UNIVERSIDAD GABRIELA MISTRAL
FACULTAD DE INGENIERIA

BIG DATA Y SALUD, UN PARADIGMA APLICADO

Memoria para optar al título de Ingeniero de Ejecución en Informática

Autor : Freddy Antonio Arce Farfán
Profesor Guía : Roberto Carú Cisternas

Santiago – Chile
Diciembre, 2017

INDICE

DEDICATORIA

Este trabajo está dedicado, a mi señora Katya quien me brinda su amor, apoyo y comprensión y a mis hijos Agustín y Simón, que sin duda son mi motor, alegría, prosecución y esperanza de vida. Los Amo

AGRADECIMIENTOS

Quiero agradecer primeramente a Dios, por permitirme estar aquí, a mi Familia: mis Padres y Hermanos. Y expresarles a ellos que este es el resultado de una etapa que comenzó en mi infancia, adolescencia y ahora con alas propias me permiten dar este paso, que las enseñanzas y virtudes aprendidas por ustedes, serán replicadas en mis hijos, agradecido estoy queridos Padres (Juan y Herminda), son lo mejor.

I. INTRODUCCION

¿Prevenir enfermedades antes de que se diagnostiquen?

¿Recibir seguimiento de nuestras enfermedades en tiempo real sin necesidad de acudir a la consulta del médico?

¿Tener al alcance de la mano un mapa de nuestra salud para que sea analizado por especialistas?

Todas estas preguntas están cada día más cerca de convertirse en realidades cotidianas gracias, entre otros elementos, al Big Data.

Las tecnologías basadas en Big Data están revolucionando todos los aspectos de los negocios y también de la vida cotidiana. Una de las áreas en el que los cambios tecnológicos están teniendo un impacto mayor es en la medicina, donde la biotecnología, los wearables[1], la genómica[2], los robots cirujanos o la biónica son realidades hoy en día.

El desafío tecnológico no es pequeño, pero el Big Data nos habilita para manejar estos grandes volúmenes de datos y sacar partido de toda la información recopilada aplicando inteligencia a los datos. Bajo este contexto entonces la aplicación de Big Data es cada vez más evidente y necesaria en el mundo de la salud y la medicina.

Sin duda esta técnica de evaluación al paciente nos dirige hacia un híper personalización, la cantidad de datos médicos recopilados en el historial médico de una persona va a aumentar de forma exponencial y esto nos abrirá una nueva puerta de conocimiento, donde podremos tomar decisiones a medida para cada paciente.

Por otro lado, comienza una nueva era para la ciencia de datos, gracias al gran volumen de información disponible podremos aplicar técnicas de inteligencia artificial,

[1] Conjunto de aparatos y dispositivos electrónicos que se incorporan en alguna parte de nuestro cuerpo interactuando de forma continua con el usuario (relojes inteligentes, Smartphone, pulseras de ritmo, zapatillas con GPS, sensores, etc.).
[2] Conjunto de ciencias y técnicas dedicadas al estudio del campo de la biología molecular.

como machine learning[3], para realizar analítica avanzada y tomar decisiones en tiempo real, también en el campo de la medicina.

Es en este punto donde se centra este proyecto, cuyo objetivo principal por un lado es definir fundamentos y las principales técnicas de analítica avanzada, contextualizar las diferencias entre Business Intelligence y Big Data, analizar las posibilidades que ofrece la adopción de Big Data en el área de la salud, exponer las principales conclusiones que se hayan suscitado del proyecto y proyectar el desafío de esta tecnología a un centro hospitalario de la Región Metropolitana.

Consecuente con lo anterior, este proyecto es visionario a largo plazo y está dirigido hacia la mejora continua del lugar donde me desempeño laboralmente que es el Hospital Clínico de la Fuerza Aérea de Chile, proyecto que supone importantes avances, pero también grandes desafíos para el Hospital, el cual exige un considerable esfuerzo de adaptación por parte de la organización, cuyo recinto hospitalario atiende aproximadamente 40 mil pacientes al año (73% beneficiarios y 27% no beneficiarios) en las diversas áreas clínicas ambulatorias, urgencia, pabellón y hospitalización.

El Hospital Clínico de la Fuerza Aérea de Chile, es un establecimiento de alta complejidad, cuya misión será proporcionar atención medica preventiva y curativa al personal institucional, a sus cargas familiares y a otros autorizados que sean beneficiarios del Sistema de Salud Institucional de acuerdo a la ley. Asimismo, desarrollara capacidades para actuar como Hospital de Trauma en la guerra, participar en operaciones de evacuación Aero médicas estratégicas y tácticas y redesplegarse en apoyo a una emergencia institucional o nacional. Además, desarrollara actividades de docencia, investigación y capacitación para elevar el nivel de especialización de los integrantes de Hospital Clínico.

[3] Disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente.

1.1. Motivación

En la actualidad existen numerosas fuentes de datos heterogénea que arrojan una gran cantidad de información relacionada con los pacientes y las enfermedades. Esta información bien analizada, resulta de gran utilidad para los médicos.

Mi motivación radica precisamente ahí, en la técnica que presta el Big Data, la que permite inferir una capa de inteligencia, en la que resulta de especial relevancia la aplicación de modelos predictivos[4] que ayuden a anticiparse a las necesidades sanitarias y que ofrezcan una atención medica más eficaz.

1.2. Planteamiento del Problema

El Hospital Clínico en sus últimos años ha orientado sus esfuerzos para cumplir estándares de calidad asociados a la atención médica de sus usuarios, alcanzando un suficiente nivel de infraestructura del establecimiento y la incorporación de moderno equipamiento médico de última tecnología, lo cual contribuye de manera efectiva a optimizar el diagnóstico y tratamiento de patologías para los pacientes.

No obstante, para que éste pueda beneficiarse de las posibles ventajas que ofrece Big Data, es necesario previamente realizar una serie de cambios en la mayoría de los sistemas de información. Estos cambios deben estar orientados, entre otros objetivos, a conseguir gestionar y analizar grandes volúmenes de datos procedentes de fuentes muy diversas y registrados en formatos muy heterogéneos.

[4] Clave para poder mediante un esfuerzo analítico, detectar anticipadamente un comportamiento o resultados futuros.

Queda la sensación que los pacientes, cuyo propósito es acudir a la consulta por sentirse enfermos o proseguir una historia clínica, permaneciesen a la deriva en cuanto al tratamiento y/o futuros hallazgos complementarios a su salud, debido principalmente a que no existe la herramienta complementaria para realizar un seguimiento, y esto ocurre tanto por el abandono del tratamiento por parte del paciente o por la demanda de atender cada vez más.

Con esta tecnología se busca optimizar tanto la gestión clínica (para predecir cómo utilizar los recursos sanitarios de forma más eficiente: frecuencia de asistencia a consultas médicas, ingresos en el hospital, etc.) como el tratamiento y la atención al paciente (dando apoyo a la medicina personalizada), servicios de alertas, predicción de necesidades, generación de recomendaciones, descubrimiento de nuevos fármacos, diagnósticos, tratamientos, mejora de la atención al paciente y quizás una reducción del costo.

