

Tecnología Big Data para el Sector Salud del Estado de Guerrero

Elías Marino Escobar Ayona

Universidad Autónoma de Guerrero
Foro Antiemperialista No. 227. Colonia Tribuna Nacional,
C. P. 39038, Chilpancingo, Guerrero.
emeayona@gmail.com

Abstract. In this paper presents the problem that has the Health Sector in the State of Guerrero in their management, storage and analysis of information generated by medical units that account, so it intends to implement Big Data technology using Hadoop presents this paper, Hbase, MapReduce, Hcatalog and thus reduce costs, with what may be called digital health, have digitized medical records, where could explore different treatments, with the intent to build models applicable to other patients not have any diagnostic.

Keywords. Aplicación de tecnología Big Data en el Sector Salud con software libre.

1 Introducción

Con el paso del tiempo las fuentes de información que manejan las organizaciones no sólo provienen de sus propias bases de datos, hoy en día, reciben importante información de una gran variedad de mecanismos de interacción internos y externos, como son las redes sociales, los blogs, de audio y de video.

Todas estas nuevas fuentes de datos conforma un repositorio muy grande de información que se conoce como Big Data; es decir crecimiento muy rápido de la información en cuanto a volumen, con variadas estructuras y que, por lo mismo, requiere de nuevos elementos tecnológicos para su análisis.

De acuerdo a [8] Big Data implica centros de datos, que reciban, almacenen y procesen información de actividad en las redes sociales, aplicaciones cloud, analíticas, sensores y dispositivos móviles, datos de actividades profesionales, usuarios, clientes y proveedores.

En [1] capturar valor a partir de grandes volúmenes de datos, las organizaciones tendrán que implementar nuevas tecnologías (para el almacenamiento, cómputo y software de análisis) y técnicas (nuevos tipos de análisis). La gama de retos tecnológicos y las prioridades serán diferentes, dependiendo de la madurez de datos de la institución. Los sistemas heredados, los estándares y formatos incompatibles, con demasiada frecuencia impiden la integración de los datos y los análisis más sofisticados que crean valor a partir de grandes cantidades de datos. Nuevos problemas y cada

vez mayor potencia de cálculo estimularán el desarrollo de nuevas técnicas de análisis. También hay una necesidad de innovación continua en tecnologías y técnicas que ayudarán a los individuos y a las organizaciones a integrar, analizar, visualizar, y consumir el creciente torrente de información.

Como es el caso del sector salud en el Estado de Guerrero en donde, las oficinas administrativas de las unidades médicas, generan un gran cumulo de información.

2 Sector Salud del Estado de Guerrero

La Secretaría de Salud en el Estado de Guerrero es la encargada de coordinar, establecer, conducir y ejecutar el Sistema Estatal de Salud e impulsar íntegramente los programas de salud en la entidad, así como ejercer facultades de autoridad sanitaria en su ámbito de competencia (salud pública).

El Sistema Estatal de Salud está orientado a establecer los medios y acciones que tengan por objeto prevenir la enfermedad, proteger y asegurar la salud de los guerrerenses, de acuerdo con el Sistema de Protección Social en Salud [4].

Cuenta con 7 (siete) Jurisdicciones Sanitarias distribuidas en las diferentes regiones del Estado (Región Tierra Caliente, Zona Norte, Zona Centro, Montaña, Costa Grande, Costa Chica y Acapulco), actualmente cuenta con 1,066 unidades (hospitales generales, hospitales materno-infantil, centro estatal de transfusión sanguínea, laboratorio estatal, centro estatal antirrábico, laboratorio de investigación de enfermedades tropicales, centro de prevención y atención de enfermedades de transmisión sexual, institutos estatales de cancerología y oftalmología, así como de alrededor de mil centros de salud de atención básica).

Debido a la creciente poblacional, a la prisa en las salas de emergencia ocupadas y la saturación en los niveles de hospitalización, muchos pacientes sólo reciben una rápida revisión superficial de su salud, lo que en muchas ocasiones se realiza un diagnóstico equivocado y repercute en que el paciente regrese con algún problema de salud más severo, fenómeno que se puede volver recurrente.

Imagine la consulta convencional con el doctor. La labor que hace es documentar la atención médica tomando nota en un papel, poniéndola en un expediente y guardándola en un archivo, pero no hay mucho que hacer con esa información desde el punto de vista de análisis.

Con ese sistema obsoleto, no se puede analizar a toda la población de pacientes que se está atendiendo y menos hacer búsquedas dentro de esa información para descubrir nuevos conocimientos. Para hacerlo, se tendrían que sacar todos los expedientes que hay archivados en los departamentos de los hospitales y centros de salud y pasarlos uno por uno al sistema electrónico para tratar de extraer algún dato en particular.

Actualmente, se está pasando de expedientes físicos a historias clínicas electrónicas y esa información se puede aprovechar para mejorar la atención clínica a los enfermos.

