Artículo **Revista de Tecnologías Computacionales**
Marzo, 2019 Vol.3 No.9 8-14

Big Data para ayudar a generar alertas tempranas en salud pública. Diseño de una arquitectura de software para sistemas Big Data

Big Data to help generate early public health alerts. Design of a software architecture for Big Data systems

MENDOZA-GONZÁLEZ, Omar†*, HERNÁNDEZ-CABRERA, Jesús y DE LA GARZA-BARROSO, Ana Lucia

Universidad Nacional Autónoma de México

ID 1er Autor: *Omar, Mendoza-González* / **ORC ID**: 0000-0002-3492-4549, **CVU CONACYT ID:** 972783

ID 1er Coautor: *Jesús, Hernández-Cabrera* / **ORC ID**0000-0002-7850-858X, **CVU CONACYT ID:** 104580

ID 2do Coautor: *Ana Lucia, De La Garza-Barroso* / **ORC ID** -0000-0002-2184-2587, **CVU CONACYT ID:** 428562

DOI: 10.35429/JOCT.2019.9.3.8.14 Recibido 22 de Enero, 2019; Aceptado 12 Marzo, 2019

Resumen

La producción masiva de datos en distintos formatos y provenientes de diversas fuentes, gubernamentales, sociales, legales ha creado la posibilidad de que las instituciones de gobierno en México, puedan tener una visión clara de lo que la sociedad opina sobre temas específicos. En salud pública, estos datos son la base para generar indicadores de alertas sobre brotes de enfermedades en diversas regiones o comunidades, basados en conceptos de inteligencia epidemiológica. El problema al que se enfrentan estas instituciones es la falta de una arquitectura de sistemas de software adecuada para colectar, catalogar y analizar los enormes volúmenes de datos generados a fin de ser integrados al proceso de crear alertas tempranas que generen en tiempo real indicadores que permitan tomar mejores decisiones y rutas de acción a las autoridades sanitaria de nuestro país. El objetivo es diseñar un sistema de Big Data cubriendo los cuatro requisitos principales del tratamiento a gran escala de datos: 1 Soporte de grandes cargas de trabajo de escritura desde fuentes diversas. 2 una arquitectura elástica, capaz de soportar picos ocasiones de carga de trabajo y agregar o liberar recursos según sea necesario. 3 Soportar el análisis intensivo, para poder admitir grandes y diversas solicitudes de lectura. Alta disponibilidad para soportar errores presentes en hardware y software.

Big Data, Ingeniería de Software, Salud Pública, Inteligencia Epidemiológica

Abstract

The massive production of data in different formats and sources, governmental, social, and legal has created the possibility that government institutions in México can have a clear vision of what society thinks about specific issues. In public health, these data are the base for generating alert indicators on outbreaks of diseases in various regions or communities, based on epidemiological intelligence concepts. The problem that institutions face is the lack of an architecture of software systems suitable for collect, catalog and analyze for to take better decisions and action routes to the health authorities of our country. The objective is to design a Big Data system covering the four main requirements of large-scale data processing. 1 Support large writing workloads from various sources. 2 An elastic architecture, capable of withstanding the peak load times of work and adding or releasing resources as needed. 3 Support intensive analysis, to be able to admit large and diverse reading requests. 3 High availability to support errors in hardware and software.

Big Data, Software Engineering, Public Health, Epidemiological Intelligence

Citación: MENDOZA-GONZÁLEZ, Omar, HERNÁNDEZ-CABRERA, Jesús y DE LA GARZA-BARROSO, Ana Lucia. Big Data para ayudar a generar alertas tempranas en salud pública. Diseño de una arquitectura de software para sistemas Big Data. Revista de Tecnologías Computacionales. 2019 3-9: 8-14

* Correspondencia del Autor: (Correo electrónico: omendoza@comunidad.unam.mx)
† Investigador contribuyendo como primer autor.