1.3. Hipótesis

La Hipótesis de esta tesis está fundada en que el Hospital Clínico de la Fuerza Aérea de Chile, carece de una tecnología de seguimiento o modelo predictivo para sus pacientes, de ahí que Big Data ofrece mayores oportunidades, tanto a pacientes como cuerpo médico.

No obstante, para que los usuarios que forman parte del entorno puedan tomar las mejores decisiones a partir de esta información, es necesario también disponer de una infraestructura de Business Intelligence (BI), que integre las herramientas adecuadas para facilitar la extracción y visualización de información de Big Data. En ese sentido, ya los principales proveedores de software de BI (Oracle, IBM, Microsoft, etc.), conscientes de las limitaciones que presentaban las soluciones tradicionales, han comenzado a incorporar este tipo de herramientas

en sus últimas versiones, lo que conlleva a la pavimentación y el impulso a remontar a esta tecnología que sin duda posicionara y tendrá la capacidad de enfrentar los nuevos desafíos del mercado de la salud, para seguir otorgando una mayor atención de calidad y optar a herramientas de diagnóstico de alta tecnología.

1.4. Objetivo General

El objetivo general de este estudio será analizar las posibilidades que ofrece Big Data al Área de Salud del Hospital Clínico de la Fuerza Aérea de Chile, siguiendo un enfoque metodológico basado principal y preferentemente en la revisión y estudio de recursos bibliográficos.

1.5. Objetivo Especifico

Los objetivos específicos de este trabajo son los siguientes:

- Describir el concepto de Big Data en una doble postura, tanto a las características de los datos masivos como a las herramientas tecnológicas que posibilitan su captura, almacenamiento y análisis.
- Definir la analítica predictiva, tanto de su técnica, modelos y métodos de analítica, como su fase de construcción de un proyecto. Herramientas de Software para el análisis predictivo.
- Estrategia de integración de Big Data en un Sistema de Información para trabajar de forma complementaria con una solución Business Intelligence.
- La posibilidad que ofrece Big Data al Área de la salud del Hospital clínico, reparando en la práctica clínica y en la investigación biomédica, considerando los retos y barreras que se deben afrontar y superar, particularmente en el sector clínico.

Observando que Big Data está consiguiendo ya logros interesantes en el campo de la investigación biomédica.

2. MARCO TEORICO

En el siguiente apartado se definirá el concepto de Big Data bajo un doble significado:

- El de la descripción de los principales rasgos que distinguen los denominados “datos masivos”.
- Y el de las herramientas tecnológicas que permiten manejar esos datos de modo que sirvan para adquirir conocimiento y la toma de decisión.

2.1. ¿Qué es el Big Data?

Durante los últimos años se ha intentado explicar la complejidad que encierra el termino Big Data o datos masivos, aludiendo tanto a las características de los datos como a las herramientas de las tecnologías de la información para manejarlos en provecho de las organizaciones. Los autores Mayer-Schönberger y Neil Cukier[5] suscitan en su libro, “el volumen de información había aumentado tanto que la que se examinaba ya no cabía en la memoria que los ordenadores emplean para procesarla, por lo que los ingenieros necesitaban modernizar las herramientas para poder analizarla”. Esta sería la razón principal para impulsar el desarrollo de nuevas tecnologías que lo hicieran posible.

[5] Autores del Libro Big Data: la revolución de los datos masivos.

A pesar de la importancia y popularidad que ha alcanzado el fenómeno de Big Data y la abundante bibliografía que se ha originado en torno al tema, cita Joyanes Aguilar[6], “no existe una unanimidad en la definición de Big Data, aunque si un cierto consenso en la fuerza descriptiva que suponen los grandes volúmenes de datos y la necesidad de su captura, almacenamiento y análisis” En su estudio selecciona algunas de las definiciones más significativa propuestas por distintas instituciones relevantes en este ámbito, Entre ellas figura la consultora tecnológica IDC[7], que dice:

> “Big Data es una nueva generación de tecnologías, arquitecturas y estrategias diseñadas para capturar y analizar grandes volúmenes de datos provenientes de múltiples fuentes heterogéneas a una alta velocidad con el objeto de extraer valor económico de ellos”:

Otra de las seleccionadas por Joyanes es la propone la consultora Gartner:

> “Big Data son los grandes conjuntos de datos que tienen tres características principales: **Volumen** (cantidad), **velocidad** (velocidad de creación y utilización) y **Variedad** (tipos de fuentes de datos no estructurados, como la interacción social, videos, audio, toda cosa que se pueda clasificar en una Base de datos”

Así, el interés cada vez más generalizado por desentrañar lo que significa Big Data y la importancia que pueda tener su adopción para las organizaciones no cesa. Como lo evidencian los numerosos artículos, informes y monografías publicadas sobre el tema, especialmente a partir de un par de años atrás.

[6] Autor del Libro, Big Data: Análisis de grandes volúmenes de datos en las organizaciones.
[7] International Data Corporation (**IDC**) es la principal firma mundial de inteligencia de mercado, servicios de consultoría, y conferencias para los mercados de Tecnologías de la Información, Telecomunicaciones y Tecnología de Consumo.

2.2. Características del Big Data

Como ya se ha anticipado, las características principales que define los datos masivos son las llamadas originalmente las “tres V”, las dimensiones **volumen**, **velocidad** y **variedad**. Sin embargo, con posterioridad se han añadido algunas más, como **veracidad** y **valor**.

Según las previsiones de Gartner, en 2020 más de 25 mil millones de dispositivos estarán conectados a Internet, acrecentando un volumen de datos que a finales de 2013 ya se estimaba en 4,4 billones de GB y que llegará, según los pronósticos, a multiplicarse por 10 en tan solo 6 años. Por supuesto, el impacto que tendrá este crecimiento exponencial del volumen de datos contenidos en Big Data sobre la inversión en TIC[8] dará mucho que hablar en un futuro muy cercano.

Además de ofrecernos una idea aproximada de las incontables aplicaciones que pueda tener su correcta gestión, el volumen no es la única magnitud que define Big Data; no debemos olvidar que los datos contenidos en él no poseen una única característica en común que permita homogeneizarlos y tratarlos por igual, de una sola vez.

Efectivamente, el valor de Big Data no se agota con la mera consideración de su volumen: también la variedad de los datos que contiene es una de sus grandes riquezas. Pero además de abrir las puertas a nuevas oportunidades de negocio, ambas magnitudes también plantean uno de los grandes retos que actualmente afrontan las TIC: gestionar un volumen y una variedad inmensa de datos facilitando el acceso inmediato a los mismos.

[8] *Tecnologías de Información y Comunicación.*

Así es, la velocidad en el acceso y el flujo de datos es el gran reto que plantea Big Data no solo hoy, sino también y sobre todo de cara al futuro. El indudable valor que aporta a los proyectos de Business Intelligence está determinado por sus tres “V”, el manejo de las cuales pone encima de la mesa cuestiones de tanta importancia como, por ejemplo, si las herramientas de almacenamiento tradicionales deben quedar a un lado para dar paso a sistemas de gestión que no pretendan capturar datos indiscriminadamente para su posterior tratamiento y estructuración, sino que realicen esta tarea en tiempo real captando únicamente aquellos datos que aporten valor a los sistemas BI.

