UNIVERSIDAD DE VALLADOLID

E.T.S.I. TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

GRADO EN INGENIERÍA DE TECNOLOGÍAS ESPECÍFICAS DE TELECOMUNICACIÓN. MENCIÓN EN TELEMÁTICA

SOLUCIONES DE PRIVACIDAD Y SEGURIDAD PARA DIFERENTES ESCENARIOS DE BIG DATA EN MEDICINA

Autor:

D. Héctor Merino Cosgaya

Tutor:

Dña. Isabel de la Torre Díez

Valladolid, 22 de Julio de 2016

Agradecimientos

Quisiera agradecer a mi familia todo el apoyo que me han dado durante estos años de carrera, que cada vez que lo veía mal me animaban. Agradecer a mis compañeros y amigos los trabajos compartidos juntos durante todos los laboratorios y prácticas, y a mi tutora, Isabel de la Torre, por su ayuda para realizar este trabajo.

Universidad de Valladolid

TÍTULO:	**SOLUCIONES DE PRIVACIDAD Y SEGURIDAD PARA DIFERENTES ESCENARIOS DE BIG DATA EN MEDICINA**
AUTOR:	**D. Héctor Merino Cosgaya**
TUTOR:	**Dña. Isabel de la Torre Díez**
DEPARTAMENTO:	**Departamento de Teoría de la Señal y Comunicaciones e Ingeniería Telemática**

TRIBUNAL

PRESIDENTE:	**D. Miguel López-Coronado Sánchez-Fortún**
VOCAL:	**Dña. Isabel de la Torre Díez**
SECRETARIO	**Dña. Beatriz Sainz de Abajo**
SUPLENTE	**D. Carlos Gómez Peña**
SUPLENTE	**D. Salvador Dueñas Carazo**

FECHA:	**22 de Julio de 2016**
CALIFICACIÓN:	

Resumen de TFG

Con Big Data se procesan grandes volúmenes de data con el fin de obtener información y poder generar conocimiento de ellos. En el campo de la sanidad, la principal fuente de información es la Historia Clínica Electrónica (HCE). Otras fuentes son las redes sociales y el Internet de las cosas. Los datos de salud son almacenados en grandes bases de datos en la actualidad y compartidos en múltiples medios electrónicos. Pero, ¿por qué dichos datos pueden despertar el interés de mafias organizadas y sumamente peligrosas? Los usos que estas mafias pueden darle a los datos son entre otros: chantajear a personas a partir de la información sobre sus enfermedades, vender información sanitaria a empresas de marketing, etc. En este artículo se analiza el problema de la seguridad de big data en el contexto de la sanidad y diferentes soluciones son propuestas. Hay muchas técnicas diferentes para preservar la seguridad, como pueden ser: técnicas de modificación de datos, métodos de cifrado y protocolos para el compartimiento de datos, y otros. Estos son analizados en el trabajo de investigación. Aún queda mucho por hacer en el campo de la seguridad en big data pero poco a poco se va avanzando en un campo de gran interés comercial y científico.

Palabras clave:

Big data

eSalud

HCE

Privacidad

Sanidad

Seguridad

Universidad de Valladolid

Tabla de contenidos

Capítulo 1. INTRODUCCIÓN

Sección 1.1. Introducción a los objetivos del proyecto

Sección 1.2. Qué es big data

Sección 1.3. Crecimiento

Sección 1.4. Almacenamiento y bases de datos

Capítulo 1. INTRODUCCION

Sección 1.1.- Introducción a los objetivos del proyecto

Con la evolución de la tecnología, cada vez existen nuevos medios y más eficientes sistemas para la obtención y almacenamiento de información. Con este trabajo, lo que he tratado de recoger es cómo está la sanidad, en la actualidad, recogiendo estos datos, que los llamaremos a partir de ahora "big data" y cómo los está almacenando para su posterior análisis. Además, se pretende explicar diferentes formas para proteger estos datos de aquellas personas que deseen dar un uso ilícito de los mismos.

Sección 1.2.- Qué es big data

Big data, un término de nuevo surgimiento que nació por el incremento masivo de volumen de información generado. Este nuevo concepto, hace referencia a información estructurada y no estructurada, supone un volumen de información tan grande que la solución a su almacenamiento es la Nube. Por tanto, una definición que dan a big data es "Big data is a massive collection of shareable data originating from any kind of private or public digital sources, which represents on its own a source for ongoing discovery, analysis, and Business Intelligence and Forecasting" por Banica et al. (2015) [1].

Una parte de los expertos en big data mencionan que tiene 3 características, mientras que otros las aumentan hasta las 5. Estas características también

las llaman las “5 V´s”. Estas son: Volumen, Variedad, Velocidad, Veracidad y Valor. La definición de estas características se puede ver resumida en la Tabla 1 [2-4].

Característica	Definición
Volumen	Enorme cantidad de datos creados a partir de multitud de fuentes.
Variedad	Los datos generados vienen en multitud de formatos diferentes, ya sea un formato estructurado o formato no estructurado.
Velocidad	Rapidez con la que se crea el volumen de información además de que dicha información sea sensible al tiempo, de tal forma que debe ser almacenada, procesada y analizada rápidamente.
Veracidad	Hace referencia a la precisión de los resultados y el análisis de los datos.
Valor	Es el valor añadido una vez analizados los datos.

Tabla 1: *Características de big data. Fuente: Propia*

A pesar de caracterizar el termino big data, no se ha encontrado información que establezca un umbral mediante el cual se pueda diferenciar si la información almacenada alcanza las cualidades necesarias para llamarlo big data o no.

Sección 1.3.- Crecimiento

El termino big data nació en 2009, pero su crecimiento a lo largo de estos años ha sido imparable. El aumento de artículos referidos a big data y en concreto en el ámbito sanitario ha crecido exponencialmente. Esto se puede

apreciar mejor si vemos la gráfica que nos muestran Andreu-Perez et al. (2015) en su artículo [4].

Figura 1: *Crecimiento de big data. Fuente:* Andreu-Perez [4]

Esto nos hace ver que se trata de un tema muy importante y que va a crecer aún más en los próximos años. Esta información solo está relacionada con el big data en sanidad, por lo que se puede suponer que el crecimiento de otros ámbitos, como son economía, tecnología o en empresas está creciendo o va a crecer de igual forma.

Sección 1.4.- Almacenamiento y bases de datos

Hablamos de cifras muy altas de volúmenes de información, ahora vamos a darlas un valor. En 2012, el volumen de datos registrado para la sanidad mundial era de 500 petabytes. Para 2020, hay una previsión que este volumen de datos pueda aumentar hasta los 25000 petabytes, lo que supone

un aumento de 50 veces en apenas 8 años. Y un problema a mayores es que no es información simple, sino que existen multitud de formatos de información entre todos estos datos sanitarios [5]. Habitualmente, toda esta información estaría recogida en bases de datos tradicionales, es decir, aquellas que utilizan tablas relacionales. Con el big data, esto ya no es eficiente de usar. La razón de ellos es que hay que pasar toda esta información a un formato de tablas ya predefinido y después crear una relación entre todo bajo unos criterios. Por tanto, la estructura que se crea es muy compleja, nada eficiente en búsqueda, lenta, poco escalable y costosa [6]. Por ello, la forma de mejorarlo es usando bases de datos NoSQL, que de forma popular se traduce como "Not only SQL". Estas nuevas bases de datos son no relacionales, por tanto resulta simple almacenar todos los datos en ellas. Además, son muy escalables, por tanto no importa la cantidad de información que introduzcamos en la misma [7]. Estas nuevas bases de datos nos traen más ventajas además de su escalabilidad y facilidad de almacenaje. Nos permiten procesar datos mucho más rápido. A diferencia de las relacionales, las NoSQL no están sujetas al denominado ACID. ACID viene de las siglas en ingles de "atomicity, consistency, isolation, durability". Atomicity, significa que las operaciones deben hacerse en su totalidad o no hacerse, "cosistency" significa que ninguna operación realizada en la base de datos tiene los permisos para romper las reglas de la misma, "isolation" significa que las

operaciones pueden realizarse independientemente de otras aplicaciones u operaciones y “durability” significa que las transacciones completadas deben persistir. Que no estén sujetas a estas reglas, hace de las NoSQL que sean más rápidas además, que en general los modelos de datos son más simples [6]. Algunas de las principales bases de datos NoSQL son MongoDB, Cassandra, BigTable o HBase [8].