Introducción

La globalización es una parte importante de la salud pública, vista como un determinante que participa en la gestión de riesgos para la transferencia de enfermedades a nivel internacional(Franco-Giraldo, 2009). Es importante resaltar, que en los últimos años, la evidencia ha concluido que la propagación de enfermedades infecciosas en el mundo es cada vez mayor, así mismo, la velocidad con la que se propagan dichas enfermedades también aumenta con el tiempo; según la Organización Mundial de la Salud (OMS)(Organización Mundial de la Salud, 2007), desde los años setenta, se han identificado una o más enfermedades nuevas por año y se han verificado más de 1,100 eventos epidémicos(Olaya, 2010).

La respuesta humana ante estas enfermedades depende precisamente de la rapidez con la que se detectan; el conocer oportunamente que existe un brote o una epidemia en la población generalmente depende de sistemas de vigilancia establecidos, a través de los cuales es posible observar el comportamiento de los padecimientos en la población(Woolhouse ME, 2015). De manera clásica, la vigilancia se define como un proceso continuo y sistemático de recolección, integración, administración, análisis, interpretación y difusión de datos para la identificación de patrones de la enfermedad, así como determinantes en tiempo, lugar y persona(Castillo, 2002).

Los sistemas de vigilancia epidemiológica que realizan este proceso de vigilancia están basados en indicadores preestablecidos en cada país; este tipo de vigilancia puede ser considerada rutinaria y sistemática ya que utiliza datos recolectados de forma pasiva; lo cual, puede resultar relativamente complicado para tomar decisiones rápidas en la implementación de acciones de control, por ejemplo, en enfermedades emergentes y en eventos de salud pública que requieran de una respuesta acelerada. En este contexto, se han desarrollado nuevos métodos y herramientas orientadas a detectar anticipadamente patrones o señales que revelen la aparición o propagación de amenazas a la salud pública con mayor antelación. Esta modalidad de detección, alerta y respuesta se denomina alerta temprana(Castillo, 2002).

Una alerta temprana puede indicar la ocurrencia de un riesgo o peligro de algún tipo para la salud de la población. El concepto de alerta temprana en inteligencia epidemiológica puede enmarcarse en la vigilancia basada en eventos. A través de este tipo de vigilancia es posible detectar señales que potencialmente indiquen la ocurrencia de un evento de importancia para la salud pública anticipadamente(Castillo, 2002). Como tal, una señal puede definirse como aquellos "datos o información que el mecanismo de alerta temprana y respuesta consideren indicativos de un posible riesgo agudo para la salud humana"(Paquet C, 2006).

Con el advenimiento de la World Wide Web, se ha hecho necesario aprovechar este cúmulo de datos con el objetivo de detectar información que potencialmente pueda traducirse como una "alerta temprana"(Castillo, 2002). Se ha definido al internet como una red mundial interconectada e integrada por múltiples redes de todo el mundo que es utilizada como una herramienta de acceso a *datos e información* de múltiples temáticas, entre las que se destaca la salud(Organización Mundial de la Salud, 2014).

Esta red, "se ha ido expandiendo cada vez más a diferentes campos relacionados con la salud pública"[11] convirtiéndose también en un importante instrumento para el intercambio de información en salud, por lo que puede constituir una oportunidad importante para establecer nuevos enfoques o metodologías para fortalecer a la vigilancia basada en eventos, y con ello al campo de la vigilancia en salud pública, disciplina, que se ha basado tradicionalmente en sistemas establecidos en la notificación obligatoria y voluntaria de las enfermedades. Las innovaciones para detectar alertas tempranas en fuentes de información en internet podrían conducir a un reconocimiento más rápido de riesgos para la población, incluidas las enfermedades transmisibles como eventos emergentes o reemergentes.[12, 13]

Toda esta colección masiva de datos que se genera en la Web hace referencia al término Big Data (Grandes Datos) y "representa a los activos de información que se caracterizan por un volumen alto, velocidad y variedad que requieren de tecnología específica y métodos analíticos para su transformación en valor".