Indudablemente, la integración de herramientas de almacenamiento en la nube ha revolucionado el mundo del Business Intelligence, permitiendo acceder con gran velocidad a grandes volúmenes de datos desde distintos dispositivos simultáneamente, preservando su variedad y proporcionando información útil para la toma de decisiones. Sin embargo, teniendo en cuenta el futuro que ya se avizora en el horizonte más cercano y todavía sin haber explotado todas las posibilidades que ofrecen los cloud, puede que incluso se deba repensar el modo de operar a través de ellos.

Así como se comentaba anteriormente de la inclusión posterior de nuevos términos, encontramos que la Veracidad, característica, añadida por IBM, se refiere a la incertidumbre que comporta la presencia de ciertos tipos de datos que se encuentran en el conjunto extraído. Por mucho esfuerzo que se dedique a la limpieza de los mismos para obtener una mejor calidad, en algunos de ellos no se puede suprimir ni la imprevisibilidad (del tiempo, de la economía) ni la incertidumbre (de los sentimientos, de la sinceridad de las personas o de sus reacciones ante determinados hechos). No obstante, esta dimensión de los datos masivos debe ser tenida en cuenta. Los directivos necesitan abordar esta incertidumbre y determinar cómo planificarla para el beneficio de su organización.

En cambio, el Valor en Big Data podría entenderse conceptualmente como una combinación de las dimensiones anteriormente mencionadas. Cada empresa u organización podrá adoptar esta tecnología bajo diferentes enfoques, pero con un objetivo común: mejorar el rendimiento y la toma de decisiones. En definitiva, se trataría de obtener valor mediante el uso estratégico de la información que pueden proporcionar el proceso y análisis de los grandes conjuntos de datos que se han ido almacenando en su entorno.

Mayer y Cukier, en su documentado estudio del 2013, dedican todo un capítulo a poner en evidencia el valor de los datos masivos. Consideran básicamente que dicho valor no ha de verse sólo a la luz del que tienen en su uso primario, y que justifica su almacenamiento, sino también teniendo en cuenta las posibilidades que ha abierto la tecnología para su reutilización, pudiéndose procesar una y otra vez con propósitos múltiples. Dicho de otro modo, “los datos pasan de unos usos primarios a otros secundarios, lo que los vuelve mucho más valiosos a lo largo del tiempo. Proponen tres vías para desencadenar el llamado “valor de opción” de los datos: la reutilización básica, la fusión de conjuntos de datos y el hallazgo de combinaciones “.

Sin dudad entonces, la adaptabilidad a los nuevos contextos que proliferen con la evolución de estas magnitudes será la clave del éxito de los sistemas BI de un mañana que está a la vuelta de la esquina. De qué modo se materializarán estas adaptaciones y qué novedades traerán consigo son interrogantes que no tardaremos en poder responder.

2.3. Tipos de Datos

En función del manejo de los datos masivos, se suelen distinguir tres tipos de datos:

- Datos estructurados.
- Datos semiestructurados.
- Datos no estructurados.

Los **datos estructurados** son los que tienen un esquema definido, en formato y longitud, para poder ser incluidos en un campo fijo (fechas, números, cadenas de caracteres, etc.) y almacenados en tablas, por ejemplo: las de una hoja de cálculo o una base de datos relacional.

Los **datos semiestructurados** carecen de formato fijo o de campo determinado, pero están dotados de marcadores que permiten diferenciar los distintos elementos dato. Un ejemplo: las etiquetas de lenguaje HTML y XML. En el ámbito de la salud, ejemplo de estas modalidades serían los datos de contabilidad, facturación electrónica, algunos datos de actuario o dato clínico.

Los **datos no estructurados** no tienen un formato especifico ni se pueden asignar a un campo fijo, por lo que no es posible su almacenamiento en una tabla. Se tratan como documentos u objetos. Ejemplo de este tipo de datos son: documentos de audio, video, fotografías, e-mails o archivos PDF. En el área de la salud, cabe señalar las imágenes de radiografías, resonancias magnéticas recetas en papel, etc.

2.4. Fuentes de datos masivos

Atendiendo a los datos que las empresas deben analizar, según diferentes propósitos y con arreglo a las fuentes donde se originan, se ha establecido una clasificación, en la cual que se ha convertido en una referencia, donde se distinguen cinco categorías básicas de fuentes de datos, comprendiendo cada una de ella distintos tipos de información. La que se presenta a continuación se basa en la versión publicada por Barranco Fragosa[9], a cuyos ejemplos citados se incluyen algunos referidos al área de la salud.

- ***Web and Social Media***

  Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, blogs y diversos contenidos Web. Para el área de Salud, Big Data puede recoger información cada vez más abundante, como las redes sociales temáticas para profesionales médicos o para comunidades virtuales de pacientes[10].

- ***Machine to machine (M2M)***

  Se refiere a las tecnologías que permiten conectarse a otros dispositivos, como sensores o medidores que capturan un evento en particular (humedad, velocidad, temperatura, presión, etc.). Entre los datos procedentes de estos dispositivos se encuentran: lectura de medidores inteligentes, de RFID (Radio Frecuency Identification)[11], de sensores de plataformas petrolera, señales GPS. En el área de la salud, los sistemas que recogen los datos procedentes de sensores en dispositivos wereables o de smartphones en pacientes monitorizados en tele asistencia.

[9] Publicación mencionada en la Bibliografía.
[10] Publicación mencionada en la Bibliografía.
[11] Es una técnica de identificación por medio del empleo de ondas de radio.

- ***Big Transaction Data***

  Incluye registros de facturación, de telecomunicaciones y registros detallados de llamadas. Los datos transaccionales pueden ser semiestructurados y no estructurados.

- ***Biometrics***

  Hace referencia a datos de información biométrica, como huellas digitales, reconocimiento facial, escaneo de retina, genética (ADN), etc. Son importantes en el área de seguridad e inteligencia para las agencias de investigación.

- ***Human Generated***

  En este apartado se incluyen datos generados por personas, como los que guardan en un call center al establecer una llamada telefónica, las notas de voz, correos electrónicos, documentos, estudios y registros médicos electrónicos o recetas médicas.

2.5. Plataformas de Big Data

Como ya se ha anticipado, el uso de datos masivos requiere la utilización de nuevas herramientas tecnológicas para su captura desde las diferentes fuentes y sistemas, así como su transformación, almacenamiento, análisis, visualización, etc. Desde luego, el ángulo correcto que actualmente tiene el liderazgo en términos de popularidad para analizar enormes cantidades de información es la plataforma de código abierto ***Hadoop****. Ha sido adoptado tanto por la comunidad de desarrolladores de aplicaciones de software libre, como los principales proveedores de Software propietarios de bases de datos (Oracle, IBM, Microsoft).*

El proyecto Hadoop consta de tres componentes fundamentales: Hadoop Distributed File System (HDFS), Hadoop MapReduce., Hadoop Common.