Ahora que ya sabemos cómo almacenar toda esta información, llega la pregunta fundamental, ¿dónde? Como he explicado anteriormente, el crecimiento tan grande de la información generada será un problema. Hasta ahora, las empresas y centros han optado por almacenar su información en centros especializados, en los cuales la seguridad tanto de infraestructura como la informática eran elevadas. Pero esto deja de ser sostenible, el aumento de información hace que el coste de mantener y de infraestructura de este sistema sea cada año cada vez mayor. Esta es la razón principal por la cual han empezado a aparecer cada vez de forma más mayoritaria los servicios de almacenamiento en la nube. La computación en la nube ha ido creciendo de forma importante en los últimos años. Una razón de ello es porque big data está directamente relacionado con la nube. La importancia que está teniendo hace que la nube se esté convirtiendo en el nuevo paradigma de arquitectura de la tecnología de la información. El motivo de esta importancia está basado en tres principales características. La primera

de ellas es su arquitectura de infraestructura de hardware. Esta arquitectura está basada en los clusters, que son muy escalables y de bajo coste debido a que se utilizan un gran número de servidores y almacenamiento de bajo coste. La segunda de ella es que para conseguir el máximo uso de recursos, es necesario el desarrollo de servicios fundamentales y aplicaciones colaborativas para mantener en funcionamiento la plataforma. Y la tercera es que la redundancia de uso de múltiples servidores de bajo coste es fácilmente solucionable mediante software [9].

Al igual que en redes cableadas, en la nube también existen diferentes modelos, los cuales están diferenciados dependiendo de cuantas capas de la arquitectura estemos usando. Estas capas de la arquitectura son: Infraestructura física, Infraestructura virtual, Plataforma, Aplicación y Red. Estas capas se dividen según la tabla siguiente y están relacionadas con los modelos como se puede apreciar.

<table>
<tr><td>Usuario</td><td>Red</td><td></td></tr>
<tr><td rowspan="4">Proveedor de servicios</td><td>Aplicación</td><td>SaaS</td></tr>
<tr><td>Plataforma</td><td>PaaS</td></tr>
<tr><td>Infraestructura virtual</td><td rowspan="2">IaaS</td></tr>
<tr><td>Infraestructura física</td></tr>
</table>

***Tabla 2**: Arquitectura de la nube. Fuente: Fatemi Moghaddam [10]*

Universidad de Valladolid

Los modelos de nube mencionados son los que aparecen en la tercera columna de la tabla 2 y son: Infraestructure-as-a-Service (IaaS), Platform-as-a-Service (SaaS) y Software-as-a-Service (SaaS). Estos diferentes modelos los explico a continuación:

Infraestructure-as-a-Service: un proveedor de servicios IaaS invierte en infraestructura, desplegando y manteniéndolo para ofrecer infraestructura física y virtual. Es accedido remotamente para la configuración y monitorización de los recursos y proporciona la oportunidad de instalar al cliente sus propias aplicaciones como si de su infraestructura se tratase.

Platform-as-a-Service: este modelo proporciona un amplio espectro de servicios de aplicaciones detallados para ofrecer un entorno de desarrollo y ejecución. Además, este modelo proporciona aislamiento entre los consumidores y los recursos que se comparten.

Software-as-a-Service: en este modelo, lo que contiene son aplicaciones basados en la nube. Los proveedores eliminan la instalación de hardware, pago de licencias y configuración de middleware mientras que mejora la instalación de software, la configuración y la personalización de los mismos [10-11].

Una vez explicado cómo almacenar los datos y dónde guardarlos, queda la tarea más costosa de todas, analizarlos. Big data se trata de información principalmente sin formato, sin relación directa entre los datos como

podemos ver en bases de datos tradicionales. Con las bases de datos NoSQL tenemos el problema de que son relativamente recientes y no existe un estándar que las unifique, por tanto no son compatibles entre ellas. Esto nos lleva a la necesidad de desarrollar algoritmos y mecanismos que nos permitan un análisis de los datos, compatible con cualquier base de datos y que sea lo más eficiente y rápido posible. De esta necesidad surgen MapReduce y Hadoop.

Hadoop, una plataforma desarrollada por Apache, [12], diseñado para hacer frente a gran cantidad de datos. Para ello usa la metodología de dividir-y-conquistar para procesamiento, que le permite lidiar con grandes datos complejos y desestructurados que normalmente no encajan dentro de una base de datos relacional [13].

Figura 2: *Esquema de Hadoop. Fuente: propia*

Los datos son almacenados en una arquitectura de sistema de ficheros distribuida (HDFS), en la cual cada fichero de datos esta difundido por multitud de nodos, conectados a través de una red de alta velocidad. El modelo tiene un servidor maestro, "Nodo Maestro" el cual se encarga de regular la distribución de la información de los nodos de datos, "Nodos Esclavos". Los nodos esclavos son los responsables de las operaciones de lectura y escritura en el sistema de ficheros. Cada nodo esclavo tiene asociado un seguidor de tareas, proporcionando ayuda en el progreso de tareas del nodo maestro, haciendo que el sistema sea tolerante a fallos y reduce los datos perdidos en las operaciones [13].

En la Figura 2 se puede ver un resumen de los pasos que se siguen para la obtención de resultados usando Hadoop. Consta de cuatro principales pasos y estos son los siguientes: 1) en este paso el cliente solicita información al Nodo Maestro, el cual analiza la petición. 2) en este segundo paso, tras haber analizado la petición del cliente, el Nodo Maestro reparte el trabajo que deben realizar los Nodos Esclavos y los proporciona los datos que necesitan. 3) en el tercer paso, los Nodos Esclavos ya han terminado de procesar todos los datos y devuelven el resultado al Nodo Maestro. 4) en el último paso, el Nodo Maestro junta todos los resultados obtenidos y proporciona el resultado final al cliente.

MapReduce es una solución de Google para procesar el big data. MapReduce es un desarrollo software, escrito en Java, diseñado para ser ejecutado en un cluster distribuido de máquinas. La forma que tiene de funcionar MapReduce es la siguiente. Primero de todo, los datos se dividen en pequeños trozos y son distribuidos sobre miles de computadoras, es a lo que llaman Google File System (GFS). Una vez dividido en pequeños trozos, se procede a aplicar sus dos funciones principales, Map y Reduce. La función Map distribuye los cálculos hacia donde están situados los pequeños trozos de datos que hemos dividido antes. La función Reduce hace una recopilación de todos los resultados obtenidos al final [14,11].

Capítulo 2. BIG DATA EN SANIDAD

Capítulo 2. BIG DATA EN SANIDAD

Uno de los grandes propósitos del big data se centra en la sanidad. Con el big data y los cada vez más potentes computadores, se quiere conseguir que la sanidad pase de ser genérica a personalizada. Para ello se pretende decodificar las secuencias genéticas de cada paciente para de esta forma personalizar y saber con certeza el tratamiento que resultaría más eficaz en su caso. Uno de los avances conseguidos gracias a la evolución de la tecnología y el big data es la inversión necesaria para secuenciar el genoma, que ha decrecido de millones inicialmente a unos miles de dólares por genoma [15].

Otro de los propósitos de big data es el del control de epidemias. A las bases de datos tradicionales se les añade un factor de georreferencia y gracias a la información obtenida tanto de medios oficiales como informales, blogs, búsquedas, redes sociales, etc. se puede hacer una estimación de la evolución de la enfermedad [16]. Así por ejemplo Google desarrolló la herramienta, Google Flu Trends [17], la cual tenía en consideración el volumen de búsquedas de palabras clave y comentarios en redes sociales como Twitter y Facebook para ver la evolución, inicialmente de la gripe H1N1 [18].

