de 2013). *CSS World*. Obtenido de Using Big Data to Defend Against Cyber Security Threats: http://www.csc.com/cybersecurity/publications/93325/104033-using_big_data_to_defend_against_cyber_security_threats

Marcos D. Assunção a, ∗. R. (2013). *Data computing and clouds: Trends and future directions.* Brazil: Elsevier.

Marqués, M. P. (2015). *Business Intelligence. Técnicas, herrameintas y aplicaciones.* Mexico D.F: Alfaomega.

Matt Turck, J. H. (Febrero de 2016). *Big Data Landscape 2016.* Obtenido de http://mattturck.com/wp-content/uploads/2016/01/matt_turck_big_data_landscape_full.png

McAfee. (2014). *Needle in a Datastack Report*. Obtenido de http://www.mcafee.com/us/resources/reports/rp-needle-in-a-datastack.pdf

Meulen, R. v. (16 de Septiembre de 2015). *Gartner*. Obtenido de http://www.gartner.com/newsroom/id/3130817

Mircea Răducu TRIFU, M. L. (2014). Big Data: present and future. *Database Systems Journal*, 32. Obtenido de http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=d647a8f3-854a-470b-bfca-8be3007465dc%40sessionmgr4001&vid=1&hid=4101

OBS Business School. (03 de 02 de 2015). Obtenido de http://www.obs-edu.com/noticias/estudio-obs/en-2020-mas-de-30-mil-millones-de-dispositivos-estaran-conectados-internet/

Open Data, M. (30 de Junio de 2015). *Open Data*. Obtenido de http://www.mintic.gov.co/portal/604/w3-article-5664.html

Pagliery, J. (03 de Diciembre de 2016). *CNN Tech*. Obtenido de http://money.cnn.com/2016/12/02/technology/russia-central-bank-hack/

Revista HelpDesk, T. (6 de Enero de 2015). *Revista Help Desk Tic.* Obtenido de http://revista.helpdesktic.com/cuadrante-magico-de-gartner/

Robert Kosara, J. M. (2013). *Kosara.* Obtenido de Storytelling: The Next Step for Visualization: http://kosara.net/papers/2013/Kosara_Computer_2013.pdf

S. Bateman, R. M. (2010). “Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Chart.

Taringa. (s.f.). *Unidades de almacenamiento informático*. Obtenido de http://www.taringa.net/post/info/10843787/Unidades-de-almacenamiento-informatico-o-electronico.html

Technet, M. (02 de Abril de 2011). *Kevin Remde's IT Pro WebLog*. Obtenido de https://blogs.technet.microsoft.com/kevinremde/2011/04/03/saas-paas-and-iaas-oh-my-cloudy-april-part-3/

Tecnosfera. (3 de noviembre de 2015). El 'Big Data' es clave para el futuro. *El Tiempo*.

Tecnósfera, R. (28 de enero de 2016). *El Tiempo*. Obtenido de http://www.eltiempo.com/tecnosfera/tutoriales-tecnologia/cuantos-delitos-informaticos-se-denuncian-en-colombia/16493604

Usama Fayyad, G. P.-S. (1996). he KDD process for extracting useful knowledge from volumes of data. *Magazine Communications of the ACM*, 27-34.

Varonis. (21 de julio de 2015). *What is User Behavior Analytics?*

Verizon. (2012). *Data Breach Investigations Report.* Obtenido de https://www.wired.com/images_blogs/threatlevel/2012/03/Verizon-Data-Breach-Report-2012.pdf

Verizon. (2016). *Regmedia.* Obtenido de Data Breach Investigations Report: https://regmedia.co.uk/2016/05/12/dbir_2016.pdf

wikiversity. (s.f.). Obtenido de https://es.wikiversity.org/wiki/Modelo_autorregresivo_integrado_de_media_m%C3%B3vil

Williams, R. (Junio de 2014). *The Telegraph*. Obtenido de Cyber crime costs global economy $445 bn annually: http://www.telegraph.co.uk/technology/internet-security/10886640/Cyber-crime-costs-global-economy-445-bn-annually.html