- ***Hadoop Distributed File System (HDFS)***

  Los datos en el clúster de Hadoop son divididos en pequeñas piezas llamadas *bloques* y distribuidas a través del clúster; de esta manera, las funciones map y reduce pueden ser ejecutadas en pequeños subconjuntos y esto provee de la escalabilidad necesaria para el procesamiento de grandes volúmenes.

- ***Hadoop MapReduce***

  Se considera como el núcleo de Hadoop. El término MapReduce en realidad se refiere a dos procesos separados que ejecuta Hadoop. El primer proceso es map, el cual toma un conjunto de datos y lo convierte en otro conjunto, donde los elementos individuales son separados en tuplas (pares de llave/valor). El segundo proceso es reduce obtiene la salida de map como datos de entrada combinando las tuplas en un conjunto más pequeño de las mismas. Una fase intermedia es la denominada Shuffle la cual obtiene las tuplas del proceso map y determina que nodo procesará estos datos dirigiendo la salida a una tarea reduce en específico. A continuación, un ejemplo de MapReduce.

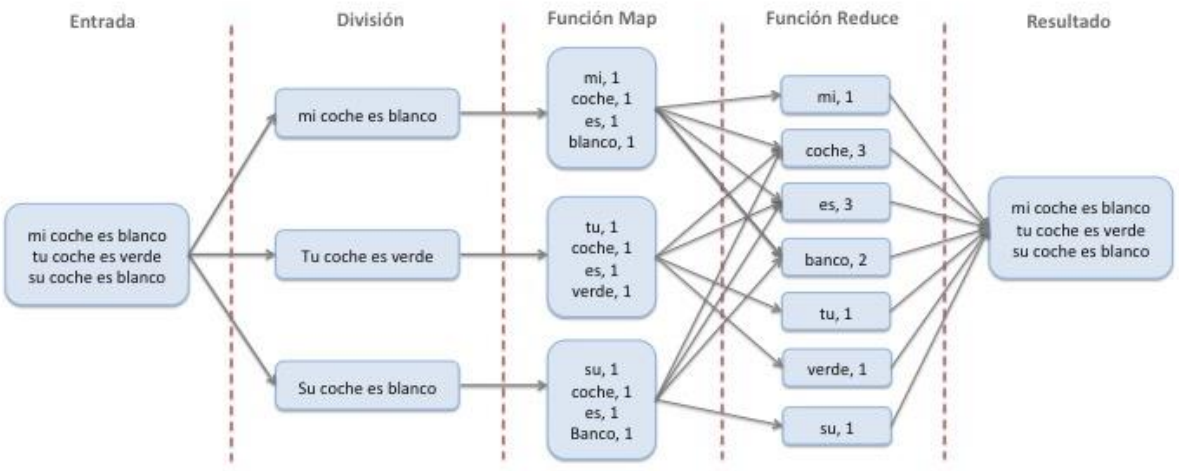


Ilustración Hadoop MapReduce

- ***Hadoop Common***

  son un conjunto de librerías que soportan varios sub-proyectos de Hadoop.

  Además de estos tres componentes principales de Hadoop, existen otros proyectos relacionados los cuales son definidos a continuación:

  - ***Avro***

    Es un proyecto de Apache que provee servicios de serialización. Cuando se guardan datos en un archivo, el esquema que define ese archivo es guardado dentro del mismo; de este modo es más sencillo para cualquier aplicación leerlo posteriormente puesto que el esquema está definido dentro del archivo.

  - ***Chukwa***

    Es un sistema de colecciones de datos para la gestión de grandes sistemas distribuidos.

- ***Cassandra***

  Es una base de datos no relacional (NoSQL) distribuida y basada en un modelo de almacenamiento de (clave-valor), desarrollada en Java. Permite grandes volúmenes de datos en forma distribuida. Su arquitectura está diseñada para ser altamente escalable. Twitter es una de las empresas que utiliza Cassandra dentro de su plataforma.

- ***Flume***

  La tarea principal es dirigir los datos de una fuente hacia alguna otra localidad, en este caso hacia el ambiente de Hadoop. Existen tres entidades principales: sources, decorators y sinks. Un **source** es básicamente cualquier fuente de datos, **sink** es el destino de una operación en específico y un **decorator** es una operación dentro del flujo de datos que transforma esa información de alguna manera, como por ejemplo comprimir o descomprimir los datos o alguna otra operación en particular sobre los mismos.

- ***HBase***

  Es una base de datos NoSQL columnar (column-oriented data base) que se ejecuta en HDFS. Cada tabla contiene filas y columnas como una base de datos relacional. HBase permite que muchos atributos sean agrupados en las denominadas familias de columnas, de modo que los elementos de una misma familia de columnas sean almacenados en un solo conjunto. Eso es lo que las distingue de las bases de datos relacionales orientadas a filas, donde todas las columnas de una fila dada son almacenadas en conjunto. Facebook utiliza HBase en su plataforma desde noviembre del 2010.

- ***Hive***

  Es una infraestructura de data Warehouse que facilita administrar grandes conjuntos de datos que se encuentran almacenados en un ambiente distribuido. Hive tiene definido un lenguaje similar a SQL llamado Hive Query Language (HQL), estas sentencias HQL son separadas por un servicio de Hive y son enviadas a procesos MapReduce ejecutados en el cluster de Hadoop.

- ***Jaql***

  Donado por IBM a la comunidad de software libre. Es un lenguaje de consultas, funcional y declarativo, que permite la explotación de datos en formato JSON (Javascript Object Notation), diseñado para procesar grandes volúmenes de información. Jaql posee una infraestructura flexible para administrar y analizar datos semiestructurados como XML, archivos CSV, archivos planos, datos relacionales, etc.

- ***Lucene***

  Es un proyecto de Apache bastante extendido para realizar búsquedas sobre textos. Proporciona librerías para indexación y búsqueda de texto. Los documentos (document) se dividen en campos de texto (fields) y posteriormente se genera un índice sobre estos últimos. La indexación es el componente fundamental de Lucene, ya que le permite realizar búsquedas rápidamente con independientemente del formato del archivo, ya sean PDFs, documentos HTML, etc.

- ***Oozie***

  Es un proyecto de código abierto que simplifica los flujos de trabajo y la coordinación entre cada uno de los procesos que deben ser ejecutados en distintos momentos. Permite que el usuario pueda definir acciones y las dependencias entre dichas acciones.

- ***Pig***

  Es un lenguaje de programación que simplifica las tareas comunes de Hadoop, como son la carga de datos, la expresión de las transformaciones sobre los datos y el almacenamiento de los resultados finales.

- ***ZooKeeper***

  Proporciona una infraestructura centralizada de servicios que pueden ser utilizados por aplicaciones para asegurarse de que los procesos a través de un clúster sean serializados o sincronizados.

La lista que se ha presentado no pretende ser íntegra, ya que el interés por la tecnología Big Data sigue aumentando y no dejan de surgir continuamente proyectos nuevos o evolucionando los ya existentes y conocidos.

2.6. ¿A qué se debe el auge actual del Big Data?

En los últimos años han sido numerosos los autores que han tratado de explicar la enorme popularidad que ha experimentado la utilización del Big Data por parte de las organizaciones en comparación con el uso de las tecnologías precedentes.