Uno de los principales campos en el cual big data está ganando gran peso es la genómica. Se trata de un campo que está en constante evolución y

descubrimiento. En este campo es donde se trata de descubrir y analizar el genoma individual de cada persona y, como mencionamos al principio, conseguir una sanidad personalizada. En este campo también se incluye el estudio del ADN y por qué las bacterias resisten los antibióticos [19]. Otro campo son los hospitales, en los cuales analizar por medio de big data los análisis realizados a los pacientes puede dar como resultado obtener un diagnóstico más temprano. De esta manera se podrían reducir la dosis de medicamento que el paciente debe recibir. Relacionado con los hospitales está la detección de focos infecciosos o la evolución de posibles virus que pueden llegar a convertirse en epidemia, como paso con el virus H1N1. Cuando surgió esta epidemia, algunas empresas ya estaban usando el big data para analizar cómo iba evolucionando. Una de estas empresas era Google, que utilizaba su buscador como fuente de datos. Mediante la búsqueda de los usuarios de palabras clave, Google analizaba por medio de geolocalización y cantidad de búsqueda hacia donde se extendía el virus y así encontrar un posible patrón que predijera su evolución y advertir a los Centros de Control de Enfermedades (*Centers for Disease Control and Prevention* (CDC)).

Por último, aunque se use big data y cada vez más en todos los campos sanitarios, es importante mencionar el campo de la diabetes [20-22]. Se trata de una enfermedad que cada vez está afectando a una cantidad mayor

de personas en todo el mundo. Es causa de una considerable cantidad de muertes que se producen por enfermedades no contagiosas. La diabetes es una enfermedad que se puede desarrollar en cualquier momento, incluso de nacimiento, y conlleva un tratamiento de por vida [23]. Por ello se está usando el big data para tratar de encontrar una cura o remedio a esta enfermedad ya que la información y datos obtenidos son de un gran volumen [24-27].

Capítulo 3. Métodos

Sección 3.1.- Metodología usada para la información de big data en sanidad.

Sección 3.2.- Metodología usada para la información sobre seguridad.

Capítulo 3. Métodos

La metodología usada para la obtención de la información será explicada a continuación, haciendo una diferencia entre la información obtenida relacionada con big data en sanidad y la información obtenida para poder realizar un análisis teórico de la seguridad de big data.

Sección 3.1.- Metodología usada para la información de big data en sanidad.

Para llevar a cabo la investigación se ha realizado una búsqueda exhaustiva de artículos en las diferentes bases de datos científicas, las cuales son: Scopus [28], PubMed [29], Science Direct [30] y Web of Science [31]. En cada una de ellas realizando una búsqueda especifica cómo se puede apreciar en la Tabla 3. En todos los casos el intervalo temporal fue desde 2005 hasta enero de 2016.

Base de datos	Criterio de Búsqueda	Campos de búsqueda
Scopus	"big data" AND "health"	"abstract, title, keywords" "all fields"
Pubmed	"big data" AND "health"	"title/abstract" "all fields"
Science Direct	"big data" AND "health"	"abstract, title, keywords" "all fields"
Web of Science	"big data" AND "health"	"title" AND "topic" " topic"

Tabla 3. Búsquedas realizadas en las diferentes bases de datos. Fuente: Propia.

La Figura 3 muestra los 9724 resultados obtenidos con las búsquedas realizadas. También indicamos cuantos artículos descartamos por estar

duplicados o tienen un título no relacionado con nuestro interés. Por último nos quedamos con 46 artículos tras leer los 209 y ver cuáles de ellos por su "abstract" no nos resultan beneficiosos. Al final nos quedamos con estos artículos porque los demás, aunque inicialmente podían tratar de big data en sanidad, tras leerlos, únicamente mencionan el big data sin entrar en profundidad y explican aspectos sanitarios únicamente. Por tanto todos estos los descartamos al no poder aportarnos información relevante.

Para realizar esta selección, hemos considerado artículos en inglés o español. Tras leer los títulos y los abstract y quedarnos con los 209 mencionados, hemos procedido a leer su contenido y de esta forma determinar cuáles de ellos nos aportaban información relacionada con big data en sanidad y cuales, como sucede en los descartados, únicamente ofrecen información sanitaria en la cual mencionan sin detalles el big data.

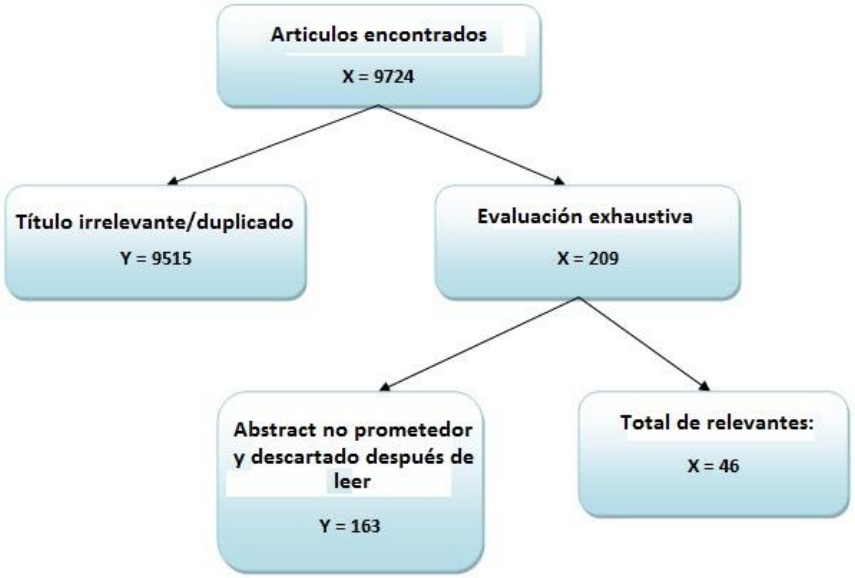

Figura 3. Diagrama de flujo de los pasos dados para la búsqueda de información. Fuente: propia

Sección 3.2.- Metodología usada para obtener información sobre seguridad de big data.

Para la llevar a cabo la investigación se ha realizado una búsqueda exhaustiva de artículos en las diferentes bases de datos científicas, las cuales son: IEEE Xplore [32], Scopus [28], Science Direct [30] y Web of Science [31]. En cada una de ellas realizando una búsqueda especifica cómo se puede apreciar en la Figura 4. En todos los casos el intervalo temporal fue desde 2008 hasta marzo de 2016.

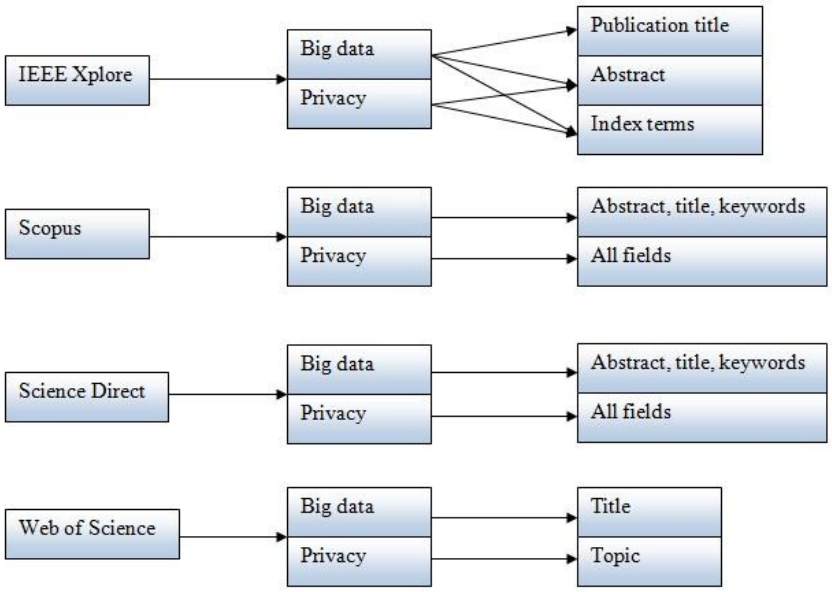

Figura 4. Relación de búsquedas en las bases de datos Fuente: propia

La Figura 5 muestra los 3168 resultados obtenidos con las búsquedas realizadas. También indicamos cuantos artículos descartamos por estar duplicados o tienen un título no relacionado con nuestro interés. Por último

nos quedamos con 21 artículos tras leer los 133 y ver cuáles de ellos por su "abstract" no nos resultan beneficiosos.