ISSN 2523-6814

Estos datos, aunque difíciles de procesar, su análisis es posible mediante el uso de herramientas especiales para el procesamiento de datos.[14]

El término Big Data surgió inicialmente en 2001 como activos formados por información de gran volumen, velocidad y variedad[19], las descripciones de los elementos que componen el término Big Data generalmente se explican como activos formados por información. Se caracteriza por tener un gran volumen de datos que forman la velocidad con la que se forman y la variedad que adquieren, requieren tecnologías específicas y métodos analíticos propios para procesar el valor[20].

El Big Data en la Web puede ser generado a partir de una diversidad de distintas fuentes y métodos, desde los "clicks" en páginas de Internet hasta en transacciones móviles, o en contenidos que son generados por usuarios de internet o a través de medios sociales de comunicación.[15]

El uso de la información y datos provenientes de internet y de los medios sociales tienen múltiples beneficios, se destacan algunas ventajas que hacen referencia a las características de Big Data. Los datos son obtenidos directamente de la población (incluye expertos, hospitales, instituciones, medios de comunicación social) y existe una gran cantidad de los mismos (Big Volume); además existe gran variedad de datos circulando (Big Variety); en cuanto a la veracidad de los mismos puede variar ya que algunos datos pueden tener un valor bajo (Veracity) pero mediante las técnicas de minería de datos para su extracción es posible agregar mayor valor a la información (Big Value).[16]

La detección oportuna y alerta temprana de enfermedades o eventos de interés para la salud pública son atributos sumamente importantes de la Inteligencia Epidemiológica; a diferencia de la vigilancia convencional, el marco de la Inteligencia permite la utilización de información fuera de los canales establecidos a través de la vigilancia basada en eventos, debido a esta característica existe la posibilidad de detectar más rápidamente amenazas para la salud pública, la exploración continua de datos contenidos en diferentes fuentes de información es necesaria con este propósito.

Para este caso, se analizará información proveniente de medios sociales virtuales debido a que representan plataformas en donde existe un intercambio de abundante información de forma permanente y constante. Además, se destacan otras ventajas de la misma como es, su fácil acceso y su procedencia directa de usuarios o instituciones quienes publican directamente sobre su estado de salud o sobre eventos de salud pública en su entorno. Monitorear este tipo de datos puede agilizar el proceso de detección oportuna de situaciones de salud que ocurren entre la población.

El uso de este tipo de datos es un área de reciente estudio para la salud pública por lo que, además, se requiere de investigación para el establecimiento de metodologías específicas para su análisis.

Problemática

La detección oportuna y alerta temprana de enfermedades o eventos de interés para la salud pública son atributos sumamente importantes de la Inteligencia Epidemiológica; a diferencia de la vigilancia convencional, el marco de la Inteligencia permite la utilización de información fuera de los canales establecidos a través de la vigilancia basada en eventos, debido a esta característica existe la posibilidad de detectar más rápidamente amenazas para la salud pública, la exploración continua de datos contenidos en diferentes fuentes de información es necesaria con este propósito.

El problema al que se enfrentan las instituciones de salud pública en México es la falta de una arquitectura de software adecuada para colectar, catalogar y analizar los enormes volúmenes de datos generados en las diversas fuentes descritas, a fin de ser integrados al proceso de generación de alertas tempranas que produzcan en tiempo real indicadores, como por ejemplo Notificación inmediata de brotes y eventos en menos de 24 horas, que a su vez permitan tomar mejores decisiones y rutas de acción a las autoridades sanitarias de nuestro país. La creciente complejidad de las nuevas y diversas fuentes y una extensa variedad de tipos de datos crea un reto de alta complejidad, debido a que los datos son cada vez más inciertos e impactan directamente en la recolección, almacenamiento, procesamiento y análisis de los mismos.