Así como lo señala, José Ramón Rodríguez, en sus artículos “Octubre, el mes de los grandes datos”[12]: ¿Qué hay de nuevo con los Big Data comparado con lo que siempre hemos llamado Inteligencia de Negocio y más recientemente Inteligencia Analítica?

[12] Mencionado en Bibliografía.

La respuesta ofrecida es que la diferencia principal de esta tecnología con respecto a las anteriores radica en “la multiplicación de la velocidad, el volumen, la variedad, la tipología de datos, la división del costo de su producción, tratamiento y almacenamiento”

Por otro lado, IBM, en su informe ejecutivo, señala dos tendencias que explican en buena medida las diferencias entre las actividades actuales de Big Data y las anteriores:

- La digitalización de prácticamente todo, que ha generado en diferentes sectores nuevos tipos de grandes datos en tiempo real, muchos de ellos no normalizados (datos en streaming, geoespaciales o generados por sensores) que no pueden ser correctamente procesados por los warehouses relacionales, tradicionales y estructurados”.

- Las tecnologías y técnicas de análisis avanzado de hoy en día, que hacen posible que las organizaciones extraigan conocimientos de los datos con un nivel de sofisticación, velocidad y precisión nunca antes visto.

Más recientemente, Luis Joyanes, en su libro Big Data en el apartado ¿Cómo se ha llegado a la explosión de Big Data?, explica este fenómeno a partir de la diferencia entre los datos estructurados que se utilizaban antes y el enorme volumen de datos de carácter no estructurado manejado en la actualidad, proveniente de todo tipo de fuentes (redes sociales, dispositivos móviles, Internet de las cosas, etc.).

3. INTEGRACION DE BIG DATA EN LA ORGANIZACIÓN

En este capítulo se ha querido señalar la relación de la integración de Big Data y el Business Intelligence en las organizaciones a nivel global, tales como las fases de adopción empresarial y/o las estrategias de integración de Big Data dentro de una solución BI.

3.1. Adopción empresarial de Big Data

Diversos estudios de los últimos años dejan ostensible la progresiva adopción de Big Data por parte de las organizaciones y empresas. La información publicada en los mismos permite conocer el proceso de implantación y afianzamiento de Big Data, desde las fases iniciales hasta el comienzo de una etapa de consolidación. Por otro lado, esta información también hace referencia a las barreras que han obstaculizado su adopción empresarial. A continuación, se enumeran algunas de las principales conclusiones recogidas en diferentes informes recopilados en la red.

En el informe de IBM[13], citado en la Bibliografía de este trabajo, los autores sugieren cuatro fases principales en el proceso de adopción y evolución de Big Data, que denominan **educar**, **explorar**, **interactuar** y **ejecutar**, las cuales se resumen a continuación:

- En la fase **educar** se trata de crear una base de conocimiento y observaciones del mercado. Las empresas que se encuentran en esta primera etapa estudian las posibles ventajas de las tecnologías y la analítica de Big Data, e intentan entender cómo estas pueden ayudarles a abordar oportunidades de negocio en sus correspondientes sectores y mercados.

[13] Analytics: El uso de Big Data en el mundo real.

- En la fase **explorar** se debe desarrollar una estrategia y una hoja de ruta basándose en las necesidades de negocio y los retos empresariales. Las organizaciones incluyen entre sus principales objetivos desarrollar un caso de negocio y crear un proyecto de Big Data, donde tendrán en cuenta los datos, la tecnología y las habilidades existentes, para luego establecer cuándo y dónde comenzar y cómo desarrollar el plan conforme a la estrategia de negocio de la empresa.

- En la fase **interactuar** las empresas comienzan a comprobar el valor de negocio que proporciona Big Data, así como sus tecnologías y habilidades. En esta etapa las organizaciones desarrollan pruebas de concepto o proyectos piloto que les permiten interactuar con la nueva tecnología, dentro de un ámbito definido y limitado, y de esta forma evaluar si cumple con los requisitos y los resultados esperados.

- La última fase, **ejecutar**, tiene como objetivo implementar Big Data a escala. El nivel de operatividad e implementación de las funciones analíticas es mayor dentro de la empresa, habiendo desplegado dos o más iniciativas de Big Data.

Quizás una de las mayores incertidumbres a la hora de operar con esta tecnología es la relacionada con la falta de profesionales con experiencia, este punto constituye uno de los principales obstáculos para la adopción de Big Data, entre las recomendaciones que se indican a las empresas en dicho informe se encuentra la mejora de los conocimientos de sus empleados mediante programas de formación y desarrollo.

Por último, el reciente estudio de septiembre de 2015, publicado por la consultora Gartner[14], revela que la inversión en Big Data por las empresas a nivel global, sin mencionar explícitamente el ámbito sanitario, ha superado ya la fase inicial de rápido crecimiento de años anteriores, para entrar en una etapa de consolidación con un crecimiento bastante más lento. Se espera que en los próximos años el 75% de las empresas invierta en Big Data, lo que supone solo un 3% de incremento respecto a 2014.

Asimismo, este informe señala un cierto cambio hacia la especialización en el análisis de datos. Un 70% de las compañías que han invertido en Big Data están analizando o planificando analizar datos de localización, y un 64% lo están haciendo para analizar texto en formato libre.

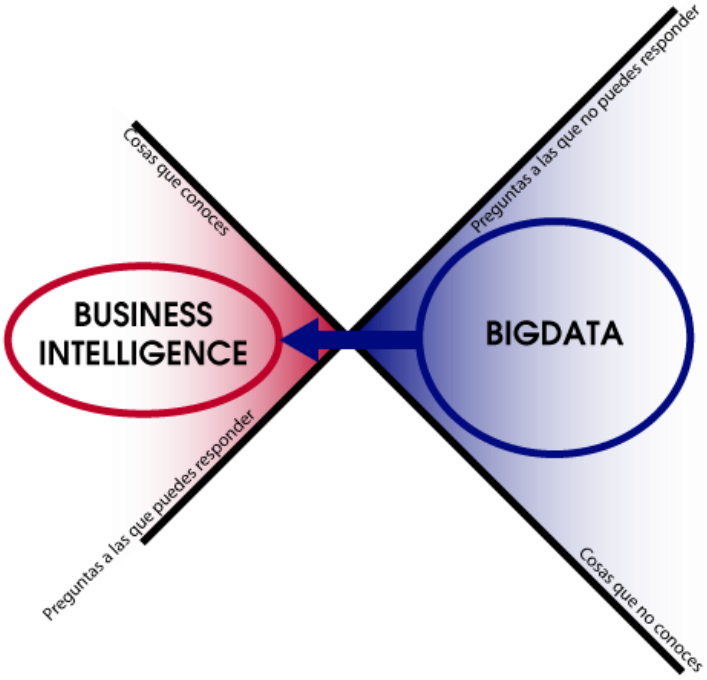
3.2. Business Intelligence y Big Data, ¿Tecnologías complementarias o sustitutivas?