Al final nos quedamos con estos artículos porque los demás, aunque inicialmente podían tratar de big data y su seguridad, tras leerlos, únicamente mencionan situación de big data o cómo está la legislación en Europa o EEUU respecto a datos en la nube. Por tanto todos estos los descartamos al no poder aportarnos información relevante.

Para realizar esta selección, hemos considerado artículos en inglés. Tras leer los títulos y los abstract y quedarnos con los 133 mencionados, hemos procedido a leer su contenido y de esta forma determinar cuáles de ellos nos aportaban información relacionada con big data y su seguridad en la nube o en las bases de datos que utiliza para almacenar los datos.

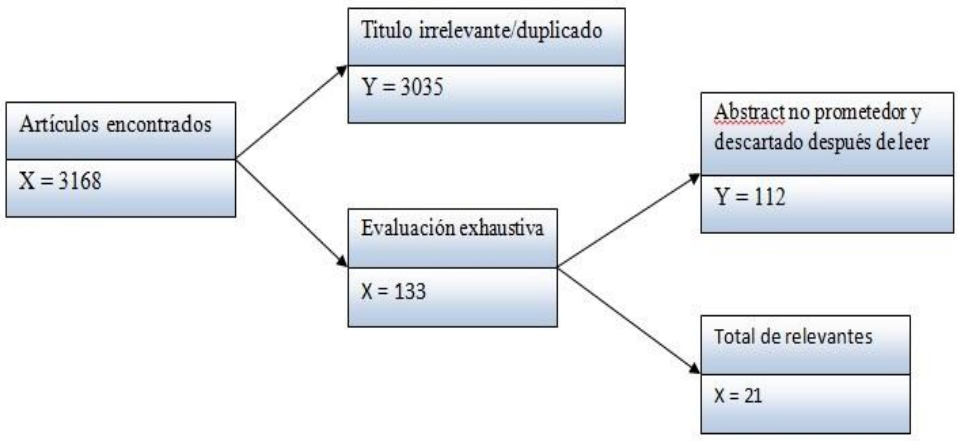

Figura 5. Diagrama de flujo de los pasos dados para la búsqueda de información Fuente: propia

Capítulo 4. Resultados

Sección 4.1. Sanidad

Sección 4.2. Seguridad de Big data en sanidad

Capítulo 4. Resultados

Sección 4.1. Sanidad

Tras el análisis de los artículos recogidos según la metodología del capítulo anterior, los más importantes son los que aparecen a continuación, recopilando las conclusiones más relevantes acerca del big data en sanidad.

Awais Ahmad et al. (2016) proponen un sistema de análisis de datos de big data basado en Machine-to-Machine (M2M) y que usa el mecanismo de dividir-y-conquistar para aumentar la eficiencia del sistema además de la velocidad para procesar los datos [19]. Victor Chang et al. (2016) realizan una comparación entre sí los datos generados se almacenan más eficientemente en la nube o en la no-nube. Tras realizar un experimento en el cual transferían 10000 ficheros de 1 GB cada uno y otro estudio en el cual transferían 1000 ficheros de 10 GB cada uno, determinaron que el tiempo era menor para la transferencia hacia la nube. Además, la mejora de rendimiento era superior para la nube que para la no-nube mientras que el control de riesgo era el mismo para ambos casos [2]. Roger Clarke (2016) comenta el gran énfasis en las oportunidades que ofrece big data y en la poca atención que se ha prestado a las amenazas que están surgiendo. Una de estas oportunidades es la de utilizar, como el SETI@, los ordenadores inactivos de la población para analizar pequeños trozos de datos que están almacenados [29].

Parra Calderón (2015) explica que la implantación de big data en sanidad es imparable y que ya se pueden ver los posibles riesgos y limitaciones que pueden aparecer [15]. Javier Andreu-Pérez et al (2015) explica que big data puede servir para potenciar la investigación clínica para casos reales. Además puede proporcionar una oportunidad para habilitar una medicina más precisa y efectiva [4]. Daniel M. Trifiletti et al. (2015) creen que la inversión en big data para detección de tumores y tratamiento de cáncer sufrirá un crecimiento significativo [30].

Rosemary Wyber et al. (2015) establecen que big data en la prestación de ayuda sanitaria podrá convertirse en un hito debido a los potenciales beneficios que va a representar [31]. Hood et al. (2015) opinan que el lanzamiento *Hundred Person Wellness Project* (HPWP) va a suponer un gran avance en la medicina preventiva y mejorar el bienestar de los pacientes [32]. Wong and Lai (2015) en sus artículos tratan de usar el big data para analizar el uso de las ambulancias en emergencias y zonas a donde se desplaza. De esta manera se podría realizar predicciones de su uso en futuros días llegando incluso a los años posteriores [16]. Effy Vayena et al. (2015) explican qué es el digital desease detection (DDD) concluyen que supone también desafíos éticos que deben tenerse en consideración [33]. Gloria Pérez (2015) explica que con la llegada de big data no es necesario realizar una separación entre la investigación con big data y la

investigación tradicional [34]. Bradford W. Hesse et al. (2015) examinan las implicaciones de un rápido desarrollo en entorno de trabajo distribuido que colaboran científicos de biomedicina y sociales y ciencias del comportamiento [35]. José Manuel Martínez Sesmero (2015) hace mención a las áreas más significativas donde big data está resultando un claro potencial de desarrollo [36]. Ari Moskowitz et al. (2015) explican que con los avances tecnológicos y con la llegada de la era de big data, es necesario formar a investigadores para que sean capaces de utilizar las nuevas herramientas que se están desarrollando [37]. Tao Huang et al. (2015) cree que el big data en sanidad va a tener mayor importancia de lo que la gente cree. Llevar los nuevos dispositivos inteligentes hará que puedas almacenar las constantes vitales y almacenarlo en la nube para poder ser analizado después [8].

João Cunha et al. (2015) proponen una arquitectura basada en Hadoop para extraer información sanitaria a través de Twitter [13]. Guillermo Lafuente (2015) se da cuenta que el big data no debe verse como una pesadilla de seguridad porque la seguridad es un proceso y no un producto [38]. Ho Ting Wong et al (2015) estudiaron tres situaciones en urgencias sanitarias en las cuales demostraron que big data tiene un gran potencial en este sector [16]. Emad Elsebakhi et al. (2015) realizan un estudio experimental para desarrollar un algoritmo que permita el análisis de datos de una forma

rápida y eficiente [39]. J.Archenaa et al. (2015) concluyen que si los gobiernos se aprovechan del big data, pueden obtener grandes beneficios para mejorar la calidad de los ciudadanos [21]. Sean D. Young (2015) explica que gracias al big data y las nuevas herramientas que grandes empresas como Google van desarrollando, resulta más sencillo estudiar la evolución de algunos virus o epidemias que surjan [22]. Dr. Saravana kumar et al. (2015) explica que analizar los datos de los pacientes y la población, en este caso, para estudiar la diabetes, puede ayudar a estimar y predecir qué personas tienen mayor probabilidad de padecer la enfermedad [23]. Greg Satell (2015) explica en este capítulo situaciones en las cuales la seguridad puede verse comprometida legalmente [40].