ISSN 2523-6814

Objetivo

Diseñar un sistema de software para colectar, catalogar y analizar datos generados en fuentes no oficiales para la vigilancia epidemiológica, en un entorno big data cubriendo los cuatro requisitos principales del tratamiento a gran escala de datos

1. Soportar grandes cargas de trabajo de escritura desde fuentes diversas.
2. Crear una arquitectura elástica, capaz de soportar picos ocasiones de carga de trabajo y agregar o liberar recursos según sea necesario.
3. Soportar el análisis intensivo, para poder admitir grandes y diversas solicitudes de lectura.
4. Contar con alta disponibilidad para soportar errores presentes en hardware y software

Justificación

En México se cuenta con el Sistema Nacional de Vigilancia Epidemiológica (SINAVE), el cual requiere de un continuo esfuerzo para implementar nuevas metodologías para la detección temprana de riesgos o eventos que puedan afectar a la salud de la población. Una parte fundamental para la detección de eventos en este sistema se da en la Unidad de Inteligencia Epidemiológica y Sanitaria (UIES) en donde se pueden llegar a identificar tempranamente brotes o eventos de salud pública a través del monitoreo permanente de fuentes de información formales e informales durante las 24 horas al día y 365 días al año.

Para ello, la UIES revisa de forma rutinaria medios de comunicación, redes sociales y diversos sitios en internet, lo que permite detectar en este tipo de medios, eventos epidemiológicos críticos nacionales e internacionales y desastres naturales que pongan en riesgo la salud de la población del país, sin embargo, precisa de implementar nuevas metodologías y enfoques que mejoren la oportunidad en la detección de dichos eventos.

Los diseños de software basados en el entorno big data pueden ser utilizados como parte de la vigilancia basada en eventos en el SINAVE específicamente en la UIES; esto permitiría la introducción de un nuevo campo de investigación a las actividades propias de la unidad y la integración de nuevas tecnologías.

Los servicios de Internet se utilizan cada vez más en la vida cotidiana, como parte de esta utilización creciente de la red, en los medios sociales virtuales también se registran más usuarios diariamente; esto, ha generado que su contenido se incremente de manera exponencial y se cuente con información almacenada en diferentes tipos de formatos en estos medios.

El contenido que se encuentra en las redes sociales virtuales tiene además atributos que permiten llevar a cabo su descripción y análisis (información pública, gratuita, accesible, publicada posiblemente en tiempo real y de manera rápida) por lo que el uso de la misma puede representar una oportunidad para llevar a cabo el vigilancia epidemiológica a partir de estos datos y además la posibilidad de implementar nuevas técnicas que pudieran ser aplicadas al marco de la vigilancia basada en eventos justificado en la Inteligencia Epidemiológica enfatizando la importancia de poder contar con información y datos de manera rápida y oportuna en salud pública.

Además, la aparición de enfermedades emergentes es cada vez más frecuente y se da con mayor velocidad en todo el mundo, esto hace necesario contar con mecanismos que apoyen la identificación de forma más rápida y precisa de dichos eventos que puedan constituir un riesgo para la población; las herramientas disponibles de una arquitectura big data, son una oportunidad para fortalecer y complementar a los sistemas tradicionales de vigilancia de los países a través de nuevas herramientas informáticas reforzadas con herramientas de análisis descriptivas y predictivas mediante la búsqueda de patrones dentro de grandes repositorios de datos. La abundante información en salud que actualmente se encuentra circulando en las redes sociales representa un reto importante para los investigadores y para los profesionales de la salud, es importante continuar realizando estudios que permitan obtener más conocimiento sobre el uso de este tipo de datos y su aplicación en salud pública.

Propuesta

La arquitectura de información basada en big data que se propone pretende mejorar el nivel de confianza que los usuarios tienen en la información generada, garantizar la coherencia de los datos y establecer políticas de protección sobre la información.