La creciente implantación de Big Data en las diferentes organizaciones, sin dejar de lado el área de la salud ha tenido, durante los últimos años una suscitación un debate que continúa abierto acerca de la posible coexistencia de esta tecnología junto con Business Intelligence (BI) dentro de una misma organización. Una de las razones que explica esta controversia se encuentra en que ambas soluciones tratan de ofrecer a los usuarios de la organización, y especialmente a los profesionales que ocupan puestos directivos, herramientas que ayuden a optimizar el proceso de toma de decisiones.

Por otro lado, mientras que “BI” se ha centrado desde sus orígenes en tratar de estructurar información útil y relevante para la toma de decisiones a partir de datos consolidados (información conocida), “Big Data” es capaz de facilitar el análisis de una cantidad mucho mayor de datos de todos los tipos, y especialmente

[14] Citado en la Bibliografía Como: Survey Analysis: Practical Challenges Mount as Big Data Moves to Mainstream

no estructurados (información desconocida), los cuales han ido adquiriendo un valor estratégico cada vez mayor para las organizaciones.



Diferencia entre Big Data y Business Intelligence

La correlación en cuanto a los objetivos y los posibles avances que presenta Big Data con respecto a Business Intelligence han dado lugar a que algunos autores afirmen que esta última tecnología se verá con el tiempo desplazada por la primera, considerando a Big Data como una evolución de Business Intelligence. Sin embargo, a pesar de los argumentos que respaldan la visión anterior, son muchos los autores que defienden la posibilidad de que ambas tecnologías coexistan dentro de un mismo sistema de información, y señalan los beneficios que se pueden obtener a partir de la complementariedad de ambas soluciones: mientras que Big Data puede aportar a Business Intelligence información procedente de datos no estructurados, Business Intelligence es capaz

de realizar un análisis avanzado de esta información y de ofrecer al usuario una solución visualmente atractiva.

No obstante, hay que señalar que cada organización posee características y necesidades particulares, las cuales se deben identificar con precisión antes de tomar la decisión de implantar cualquiera de estas dos tecnologías (o una combinación de ambas). No hay que descartar, por lo tanto, que existan situaciones en las que resulta desaconsejable la adopción conjunta de ambas soluciones, si se considera que los beneficios que pueda reportar su implantación son inferiores a los costos que esta supone.

En aquellos casos en los que sí puede ser adecuada la implantación de ambas tecnologías, es necesario realizar una correcta integración de la solución planteada dentro del sistema de información de la organización. A continuación, se describen, de forma general, los aspectos fundamentales que se deberían considerar para llevar a cabo con éxito esta integración.

Integrar Business Intelligence y Big Data permite un análisis potente de una gran cantidad de datos que brinda una ventaja competitiva clave: la de transformar cualquier tipo de datos, (volumen, forma, ubicación) en informaciones de alto valor agregado para cada sector de la empresa.

Sin embargo, hay que saber explotar esta materia prima para que se transforme en el mayor capital de una organización: el conocimiento pertinente del negocio.

La alianza Business Intelligence y Big Data permite entre otras cosas, medir el sentir de los pacientes, optimizar las cadenas de suministro y detectar el engaño.

Hoy, existe una fuerte tendencia por parte de las organizaciones a complementar el uso del Business Intelligence con la funcionalidad del Big Data. Esta unión tiene como fin facilitar al cuerpo médico en la toma de decisiones ante el incremento constante de datos a analizar.

Para señalar como se complementarían, debemos señalar primero que el Business Intelligence es un sistema que analiza los datos estructurados de la organización con el fin de ayudar a tomar mejores decisiones con un único objetivo: potenciar los resultados del negocio.

En cambio, el Big Data es un conjunto de tecnología, arquitectura, y procesos que permite captar, tratar y analizar rápidamente grandes volúmenes de datos y cantidades de contenidos heterogéneos, que están en constante evolución.

Big Data agrega a su solución BI el imprescindible poder de extraer las informaciones pertinentes, de manera sencilla y accesible para todos.

3.3. Madures de las empresas con respecto a Business Intelligence

El primer paso para la integración de una solución que incorpore BI y Big Data dentro del sistema de información de una organización consiste en definir una correcta estrategia de inteligencia de negocio, midiendo a la vez el nivel de madurez de la organización, en una de las siguientes fases:

- **Fase 1:** No existe BI.

  Los datos se encuentran en los sistemas de procesamiento de transacciones en línea, repartidos en otros soportes o en algunos

casos únicamente en el know-how[15] de la organización. Las decisiones no están basadas en datos consistentes, sino en la intuición o la experiencia. No se percibe la importancia del uso de datos corporativos a través de las herramientas adecuadas en la toma de decisiones.

- **Fase 2:** No existe BI

  Aunque los datos se encuentran accesibles. Para la toma de decisiones no se realiza un procesado formal de los datos, aunque algunos usuarios pueden acceder a información de calidad y justificar decisiones con dicha información. A menudo este proceso se lleva a cabo mediante Excel o algún tipo de reporte manual. Aunque se tiene conciencia de las limitaciones de este proceso se desconocen las posibilidades de BI.

- **Fase 3:**

  Se llevan a cabo procesos formales de toma de decisiones basada en datos. Existe un equipo que controla los datos, a partir de los cuales es posible elaborar informes y tomar decisiones fundamentadas. Los datos se extraen directamente de los sistemas transaccionales sin realizar limpieza de datos ni modelización. No existe un almacén de datos, data mart o data Warehouse.

- **Fase 4:**

  Se dispone de data Warehouse, aunque los reportes se realizan de forma personal.

[15] Capacidades y habilidades que un individuo u organización poseen en cuanto a la realización de una tarea específica.

- **Fase 5:**

  Se amplían las funcionalidades de la data Warehouse y se formaliza el reporte a nivel corporativo. Se plantea la adopción de OLAP[16].

- **Fase 6:**

  Ni el reporte ni el acceso a la data Warehouse permiten responder de forma eficiente a preguntas complejas, por lo que se decide la implantación de OLAP. Las decisiones tienen un peso cada vez mayor en los procesos de negocio de toda la organización.

- **Fase 7:**

  Se formaliza el uso de Business Intelligence. Se hace necesaria la implantación de otros procesos de inteligencia de negocio como Data Mining[17] o Balanced ScoreCard[18].

3.4. Estrategias de integración de Big Data dentro de una solución de Business Intelligence empresarial

Una vez identificada la fase en la que se encuentra la organización, el siguiente paso consistirá en plantear el diseño de una plataforma que integre Business Intelligence y la analítica de Big Data. En el artículo de Peter J. Jamack[19], se abordan desde una perspectiva general los aspectos más importantes que intervienen en este diseño.

[16] Procesamiento analítico en línea
[17] Conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática.
[18] Visualiza un balance integrado y estratégico del avance, crecimiento, productividad y competitividad de una organización.
[19] Citado en la Bibliografía: Analítica de inteligencia de negocios de Big Data.

Según este autor, una de las cuestiones principales que deben considerarse en el diseño de una plataforma integrada consiste en la extracción, la transferencia y la carga (ETL). Este aspecto tiene una importancia decisiva en proyectos de integración, ya que, de no realizarse correctamente, se corre el riesgo de utilizar datos incorrectos y poco fiables, que afectarán a la calidad y al uso del sistema.