Las unidades médicas al atender a los pacientes, estos acumulan grandes cantidades de datos en distintos formatos, ya sea en papel o en versión electrónica, que por su dispersión (algunos son estructurados y otros no) resultan imposibles de utilizar y

analizar. Big Data ofrece la posibilidad de organizar información de forma efectiva, por lo que se podrían integrar a los datos estructurados ya existentes hoy en día (expedientes personales de los pacientes).

En la figura 1 se muestra gráficamente la información que se genera en las unidades médicas en los institutos, en los laboratorios de estudios e investigación, así como el espacio de almacenamiento que ocuparía, en términos computacionales.

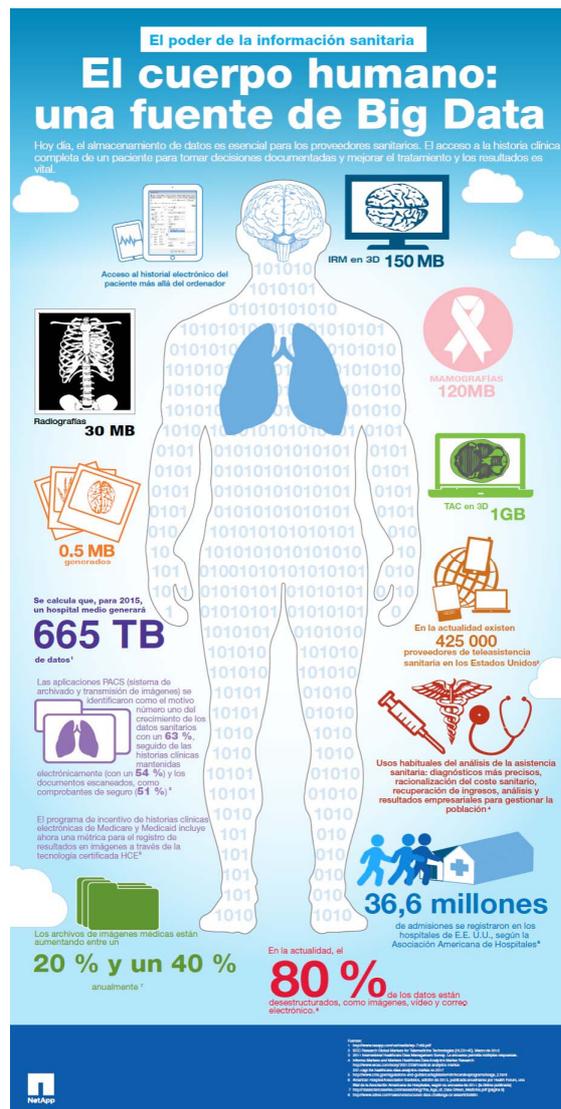


Fig. 1. El cuerpo humano: una fuente de Big Data.

3 Big Data

El concepto Big Data se refiere al tratamiento y análisis de enormes repositorios de datos, tan desproporcionadamente grandes que resulta imposible tratarlos con las herramientas de bases de datos y analíticas convencionales. La tendencia se encuadra en un entorno que no suena para nada extraño: la proliferación de páginas web, aplicaciones de imagen y vídeo, redes sociales, dispositivos móviles, apps, sensores, internet, etc., capaces de generar, según IBM, más de 2.5 quintillones de bytes al día, hasta el punto de que el 90% de los datos del mundo han sido creados durante los últimos dos años.

[8] en su investigación estima que en el año 2010 las empresas y consumidores almacenaron un total de 13 exabytes (EB) de datos, imaginemos que sucederá en el 2020 en cuanto almacenamiento de información, la tabla 1 muestra esta escala de datos.

Table 1. Escala de datos

Unidad	Tamaño	Significado
Bit (b)	1 ó 0	Código binario que utilizan los ordenaros para almacenar y procesar datos
Byte (B)	8 bits	Información suficiente para crear un caracter. Unidad básica de la informática
Kilobyte (KB)	1,024 bytes	Un página de texto ocupa 2kb
Megabyte (MB)	1,024KB	Las obras completas de Shakespare (5MB). Una canción suele tener alrededor de 4MB.
Gigabyte (GB)	1,024MB	Una película de dos horas puede comprimirse entre 1 y 2GB.
Terabyte (TB)	1,024GB	Todos los libros de la biblioteca del congreso estadounidense suma un total de 15TB.
Petabyte (PB)	1,024TB	Todas las cartas entregadas por el servicio postal estadounidense sumarán alrededor de 5PB.
Exabyte (EB)	1,024PB	El equivalente a 10,000 millones de copias de The Economist
Zettabyte (ZB)	1,024EB	Se calcula que al final del año habrá un total de 1.2ZB de información total

Big Data hace referencia a un entorno absolutamente relevante para muchos aspectos, desde el análisis de fenómenos naturales como el clima o de datos sismográficos, hasta entornos como salud, seguridad y el ámbito empresarial.