ISSN 2523-6814

Cuando se confía en la información, las organizaciones pueden optimizar los resultados. No es solo una propuesta para generar analítica confiable, es una propuesta para generar datos confiables. Los pilares de una plataforma de Big Data son:

1. Integración: con la finalidad de contar con una plataforma para administrar todos los datos, evitando silos de datos personalizados.
2. Analítica, el Big Data debe ser una plataforma viable para analizar y almacenar datos. Esta tecnología permite ir más allá del preprocesado de datos Data Warehouse y se emplea el concepto de Data Lake. El análisis profundo es un área importante de valor agregado cuidando en todo momento la sofisticación y precisión del análisis de datos.
3. Visualización: a fin de acercar el big data a los usuarios finales, por medio de tableros, gráficos y reportes claros y bien definidos acorde a las necesidades de información de los distintos perfiles de usuarios.
4. Desarrollo: uso de herramientas de desarrollo sofisticadas integradas con el motor principal para big data, Hadoop a fin de que puedan desarrollarse aplicaciones de ingesta, procesamiento y analítica de manera rápida, ágil e integrada.
5. Optimización de la carga de trabajo: mejoras en el por medio del uso de herramientas de código abierto para un procesamiento y almacenamiento eficientes de datos de todos los formatos.
6. Seguridad y gobernanza: Existen datos confidenciales que deben protegerse mediante políticas y normas de retención establecidas correctamente y en distintos niveles, aprovechando la madurez de la gobernabilidad de datos estructurados a fin de beneficiar al entorno de big data.

La propuesta de infraestructura de procesamiento de información y sus capacidades analíticas para tratar con big data contempla la obtención e ingesta de datos usando conectores en tiempo real a fuentes de datos externas no oficiales (redes sociales, páginas web, blogs) y oficiales (SINAVE, sistemas de información de salud estatales, redes de comunicación), y el procesamiento analítico en tiempo real, real-time analytic processing (RTAP).

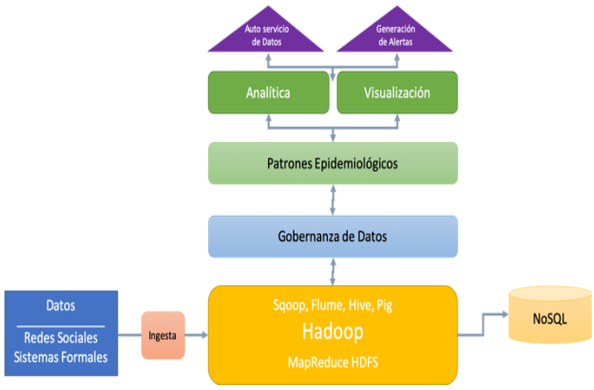

Figura 1 Propuesta de Arquitectura de Software para alertas epidemiológicas basadas en big data
Fuente: Elaboración Propia

La arquitectura general estará compuesta por varios operadores individuales que al trabajar de forma conjunta ofrecen flexibilidad y escalabilidad ya que la base es Hadoop que tiene la capacidad de trabajar de forma distribuida y paralela a fin de particionar los paquetes que manejan las aplicaciones en componentes de software distribuidos a través de hosts de hardware conectados a la red de Hadoop, de manera que si fuera necesario mayor poder de procesamiento se puedan conectar mas nodos en los que se distribuye el trabajo, los servicios de ejecución de streams pueden ayudar a distribuir la carga.

Por tanto, el diseño pretende que la arquitectura este en capacidad de:

1. Procesar cualquier tipo de datos: estructurado, no estructurado, en stream
2. Contar con motores integrados para propósitos específicos
3. Diseñada para manejar diferentes requisitos y patrones de inteligencia epidemiológica
4. Administrar y controlar los datos en el ecosistema de forma correcta por medio de políticas de gobernanza de datos
5. Integración de datos de fuentes externas oficiales y no oficiales para salud pública
6. Trabajar con componentes integrados y servicios basados en estándares
7. Visualizar de forma amigable los datos y resultados del análisis.

ISSN 2523-6814

Propuesta de trabajo práctico

En los lineamentos del SINAVE se especifica que se realiza un monitoreo de medios de comunicación no formales, en la práctica esto se lleva a cabo por un grupo de 4 o 5 personas que se encuentran extrayendo y recabando información de periódicos, noticiarios, correos electrónicos, información proporcionada por la población y medios similares de forma manual, lo que da como resultado una limitada capacidad de obtención, clasificación, procesamiento y extracción de la información que puedan complementar la generada por los medios formales.