Ashwin Belle et al. (2015) nos muestra las diferentes áreas en medicina donde se puede aplicar el big data y qué es lo que se pretende obtener en cada una de ellas [41]. Karim S. Ladha et al. (2015) concluye que la recolección de grandes cantidades de datos electrónicos va a permitir crear una nueva era de investigación [42]. Akram Alyass et al. (2015) explican que es necesario estimular las iniciativas de desarrollo en el campo del análisis e integración de big data. Además, son necesarias mayores inversiones en bioinformática, biomatemáticas y bioestadística [43]. Xiaolong Jin et al. (2015) realizan un resumen de las principales oportunidad que nos trae big data a las empresas, sanidad y demás sectores

debido al gran impacto que está siendo en la actualidad [44]. Karim S. Ladha et al. (2015) nos hacen ver que con el avance de los expedientes clínicos digitales, analizar los datos con big data y sus herramientas para investigar enfermedades raras, entre otras cosas, ha supuesto una revolución en la investigación [42].

Ivan Merelli et al. (2014) opina que los datos ya son el cuarto paradigma de la ciencia detrás de la experimentación, la teórica y las ciencias computacionales [18]. Min Chen et al. (2014) fundamenta que big data nos traerá grandes oportunidades. Hasta ahora la tecnología era desarrollo pero en un futuro los datos serán el progreso tecnológico [20]. Nir Kshetri et al. (2014) establece que big data está estrechamente conectado con las preocupaciones de privacidad, bienestar y seguridad. Estas preocupaciones llevan a la necesidad de una política de seguridad fuertemente estricta [3]. Ronald Margolis et al. (2014) advierten que para los desafíos que supone big data en biomedicina es necesario encajar todas las piezas que supone el big data. Lo cual, aunque complejo, es posible su realización [45]. Xuyun Zhang et al. (2014) formulan una serie de requerimientos básicos para preservar la privacidad de los datos en la nube y de big data [46]. P. Otero et al. (2014) creen que los programas de formación informática en biomedicina y salud deben introducir los nuevos conceptos de análisis y desarrollar las habilidades para que sean capaces de usarlos [47]. Richard

Kemp (2014) observa que tener una visión competitiva de cómo gestionar grandes volúmenes de datos será un objetivo clave en la estrategia de las grandes empresas [48]. Fabricio F. Costa (2014) explica que la creciente cantidad de "omics data" generados dependerá de la capacidad para poder interpretarlo [17]. Yulong Shen et al. (2014) explican diversas situaciones en las cuales la integridad de big data está comprometida en redes inalámbricas. Para ello explican el protocolo 2HR y simulan resultados para corroborarlo [49].

C.L. Philip Chen et al. (2014) hace un resumen de lo que es big data, cuáles son sus oportunidades y desafíos así como las actuales técnicas y tecnologías [50].

David Shin et al. (2013) concluye que, en el ámbito de la seguridad, es necesario definir con claridad qué es lo importante y como está diseñado porque unos algoritmos pueden resultar más eficientes que otros, por ejemplo el AES en DaaS [51]. Aisling O'Driscoll et al. (2013) comentan que la cantidad de datos generados en genómica ha crecido enormemente y además dan a conocer diferentes técnicas y plataformas para poder analizar todos estos datos [11]. Richard Cumbley et al. (2013) dan a conocer diversas situaciones en las cuales el big data puede verse comprometido y estar en peligro. Además comentan cómo se encuentra la regulación europea para el big data [52]. Raghunath Nambiar et al. (2013) dan a

conocer que en sanidad se está empezando a entender todas las cosas innovadoras que se pueden realizar con big data. Gracias a las herramientas y tecnología permitirá generar nuevas soluciones [53].

Benjamin H. Brinkmann et al. (2009) desarrollaron un nuevo formato de fichero el cual les permitía almacenar grandes volúmenes de datos, de sistemas neurofisiológicos, procedentes de multitud de fuentes [54].

Sección 4.2. Seguridad de big data en sanidad

En esta sección tenemos un resumen de los principales avances conseguidos por la comunidad científica que nos ayudarán a mantener nuestros datos más seguros.

Liu et al. (2015) definen una serie de pasos en los cuales se verifica de forma externa la validez de los datos almacenados. Una verificación externa es tan importante como la privacidad y seguridad que proporciona el lado del servidor. Al tratarse de un agente externo, es necesario establecer unos pasos para mantener tanto la privacidad de los datos como su integridad [55].

Fabiano et al. (2015) pertenecientes a la Universidad de Wyoming, han ido desarrollando una variante del paradigma MapReduce para aplicarlo en seguridad y que cumple con HIPAA (the U.S. standard for medical record privacy) además de usar el paquete de cifrado OpenSSL. Para conseguir la máxima escalabilidad, han implementado un hibrido de paradigma de programación OpenMP-MPI. Mediante este paradigma lo que consiguen es que a cada núcleo del procesador se asignan un número de ficheros y luego cada núcleo subdivide estos ficheros dependiendo del número de hilos que esté usando [56].

Yan et al. (2015) proponen dos esquemas de seguridad en los cuales se quiere proteger la privacidad de los proveedores de confianza. El primer

esquema está enfocado en la eficiencia computacional mientras que el segundo proporciona una mayor protección a costa de coste computacional. Utilizan homomorfismo aditivo con recifrado basado en proxy para el diseño de estos esquemas para Privacy-Preserving Trust Evaluation (PPTE) [57].

Hsu et al. (2014) nos enseñan cómo desarrollar un protocolo para la transferencia segura de datos. Además, nos proponen un protocolo para la transferencia de una clave de grupo. Para ello, realiza una variante del algoritmo de Diffie Hellman el cual está diseñado para una comunicación uno a uno y no una comunicación varios a varios. La motivación del diseño de este protocolo es para preservar el refresco de la clave, la confidencialidad de la clave y la autenticación de la clave [58].

Zhou et al. (2015) proponen un algoritmo de cifrado centrado en la seguridad de imágenes. Este algoritmo está basado en el proceso estocástico de Caos y en el principio de Line Map. Este algoritmo está diseñado para que si una imagen se cifra, si se trata de descifrar sin la clave, seguirá sin verse nada. La desventaja de este algoritmo es que solo sirve para aquellas imágenes que tienen mismo ancho y misma longitud, pero han realizado pruebas y llegan a la conclusión de que es un algoritmo robusto [59].

Cho et al. (2016) nos presentan una arquitectura basada en una doble capa para el entorno de trabajo con big data. Estas dos capas que proponen son: capa pre-filtrado y la capa post-filtrado. La primera de ellas trabaja en la etapa de recolección de información de big data. Se encarga de buscar y eliminar la información personal sensible de los datos recolectados. Lo hace con la finalidad de anonimizar la información y que así resulte más complicado identificar a la persona en particular. La segunda capa, post-filtrado, enmascara la información sensible sintetizada después del análisis de big data [60].

Peng Jing (2014) comenta el uso creciente de la nube como lugar de almacenaje. Para mejorar la seguridad, propone un sistema de doble cifrado de los datos. Este doble cifrado consiste en un cifrado inicial utilizando el algoritmo de cifrado AES, y por tanto un cifrado simétrico. Posteriormente, utiliza el algoritmo RSA como cifrado asimétrico. Con el cifrado asimétrico se generan dos claves, la pública y la privada. La privada la poseen los usuarios y la utilizan para descifrar la información. De esta forma consiguen que solo ellos sean capaces de obtener los datos [61].

Subashini and Kavitha (2011) explican un modelo de seguridad que no evita que se hackee la base de datos, sino que los datos que obtengan no tengan valor. Ponen como ejemplo el login y contraseña de un usuario, en el cual estos dos datos sin relación y por separado no tienen ningún valor.

Su modelo consiste en dividir los datos almacenados en Public Data Segment (PDS) and Sensitive Data Segment (SDS). Los SDS deben ser fragmentados aún más, pero no de cualquier forma. Mediante el algoritmo que describen y explican, dividen estos datos de forma que dejen de tener valor individualmente [62].