Para cada uno de estos aspectos el autor propone diferentes soluciones:

- En el caso de los procesos ETL[20], existen diferentes herramientas y metodologías, como Sqoop[21], que permite procesar datos de sistemas de gestión de base de datos relacionales, u otras aplicaciones open source, como Flume o Scribe, que pueden ser útiles en el procesamiento de los registros. También se dispone de herramientas más visuales y que integran Big Data, como IBM InfoSphere DataStage. La selección de una u otra solución deberá estar en función de los sistemas, las fuentes, los datos, el tamaño y el personal de la organización.

- Para el almacenamiento de datos existen diferentes alternativas, como HBase, dentro del sistema de Hadoop, o bien Cassandra, Neo4j, Netezza o el sistema de almacenamiento de archivos HDFS. Jamack recomienda para la integración de Big Data el uso de una plataforma integrada que consista en Hadoop y Cassandra para datos no estructurados o semiestructurados e IBM Netezza Customer Intelligence Appliance, que utiliza en la capa del usuario el software de BI IBM Cognos, para datos estructurados.

[20] Extracción, transferencia y carga
[21] Aplicación con interfaz de línea de comando para transferir datos entre bases de datos relacionales y Hadoop.

- Los usuarios avanzados pueden utilizar herramientas como IBM SPSS Statistics, que realizan tareas de minería de datos, modelado predictivo, aprendizaje automático, desarrollo de algoritmos y modelos complejos. Otras, como Cognos, permiten a distintos tipos de usuarios explorar los datos o visualizar informes simples. Jamack también sugiere Apache Mahout, para tareas de aprendizaje automático, o Apache Hive, que permite realizar consultas de tipo SQL (Structured Query Language) sobre Hadoop. En cualquier caso, el autor recomienda no perder de vista el interés del usuario final, que consiste básicamente en disponer de los datos correctos en el momento adecuado.

Además de los ya citados, Jamack advierte de la importancia de otros aspectos que se deben considerar en el diseño de una plataforma integrada, como el traslado de los datos. En función del tamaño y la distribución geográfica de la compañía, las operaciones de traslado de datos de un lugar a otro pueden tener una gran complejidad, como en el caso de organizaciones que cuentan con sedes ubicadas en varios países, en donde los datos, que pueden ser no estructurados o semiestructurados, pueden estar repartidos en distintas bases de datos en múltiples data centers a nivel global, y protegidos con diferentes mecanismos de seguridad. Las tecnologías de Big Data como Apache Hadoop tienen en cuenta esta complejidad, además del costo temporal que supone el traslado de los datos y los posibles riesgos, como la pérdida de datos, paquetes o archivos. Por ello fomentan el traslado del sistema hasta donde se encuentran los datos en lugar de llevar los datos hacia el sistema. No obstante, el traslado de la aplicación de Big Data hacia los datos también puede ser una tarea muy compleja, sobre todo cuando se cuenta con diferentes sistemas de gestión de datos maestros (Master Data Management o MDM) para cada una de las áreas de la organización (productos, ventas, clientes), y existen dificultades para integrar los distintos MDM.

Por otro lado, una aplicación de estas características requiere una gran cantidad de espacio, memoria y velocidad, que puede provocar fallos o no ser operativa si el sistema sobre el que se instala es antiguo o no está suficientemente preparado. Para abordar esta problemática, Jamack recomienda en primer lugar analizar los sistemas existentes, los casos de uso y el grado de experiencia y competencia del personal de la organización. Una vez realizado este análisis, existen diferentes opciones, como el desarrollo de un sistema completo de código abierto partiendo de tecnologías como Vanilla Hadoop (Sistema de Archivos Distribuidos de Hadoop [HDFS] y MapReduce), Zookeeper, Solr, Sqoop, Hive, HBase, Nagios y Cacti. Otra posible opción sería el desarrollo de un sistema que incorporara tecnologías que proporcionan un soporte mayor, como IBM InfoSphere Big Insights e IBM Netezza. Una tercera alternativa consistiría en la separación de datos estructurados y no estructurados y el desarrollo de una capa de interfaz gráfica de usuario (GUI) para usuarios finales, usuarios avanzados y aplicaciones.

Por último, el autor aconseja tener en cuenta la experiencia del personal de la organización en el trabajo con Big Data para llevar a cabo la integración de Business Intelligence y la analítica de Big Data. En caso de que la organización cuente con personal experto, esta puede optar por un sistema de código abierto, ya que resulta mucho más rápido y menos costoso de implementar. En caso contrario, Jamack aconseja una aplicación de proveedor de Big Data, a pesar del mayor costo que esta supone.

4. DESARROLLO Y PROYECCION

En base a la metodología de trabajo de este proyecto de Titulación, el cual consiste en el análisis de las bondades que ofrece la tecnología del Big Data para ser implementada en el Hospital Clínico de la Fuerza Aérea de Chile y como ya se ha señalado en los puntos anteriores se estudiaron y examinaron conceptos, planteamientos, por que usar esta tecnología en el área de la salud, las implicancias que trae, sociabilización del Big Data en el mundo empresarial, modelo de solución propuesto.

Por otro lado, este capítulo también estará dedicado y orientado íntegramente a conceptuar el Big Data en el área de la salud, partiendo de dos estudios fundamentales, en el primero se analizan las posibilidades de la analítica de Big Data en el sector de la salud. En el siguiente, se expone el estado en que se encuentra actualmente su adopción en nuestro entorno sanitario. Y finalmente se presentan algunos casos de éxito.

Si bien el Hospital Clínico, estuvo un tiempo operando con la plataforma BI de qlikview el cual permitía recolectar los datos desde diferentes orígenes como bases de datos SQL, datos de Excel, etc., se lograron modelamientos personalizados a nuestro gusto facilitando el entendimiento y manejo de los mismos, pero esto con el tiempo y por falta de personal capacitado se dejó de utilizar quedando a la deriva y volviendo a analizar los datos de forma manual, pero no óbstate a lo anterior esta solución de BI nunca estuvo enfocado y orientado a lo que se pretende en este proyecto, que es al estudio de los datos del pacientes y a la información a recabar por parte del cuerpo médico, sin dejar de lado la farmacología, en definitiva al área de salud, donde el paciente no quedara desvalido por falta de información o intermitencia en su tratamiento. A continuación, se deja constancia de los resultados de un estudio, donde perfectamente puede ser aplicado en nuestro país y que de seguro contribuirá al desarrollo o aplicación que pretende este trabajo de análisis.

4.1. Análisis predictivo en la tecnología Big Data

En esta sección se define conceptualmente la analítica predictiva, proceso relevante en la adopción de la tecnología de Big Data. Se presentan sus técnicas, modelos y métodos; las fases de un proyecto de construcción y mejora continua de modelos; así como soluciones disponibles y las principales aplicaciones.

4.2. Definición de analítica predictiva

Dentro de las posibilidades que ofrece la analítica de Big Data (análisis en tiempo real, grandes volúmenes de datos, datos no estructurados), en el presente trabajo se repara especialmente en el estudio del análisis predictivo, debido a la importancia que tiene su aplicación en el sector de la salud, como se pondrá de manifiesto más adelante.