En este caso, el presente trabajo de investigación busca apoyar en la práctica el monitoreo de medios no oficiales, específicamente redes sociales e integrarlos con los medios formales para automatizar la extracción de datos y de esta forma complementar y comparar los resultados de ambas fuentes.

Como trabajo práctico se plantea realizar en paralelo un proceso para colectar, catalogar y analizar de datos provenientes de medios no formales empleando la infraestructura Big Data aquí planteada e integrarlos a los datos provenientes de los medios formales.

Los resultados esperados, en primera instancia, es duplicar el número de entradas procesadas provenientes de las redes sociales y en segunda instancia disminuir el tiempo de procesamiento al menos a un tercio del actual, y de esta forma poder generar patrones de asociación entre diversos conceptos a fin de identificar comportamientos relacionados a conceptos de infectología.

Metodología

El primer paso consiste en la ingesta de datos, tomando los datos de las diversas fuentes y formatos, filtrados por conceptos específicos de inteligencia epidemiológica. Estos datos serán almacenados en un Data Lake en Hadoop y los metadatos guardados en tablas Hive.

El segundo paso genera patrones de asociación entre diversos conceptos a fin de identificar comportamientos relacionados a la inteligencia epidemiológica.

El tercer paso permitiría a los usuarios finales explorar los datos a través de una exploración visual e iterativa con herramientas de filtrado y clasificación. La visualización inicial efectiva es la clave. En lugar de tabular la información a través de hojas de cálculo, la interfaz representa los datos gráficamente, con el énfasis en fomentar un entorno de colaboración.

El usuario podrá también analizar los datos a través del análisis iterativo de SQL Like, apoyado en Hive.

Conclusiones

La utilización de datos e información provenientes de fuentes como redes sociales constituye en estos tiempos una parte fundamental para detectar tempranamente eventos relacionados con la seguridad global en salud.

La detección digital de enfermedades puede utilizarse como parte de los sistemas de vigilancia epidemiológica; utilizando datos no estructurados como parte de la vigilancia en el marco de la Inteligencia Epidemiológica, lo cual es acorde a lo que diversos autores como (Paquet C, 2006)establecen.

La información que circula en internet y en las redes sociales incrementa su volumen de forma diaria y, es fundamental implementar y fortalecer la realización de este tipo de análisis por lo que una arquitectura de software adecuada que soporte grandes cargas de trabajo, integración análisis y visualización de datos podría en definitiva ayudar a la generación de alertas tempranas en cualquier sistema nacional de vigilancia epidemiológica; automatizando este proceso de la mayor forma posible para mejorar la toma de decisiones en base a esta información.

Referencias

Castillo, C. S. (2002). Módulos de principios de epidemiología para el control de enfermedades. OPS Washington.

Franco-Giraldo, A. a.-D. (2009). Salud pública global: un desafío a los límites de la salud internacional a propósito de la epidemia de influenza humana A.

ISSN 2523-6814

Olaya, A. P. (2010). Salud Global: Política Pública, derechos sociales y globalidad. Facultad Nacional de Salud Pública, 28(3), 301-302.

Organización Mundial de la Salud. (2007). Informe sobre la salud en el mundo 2007: un porvenir más seguro Protección de la salud pública mundial en el siglo XXI.

Organización Mundial de la Salud. (2014). Detección temprana, evaluación y respuesta ante eventos agudos de salud pública: Puesta en marcha de un mecanismo de alerta temprana y respuesta con énfasis en la vigilancia basada en eventos.

Paquet C, C. D. (2006). Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Eurosurveillance.

Woolhouse ME, R. A. (2015). Lessons from Ebola: improving infectious disease surveillance to inform outbreak management. Sci Translat Med, 7:307rv5.

ISSN 2523-6814