Hingwe et al. (2014) explican un modelo de base de datos en la nube con una arquitectura de dos capas adicionales, las cuales se usan dependiendo si los datos son sensibles o no. Junto a las capas, las añaden un cifrado a los datos. Este cifrado es doble si se trata de información sensible. Para estos cifrados se necesita una clave, y utilizan una clave simétrica proporcionada por el servidor de la base de datos. Para mejorarlo, cuando se trata de información sensible, y por tanto usa las dos capas, esta clave se divide en dos por medio de un algoritmo [63].

Cheng et al. (2014) realizan un resumen de las amenazas más directas que puede sufrir un cliente de un proveedor de nube. Entre estas amenazas están que el mismo proveedor utilice sus datos para intereses propios o que atacantes adquieran estos datos. Para hacer frente a estas amenazas proponen un esquema sencillo. Se trata de dividir los datos en fragmentos. Tras realizar una función hash, estos fragmentos se empaquetan. Lo que hace que este esquema funcione es que estos paquetes son distribuidos

aleatoriamente entre diferentes puntos de almacenaje, de tal forma que por sí solos estos paquetes no poseen información útil [9].

Thilakanatha el at. (2014) nos presentan un modelo de seguridad para ser utilizado en la monitorización de pacientes a través de dispositivos remotos, como son móviles, pulseras, etc. Este modelo hace uso de un doble cifrado, un cifrado simétrico y un cifrado usando el algoritmo de ElGamal. Los dispositivos móviles generan los datos del paciente y estos datos son cifrados usando una clave simétrica. Para mejorar esta seguridad, se usa el segundo cifrado, el cual es asimétrico y su función será la de cifrar la clave pública usada. El contra del uso de una clave simétrica es que se pierde la identidad de los datos [64].

Bohli et al. (2013) presentan un modelo de diferentes patrones de arquitectura para recursos distribuidos sobre múltiples proveedores de nube. Distinguen cuatro patrones diferentes y realizan un análisis de los detalles y mejoras de estos patrones. Estos son: 1) replicación de aplicaciones, que permite recibir múltiples resultados de una operación realizada en diferentes nubes y compararlo con su propio resultado, lo cual da al usuario pruebas de la integridad del resultado. 2) división de la aplicación en diferentes niveles, que permite separar la parte lógica de los datos. 3) división de la aplicación lógica en fragmentos, que permite la distribución de la aplicación lógica en diferentes nubes. Como beneficio,

ningún proveedor de nube tiene la lógica completa de la aplicación y tampoco pueden saber el resultado final de las operaciones. 4) división de los datos de la aplicación en fragmentos, permite la distribución de los datos sobre distintas nubes, lo cual permite que ninguno tenga acceso a la totalidad de los datos [65].

Yang et al. (2013) realizan un experimento para comprobar la seguridad de plataformas de computación en la nube basado en diferentes niveles de seguridad. Para ello determinan la división de tres niveles, nivel 1) encargado de la seguridad de los datos, como el almacenamiento en la nube, los back up de datos, e-mail, etc. Nivel 2) además de tener en consideración la seguridad, lo combina con la eficiencia. Nivel 3) tiene en consideración la velocidad de procesado de los datos [66].

Ahora que hemos analizado los sistemas más recientes de seguridad en la nube, vamos a realizar de forma teórica una aproximación para llevarlos a la práctica en un entorno que los requiera. Este entorno esta explicado y desarrollado en [67] y pertenece al ámbito de la sanidad. Un ámbito que afecta a la población de todos los países. En este modelo proponen de forma teórica una agrupación de información de diferentes hospitales y clínicas situadas en la provincia de Valladolid, España, y los cuales almacenan la información de los pacientes en un servidor de almacenamiento en la nube. Además, este esquema podrá ser utilizado para

escalarlo a cualquier grupo de hospitales o clínicas de los países. Para dotarlo de seguridad, proponen sistemas como el uso de tarjetas identificativas o cortafuegos. Al tratarse de un tema de relativa importancia, creemos que es necesario aumentar la seguridad de la información.

Antes de exponer teóricamente una mejora de seguridad, vamos a mencionar las tres posibles amenazas clasificadas según origen. Estas amenazas son: a) agente interno, el cual pertenece al sistema sanitario y posee autorización para obtener la información pero la usa para final no éticos, b) agente intermediario, el cual lo identificamos como miembro del grupo poseedor del sistema de almacenamiento de información, y c) agente externo, el cual es cualquier persona ajena al sistema sanitario y/o de almacenamiento. La mejora que propondremos estará enfocada a mejorar la seguridad frente a los agentes intermediarios y los agentes externos.

La información que vamos a querer proteger consiste en un grupo de datos expresados en forma de texto o en forma de imágenes. La primera propuesta teórica es utilizar el esquema de doble capa de Subashini and Kavitha (2011) mediante el cual dividiremos los datos de texto de tal forma que no tengan valor por si solos. Con esto ya conseguimos una cierta anonimación de los datos, algo importante en sanidad. Como alternativa a este esquema, se puede utilizar el expuesto por Cho et al. (2016), el cual

trata de dividir los datos entre sensibles y no-sensible pero para conseguir la misma finalidad, anonimizar la información. Este paso será el primero de todos. El siguiente paso sería utilizar la propuesta de Peng Jing (2014) en la cual expone la utilización de un doble cifrado para la protección de datos. Ahora que tenemos los datos divididos, un cifrado ayudaría a que los agentes intermediarios y externos no obtengan información en texto plano. Aunque usáramos la división de datos, puede llegar a un punto en que estos obtuvieran toda la información y solo necesitaran reordenarla. Por tanto usamos estos cifrados para que ya no resulte tan fácil obtener los datos en texto plano. Por otra parte y como complemento, debido a que las imágenes suele no ser preferible dividirlas en fragmentos, se puede usar el algoritmo desarrollado por Zhou et al. (2015). Las imágenes médicas tienen tanto valor como los informes en texto, por tanto también son necesarias protegerlas. El algoritmo que proponen es un buen sistema para que solo personas autorizadas sean capaces de descifrar correctamente las imágenes. Por desgracia y como se explicó anteriormente, hay que tener cuidado con el tipo de imágenes pues no es posible usar el algoritmo para todas.

Con esto finalizamos la propuesta teórica de mejora, la cual es un complemento al modelo usado, no una alternativa. El haber mencionado estos esquemas y algoritmos no significa que otros no sean posible usarlos,

pero estas personas han invertido mucho tiempo y esfuerzo en desarrollarlos y si funcionan merecen ser usados.

En la Figura 6 explicamos de forma visual las dos alternativas propuestas de primera etapa en el momento de dividir la información de texto generada. Mientras que el cifrado de texto e imágenes es el mismo para ambas situaciones.