Big data plantea básicamente tres retos sobre el flujo de datos:

Volumen: saber cómo gestionar e integrar grandes volúmenes de datos, procedentes de fuentes heterogéneas.

Velocidad: poder acceder a la plataforma desde cualquier lugar, de forma autónoma por cualquier usuario de negocio, para mejorar y agilizar la toma de decisiones mediante la automatización: programación de acciones, eventos y alarmas.

Variedad: conseguir unificar contenidos dispersos y no estructurados, con datos históricos, actuales y/o predictivos para un manejo óptimo de los mismos y para extraer de ellos información de valor.

3.1 Técnicas y tecnologías.

Una amplia variedad de técnicas y tecnologías se ha desarrollado y adaptado para agregar, manipular, analizar y visualizar grandes volúmenes de datos [6]. Estas técnicas y tecnologías sacan información de varios campos, incluyendo estadísticas, informática, matemáticas aplicadas y economía. Esto significa que una organización que tiene la intención de obtener valor de big data tiene que adoptar un enfoque flexible y multidisciplinar. Algunas de las técnicas y tecnologías se han desarrollado en un mundo con acceso a volúmenes mucho más pequeños y para una gran variedad de datos, pero se han adaptado con éxito para que sean aplicables a conjuntos muy grandes de datos diversos.

Estas técnicas que se han desarrollado y adaptado para visualizar, analizar y manipular grandes volúmenes de datos, para hacer estos datos manejables, son: fusión de datos, Crowdsourcing, análisis de series de tiempo, pruebas A / B, análisis de redes, análisis de clusters, aprendizaje de reglas de asociación, aprendizaje automático y mucho otras.

Para este caso, se opta por utilizar la tecnología, siendo un marco de software de código abierto para aplicaciones intensivas de datos distribuidos, originalmente creado por Doug Cutting para apoyar su trabajo en Nutch, un motor web de código abierto.

Para cumplir con los requisitos de procesamiento multimáquina de Nutch, Cutting implementó un centro de MapReduce y un sistema de archivos distribuido que en conjunto se convirtieron en Hadoop.

A través de MapReduce, Hadoop distribuye los datos en grandes piezas sobre una serie de nodos que se ejecutan en hardware de consumo.

En la actualidad, Hadoop se encuentra entre las tecnologías más populares para el almacenamiento de los datos estructurados, semi estructurados y no estructurados que forman parte de los grandes volúmenes de datos. Estando disponible bajo Apache License 2.0.



Es la capa de almacenamiento de Hadoop. Sistema de fichero distribuido escrito en java, escalable, tolerante a fallos.

Aunque Hadoop pueda funcionar con varios sistemas de ficheros (sistema de ficheros locales de Linux, GlusterFS, S3 de Amazon...)

HDFS se desmarca de ellos por ser totalmente compatible con MapReduce y ofrecer la optimización de “localidad de los datos”, lo cual lo convierte en la solución “natural” de Hadoop.

HCatalog

Table Management

SequenceFile, JSON, etc).

Ofrece una capa de abstracción de acceso a los datos.

Permite a los usuarios de Hive, Pig o MapReduce acceder fácilmente a cualquier fichero en HDFS sin preocuparse del formato que puede tener este fichero (sea CSV,



Escrito en Java y el modelado con BigTable de Google. Apache HBase es una base de datos no relacional distribuida en columnas, diseñada para ejecutarse en la parte superior de Hadoop Distributed Filesystem (HDFS), proporciona tolerancia a fallas de almacenamiento y un rápido acceso a grandes cantidades de datos dispersos. HBase es uno de los muchos almacenes de datos de NoSQL que se han vuelto disponibles en los últimos años.

4 Implementación de Big Data en la organización

Para la implementación de Big Data en la organización se tiene que tener claro el objetivo, que en este caso es el de concentrar información de los pacientes, que generen los hospitales, institutos, laboratorios, unidades médicas de especialidad, videoconferencias que hace el área de telemedicina y consultas médicas por videoconferencia, todo esto para consultas posteriores, para revisión y análisis de personal médico, de investigación y para análisis estadístico.

Cuando se recolectan historiales individuales, estos datos se pueden utilizar para crear bases de datos inmensas, información que puede ayudar a la continuidad del cuidado del paciente; a que los tratamientos comparados con los resultados se realicen de manera eficiente y sean menos costosos, por tal motivo lo principal es pensar en el almacenamiento de información y el tipo de análisis a aplicar.

Se requiere construir un equipo de trabajo con experiencia técnica y organizativa para el proyecto. Lo ideal sería que el grupo de trabajo esté integrado por profesionales de TI que entienden la tecnología.