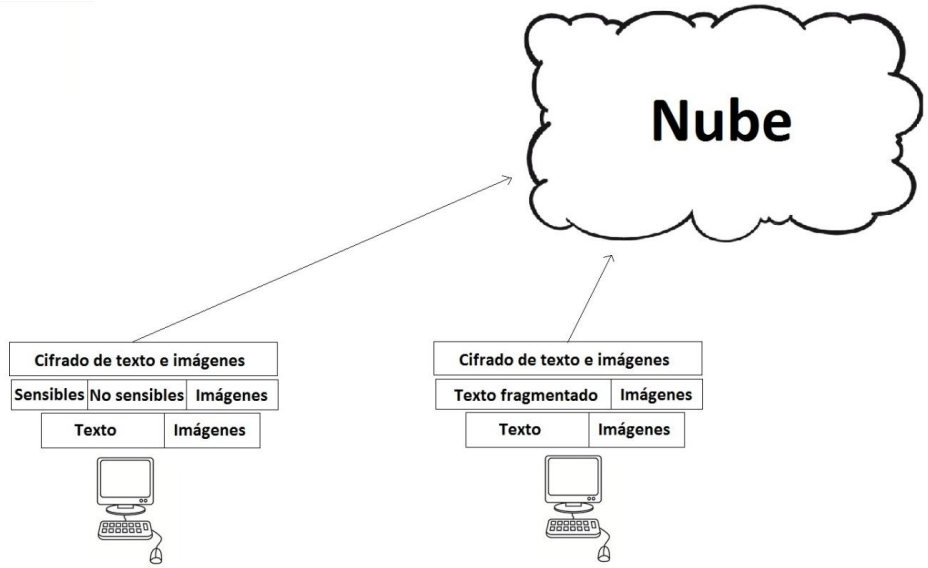

Figura 6. Esquema teórico propuesto

Estas dos alternativas se dividen en 3 capas, siendo la capa 1 la inferior, compuesta por el texto y las imágenes, y siendo la capa 3 en la cual ciframos los datos. En la primera capa simplemente representamos el tipo de información que tenemos, en este caso nuestros datos serán una composición de texto e imagen. En la alternativa de la izquierda tenemos la

utilización de la propuesta de Cho et al. (2016) en la cual el texto lo dividimos en información sensible, la cual representa datos personales, e información no sensible, la cual es algo genérico al conjunto de las personas como puede ser una etiqueta de "nombre". En la alternativa de la derecha tenemos la utilización de la propuesta de Subashini and Kavitha (2011) en la cual el texto lo fragmentamos tantas veces que deje de tener valor. Con ello conseguimos que si nos capturan algunos de estos fragmentos, no tengan significado por si solos y mantengamos la privacidad. En ambos casos, las imágenes las mantenemos sin variar ni fragmentar debido a que tenemos el algoritmo desarrollado por Zhou et al. (2015) para realizar sobre ellas un cifrado eficiente y que sin la clave adecuada no se pueda obtener ningún fragmento de la imagen. Todos estos datos ya cifrados pasaran a través de la red hasta ser almacenados en el servidor de la nube.

Capítulo 5. CONCLUSIONES

Capítulo 5. CONCLUSIONES

Big data supone una novedad en todos los sectores pero representa un gran avance en la sanidad. La cantidad de datos que se generan diariamente necesitan ser almacenados y procesados. Por ello, la arquitectura de la nube está ganando mucha fuerza. Los datos generados en sanidad son de vital importancia puesto que nos pueden permitir una sanidad personalizada, haciendo los tratamientos de las personas específicos para que alcancen su máximo nivel de eficacia al combatir la enfermedad. No solo nos van a ayudar a encontrar un tratamiento personalizado, sino a reducir los costes de la sanidad. Con ello no solo incluimos tratamientos, sino un posible diagnóstico más fiable y un análisis de los ensayos clínicos más preciso y más barato. También nos va a ayudar a determinar factores de riesgos en los ensayos realizados o en epidemias, como puede ser la gripe, especular sobre la evolución del virus y su propagación por los países que la sufran. Todos estos datos es necesario protegerlos, un mal uso de los mismos puede ser una catástrofe. Si mafias obtienen datos clínicos de pacientes importantes, pueden chantajear y ganar gran poder. Farmacéuticas que se hagan con datos de ensayos clínicos pueden alterar los resultados para falsificar lo positivo del producto y acrecentar lo efectos adversos. Por razones similares, se pueden alterar los datos sobre propagación de una epidemia haciendo que el pronóstico de evolución sea peor de lo que sería,

creando pánico entre la población. Estas suposiciones nos llevan a la conclusión de que es necesario sistemas de seguridad, algo que impida o simplemente no resulte atractivo la idea de robar estos datos. A causa de la desconfianza nace la seguridad, mecanismos que hemos resumido anteriormente. Llegamos a la conclusión de que no existe la seguridad perfecta, pues los métodos que usan se adaptan al sector en el que están y las necesidades del momento. Con el paso de los años, la tecnología avanza a grandes pasos, lo que puede ayudar a crear algoritmos que actualmente no pueden ser usados por la carga computacional que requieren, pero esta tecnología es la misma para los atacantes, haciendo que cada vez requieran menos tiempo para descubrir las claves. Como la seguridad nunca llegara a ser perfecta, solo podemos diseñar métodos para que sea más complicado. Muchos de estos métodos, como vimos en los resultados asociados a la seguridad, se desarrollan para entornos concretos y para tipos específicos de ficheros, como pueden ser imágenes o texto de características específicas. Con el paso de los años, la seguridad es algo que estará en el orden del día de cualquier empresa e invertirá en mejorarla.

Capítulo 6. BIBLIOGRAFÍA

Capítulo 6. BIBLIOGRAFÍA

1. Logica B., Magdalena R. Using Big Data in the Academic Environment. Procedia Economics and Finance 2015;33: 277 – 286
2. Chang V, Wills G. A model to compare cloud and non-cloud storage of Big Data. Future Generation Computer Systems 2016;57: 56–76
3. Kshetri N. Big data's impact on privacy, security and consumer welfare. Telecommunications Policy 2014; 38: 1134–1145.
4. Andreu-Perez J, Poon C, Merrifield R, Wong S, and Yang G. Big Data for Health IEEE Journal of biomedical and health informatics, VOL. 19, NO. 4, JULY 2015
5. Goli-Malekabadi Z, Sargolzaei-Javan M, Kazem Akbari M. An effective model for store and retrieve big health data in cloud computing. Computer methods and programs in biomedicine 2016;132: 75–82
6. Leavitt N.,Will NoSQL Databases Live Up to Their Promise, IEEE Computer Society, 2010.
7. Atzeni P., Bugiotti F., Rossi L. Uniform access to NoSQL systems. Information Systems 2014;43: 117–133
8. Huang T, Lan L, Fang X, An P, Min J, Wang F. Promises and Challenges of Big Data Computing in Health Sciences. Big Data Research 2015;2:2–11.
9. Cheng H, Wang W, Rong C. Privacy Protection Beyond Encryption for Cloud Big Data. 2nd International Conference on Information Technology and Electronic Commerce 188-191
10. Fatemi Moghaddam F., Baradaran Rohani M., Ahmadi M., Khodadadi T., Madadipouya K. Cloud Computing: Vision, Architecture and Characteristics.

IEEE 6th Control and System Graduate Research Colloquium, Aug. 10 - 11, UiTM, Shah Alam, Malaysia 2015

11. O'Driscoll A, Daugelaite J, Sleator R. 'Big data', Hadoop and cloud computing in genomics. Journal of Biomedical Informatics 2013;46:774–781.
12. https://hadoop.apache.org/
13. Cunhaa J, Silvaa C , Antunesa M ; Health Twitter Big Bata Management with Hadoop Framework ; Procedia Computer Science 64 (2015) 425 – 431
14. Blanke T ; 5 - Big data collecting ; Digital Asset Ecosystems, 2014, Pages 87-117 ;
15. Parra Calderón CL. Big data in health in Spain: now is the time for a national strategy. Gac Sanit 2016;30(1):63–65.
16. Ting Wong H, Yin Q, Qi Guo Y, Murray K, Hau Zhou D, Slade D. Big data as a new approach in emergency medicine research. Journal of Acute Disease 2015;4(3):178–179.
17. Costa F. Big data in biomedicine. Drug Discovery Today 2014;19(4):433-440.
18. Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D. Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives ; Hindawi Publishing Corporation BioMed Research International Volume 2014, Article ID 134023, 1-13 http://dx.doi.org/10.1155/2014/134023
19. Ahmad A, Paul A, Rathore M; An efficient divide-and-conquer approach for big data analytics in machine-to-machine communication; Neurocomputing 174 (2016) 439 – 453
20. Chen M, Mao S, Liu Y ; Big Data: A Survey ; Mobile Netw Appl (2014) 19:171–209

Código de campo cambiado

21. Archenaa J, Anita M ; A Survey Of Big Data Analytics in Healthcare and Government ; Procedia Computer Science 50 (2015) 408 – 413
22. Young S ; A “big data” approach to HIV epidemiology and prevention ; Preventive Medicine 70 (2015) 17–18
23. Kumar S, Eswari , Sampath , Lavanya S ; Predictive Methodology for Diabetic Data Analysis in Big Data ; Procedia Computer Science 50 (2015) 203 – 208
24. Scopus. Available from: http://www.scopus.com/ (last accessed May 2016).
25. PubMed. Available from: http://www.ncbi.nlm.nih.gov/pubmed/advanced (last accessed May 2016).
26. Science Direct. Available from http://www.sciencedirect.com/ (last accessed May 2016).
27. Web of Science. Available from https://www.accesowok.fecyt.es (last accessed May 2016).
28. Available from http://ieeexplore.ieee.org/search/advsearch.jsp (last accessed May 2016)
29. Clarke R ; Big data, big risks ; Info Systems J (2016) 26, 77–90
30. Trifiletti D, Showalter T ; Big Data and Comparative Effectiveness Research in Radiation Oncology: Synergy and Accelerated Discovery ; Front. Oncol. 5:274. doi: 10.3389/fonc.2015.00274
31. Wyber R, Vaillancourt S, Perry W, Mannava P, Folaranmi T, Celi L ; Big data in global health: improving health in low- and middle-income countries ; Bull World Health Organ 2015;93:203–208
32. Hood L, Lovejoy J, Price N ; Integrating big data and actionable health coaching to optimize wellness ; Hood et al. BMC Medicine (2015) 13:4 DOI 10.1186/s12916-014-0238-7

Código de campo cambiado
Código de campo cambiado
Código de campo cambiado
Código de campo cambiado
Código de campo cambiado

33. Vayena E, Salathé M, Madoff L, Brownstein J; Ethical Challenges of Big Data in Public Health; PLOS Computational Biology | DOI:10.1371/journal.pcbi.1003904 February 9, 2015
34. Pérez G ; Risks of the use of big data in research in public health and epidemiology ; Gac Sanit. 2016;30(1):66–68
35. Hesse B, Moser R, Riley W ; From Big Data to Knowledge in the Social Sciences ; Ann Am Acad Pol Soc Sci. 2015 May 1; 659(1): 16–32. doi:10.1177/0002716215570007
36. Martínez Sesmero JM. "Big Data"; aplicación y utilidad para el sistema sanitario. Farm Hosp 2015;39(2):69-70.
37. Moskowitz A, McSparron J, Stone D, Celi L ; Preparing a New Generation of Clinicians for the Era of Big Data ; Harv Med Stud Rev. 2015 January ; 2(1): 24–27.
38. Lafuente G; The big data security Challenge ; Network Security January 2015 12-14
39. Elsebakhi E, Leeb F, Schendela E, Haquea A, Kathireasona N, Patharea T, Syeda N, Al-Ali R ; Large-scale machine learning based on functional networks forbiomedical big data with high performance computing platforms ; Journal of Computational Science 11 (2015) 69–81
40. Satell G, "6 Things You Should Know About the Future," 10- THE FUTURE OF ONLINE SECURITY 237-258 (2014)
41. Belle A, Thiagarajan R, Soroushmehr R, Navidi F, Beard D, Najarian K ; Big Data Analytics in Healthcare ; Hindawi Publishing Corporation BioMed Research International Volume 2015, Article ID 370194, 1-16

42. Ladha K, Arora V, Dutton R, Hyder J, Potential and Pitfalls for Big Data in Health Research ; Advances in Anesthesia 33 (2015) 97–111
43. Alyass A, Turcotte M, Meyre D; From big data analysis to personalized medicine for all: challenges and opportunities ; Alyass et al. BMC Medical Genomics (2015) 8:33 DOI 10.1186/s12920-015-0108-y
44. Jin X, Waha B, Chenga X, Wanga Y; Significance and Challenges of Big Data Research ; Big Data Research 2 (2015) 59–64
45. Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, Guyer M, Green E ; The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data ; J Am Med Inform Assoc 2014;21:957–958
46. Zhang X, Liu C, Nepal S, Yang C, Chen J S., Privacy Preservation over Big Data in Cloud Systems Security, Nepal and M. Pathan (eds.) Privacy and Trust in Cloud Systems, 239 DOI: 10.1007/978-3-642-38586-5_8
47. Otero P, Hersh W, Ganesh J ; Big Data: Are Biomedical and Health Informatics Training Programs Ready? ; Yearb Med Inform 2014:177-81
48. Kemp R ; Legal aspects of managing Big Data ; computer law & security review 30(2014) 482-491
49. Shen Y, Zhang Y; Transmission protocol for secure big data in two-hop wireless networks with cooperative jamming ; Information Sciences 281 (2014) 201–210
50. Chen P, Zhang C; Data-intensive applications, challenges, techniques and technologies: A survey on Big Data ; Information Sciences 275 (2014) 314–347
51. Shin D, Sahama T, Gajanayake R. Secured e-health data retrieval in DaaS and Big Data ; Secured e-health data retrieval in DaaS and Big Data
52. Cumbley R, Church P; Is "Big Data" creepy? ; computer law & security review 29 (2013) 601-609

53. Nambiar R, Sethi A,Bhardwaj R,Vargheese R; A Look at Challenges and Opportunities of Big Data Analytics in Healthcare ; 2013 IEEE International Conference on Big Data 17-22
54. Brinkmann B, Bowera M, Stengel K, Worrell G, Steada M ; Large-scale electrophysiology: Acquisition, compression, encryption, and storage of big data ; Journal of Neuroscience Methods 180 (2009) 185–192
55. Liu C, Yang C, Zhang X, Chen J ; External integrity verification for outsourced big data in cloud and IoT: A big picture ; Future Generation Computer Systems 49 (2015) 58–67
56. Fabiano E, Seo M, Wu X, Douglas C ; OpenDBDDAS Toolkit: Secure MapReduce and Hadoop-like Systems ; Procedia Computer Science Volume 51, 2015,1675–1684
57. Yan Z, Ding W, Niemic V, Vasilakos A ; Two Schemes of Privacy-Preserving Trust Evaluation ; Future Generation Computer Systems (2015)
58. Hsu C, Zeng B, Zhang M; A novel group key transfer for big data security ; Applied Mathematics and Computation 249 (2014) 436–443
59. Zhou G, Zhang D, Liu Y, Yuan Y, Liu Q; A novel image encryption algorithm based on chaos and Line map ; Neurocomputing169(2015)150–157
60. Cho D, Kim S, Yeo S ; Double Privacy Layer Architecture for Big Data Framework ; International Journal of Software Engineering and Its Applications Vol. 10, No. 2 (2016), 271-278
61. Jing P. A New Model of Data Protection on Cloud Storage ; JOURNAL OF NETWORKS, VOL. 9, NO. 3, MARCH 2014 666-671

62. Subashini S, Kavitha V ; A Metadata Based Storage Model For Securing Data In Cloud Environment ; 2011 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery 429-434
63. Hingwe K, Bhanu M ; Two Layered Protection for Sensitive Data in Cloud ; 2014 International Conference on Advances in Computing,Communications and Informatics (ICACCI) 1265-1272
64. Thilakanathan D, Zhao Y, Chen S, Nepal S, Calvo R, Pardo A; Protecting and Analysing Health Care Data on Cloud ; 2014 Second International Conference on Advanced Cloud and Big Data 143-149
65. Bohli J., Gruschka N., Jensen M., Lo Iacono L., Marnau N. Security and Privacy Enhancing Multicloud Architectures. IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 10, NO. 4, JULY/AUGUST 2013
66. Yang F, Pan L., Xiong M., Tang S. Establishment of Security Levels in Trusted Cloud Computing Platforms. 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing
67. De la Torre- Díez I, López-Coronado M, Garcia-Zapirain Soto B, Mendez-Zorrilla A. Secure Cloud-Based Solutions for Different eHealth Services in Spanish Rural Health Centers. J Med Internet Res