Utilizando Hadoop que es un sistema que se puede implementar sobre hardware a un costo relativamente bajo, siendo a su vez totalmente gratuito para software. Siendo que aquella información que antes las empresas no podían procesar por la metodología existente, hoy puedan hacerlo con Hadoop, pudiendo de este modo sacar información nueva y a su vez descubrir y aplicar otro tipo de análisis como por ejemplo, una regresión lineal, sobre millones de registros.

Para la captación de datos, almacenamiento y análisis, como primera etapa, se incluirán a los hospitales, los institutos, los laboratorios, los centros de nueva vida, entre otras unidades médicas, para que toda la información la concentren vía internet al servidor central, establecido en oficinas centrales de la Secretaría de Salud, tal como se muestra en la figura 2.

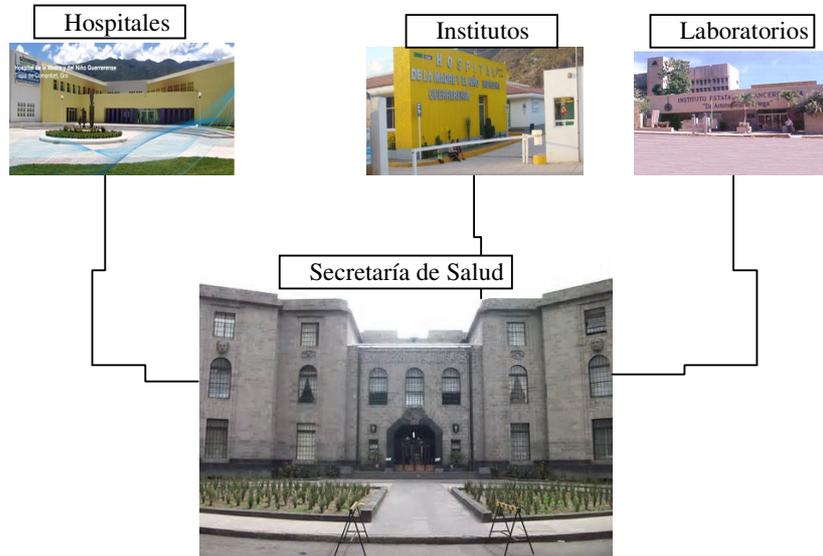


Fig. 2. Ejemplo de conexión de unidades médicas con el servidor central.

5 Conclusiones

Implementar Big Data en el Sector Salud del Estado de Guerrero, permitirá tener una mejora en los tratamientos y a los pacientes, apoyo en las decisiones médicas cada que se requiera el análisis de mayor número de variables de análisis comparativo de múltiples opciones de tratamiento y de esta manera proponer la mejor, se pueden monitorear pacientes a través de dispositivos remotos, recolectar datos individuales

Al utilizar esta tecnología se estará apoyando al sector sanitario estatal en la automatización de los sistemas de monitorización de pacientes, sistemas de soporte a las decisiones médicas, análisis comparativo de tratamientos y que los médicos puedan aplicar la medicina más personalidad de acuerdo a cada paciente.

Posteriormente se podrán incorporar más unidades médicas que continúen alimentando la base de datos con información de pacientes, enfermedades, sobre todo las denominadas del nuevo siglo, tratamiento. Dicha información podría ser compartida con investigadores externos a esta Secretaría o con farmacéuticas a través de algún convenio que se logrará firmar y a su vez estas organizaciones puedan retroalimentar a la dependencia con nuevos fármacos.

Referencias

1. Cognizant 20-20 insights. Big Data is the Future of Healthcare. With big data poised to change the healthcare ecosystem, organizations need to devote time and resources to understanding this phenomenon and realizing the envisioned benefits. (2012).
2. Gestal R., Vidal J., Carracedo J. C., Guillaumet J. Big Data. El inicio de una nueva era.
3. Institute for Health Technology Transformation (2013). Transforming Health Care through Big Data. Strategies for leveraging big data in the health care industry. (2013).
4. LOAPEG433, Ley Orgánica de la Administración Pública del Estado de Guerrero número 433. Artículo 27. Publicada en el Periódico Oficial No. 84 Alcance I, el día 21 de octubre de 2011.
5. Marz N. & Warren J. Big Data. Principles and best practices of scalable realtime data systems. (2012).
6. McKinsey Global Institute. Big Data: The next frontier for innovation, competition and productivity. (2011)
7. O'Reilly Media, Inc. Big Data Now. Current Perspectives from O'Reilly Media. (2012). 83-123.
8. TICbeat patrocinado por Informática & Actucast. Big Data. Como la avalancha de datos se ha convertido en un importante beneficio. (2012).
9. Vignesh Prajapati. Big Data Analytics with R and Hadoop. Set up an integrated infrastructure of R and Hadoop to turn your data analytics into Big Data analytics. (2013).