El autor Joyanes define la analítica predictiva[22] como “una rama de la minería de datos centrada en la predicción de las probabilidades y tendencias futuras”, que “trata de analizar hechos actuales o históricos con el propósito de hacer predicciones sobre sucesos futuros”.

Conviene precisar, que la analítica predictiva existía ya desde unas décadas antes de que su relevancia en la industria se incrementara por efecto de las posibilidades que ofrecía la tecnología Big Data: cantidad de datos que se capturaban de las personas (por ejemplo, de transacciones online y redes sociales) y sensores (por ejemplo, de dispositivos GPS móviles) así como la disponibilidad de poder de procesamiento, ya sea basado en la Nube o en Hadoop.

[22] Citado en la Bibliógrafa “Big Data: Análisis de grandes volúmenes de datos en las organizaciones”

Se podría decir, por tanto, que existe un antes y un después en la analítica predictiva con respecto a Big Data.

El hecho de aprender a predecir a partir de datos en el ámbito científico ha recibido en ocasiones la denominación de "aprendizaje automático", una disciplina basada en la informática y la estadística a la que se dedican conferencias y programas académicos, mientras que la aplicación de esa rama de la ciencia al mundo donde tiene lugar la acción real, en el ámbito comercial, industrial, político, ha cambiado el nombre por el de "analítica predictiva" para referirse a “una tecnología que aprende de la experiencia (los datos) para predecir el futuro comportamiento de los individuos con el propósito de tomar mejores decisiones”. Sin dudad así, la Analítica predictiva, es “la pionera de la tendencia que existe actualmente para tomar decisiones basadas en datos, confiando menos en el instinto personal y más en una evidencia empírica y palpable”.

Lo esencial de los datos masivos consiste en que permiten hacer predicciones. No obstante, a pesar de que se los engloba en la ciencia de la computación llamada inteligencia artificial y, más específicamente, en el área llamada aprendizaje automático o de máquinas, esta caracterización puede inducir a error. Para dichos autores, la utilización de los datos masivos “no consiste en intentar enseñar a un ordenador a pensar como un ser humano. Más bien consiste en aplicar las matemáticas a enormes cantidades de datos para poder inferir probabilidades”. Asimismo, estos autores ponen de relieve que el buen funcionamiento de estos sistemas radica precisamente en que “están alimentados con montones de datos sobre los que basar sus predicciones. Es más, los sistemas están diseñados para perfeccionarse solos a lo largo del tiempo, al estar pendientes de detectar las mejores señales y pautas cuando se les suministran más datos”

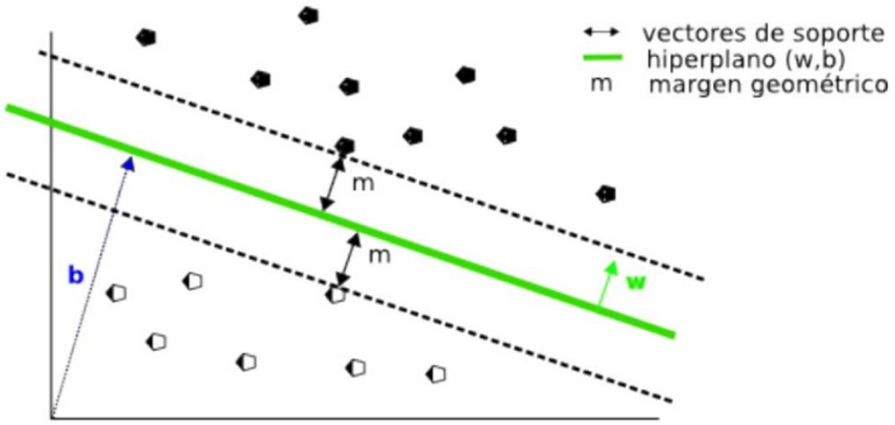
En un mar de datos en expansión constante, recolectados a partir de personas y sensores, la analítica predictiva proporciona herramientas de navegación esenciales para que las compañías e individuos alcancen exitosamente su destino, de forma que uno pueda responder del modo más conveniente para mantenerse en el curso más preciso, seguro, repetible, rentable y eficiente.

4.3. Técnicas, modelos y métodos de análisis predictivo

La analítica predictiva se materializa mediante la creación de modelos de conocimiento predictivos. Se podría definir como un mecanismo que predice un comportamiento de un individuo, como un clic, una compra, un comentario, una muerte o una mentira. Toma como datos de entrada las características del individuo y genera como salida una puntuación predictiva. Cuanto mayor sea la puntuación, más probable será que el individuo exhiba el comportamiento predictivo.

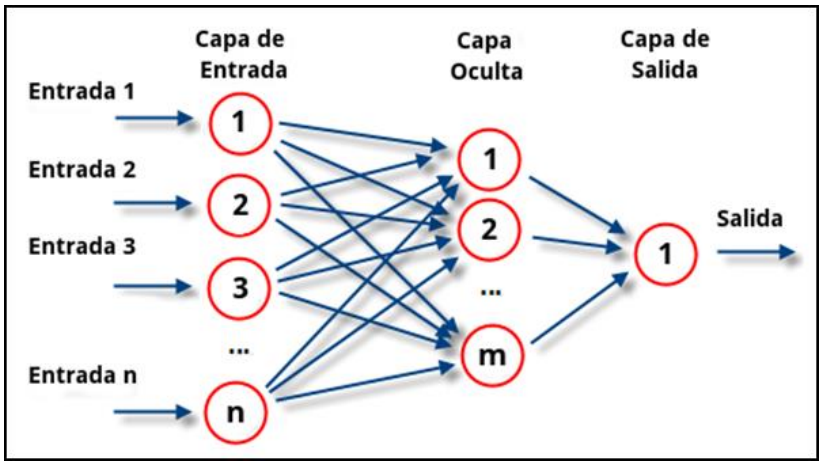
Los modelos predictivos son funciones matemáticas o algoritmos, capaces de determinar y aprender la correlación entre un conjunto de variables de datos de entrada, por lo general empaquetadas en un registro, y una variable de respuesta o de destino. Estos algoritmos forman parte de las técnicas y métodos de minería de datos. El creciente uso de la analítica predictiva para aplicarla sobre un componente importante de datos no estructurados ha impulsado a los desarrolladores de software a incluir en sus aplicaciones específicas un número cada vez mayor de algoritmos para cubrir un amplio espectro de posibles soluciones de modelado predictivo, de forma que el modelo óptimo se encuentre en la combinación de métodos. No obstante, hay un reducido grupo de algoritmos genéricos que suelen incluir tanto los fabricantes de software de código abierto como comercial, como los señalados a continuación:

- ***Máquinas de vectores de soporte (SVM)***. Se trata de un conjunto de algoritmos de aprendizaje supervisado que dan solución a problemas de clasificación y regresión. Una SVM construye uno o varios híper-planos en un espacio de dimensión mayor que el conjunto hallado calculando aquel que proporcione la mayor separación entre dos subconjuntos diferenciados, que será el "hiper-plano óptimo". Eso los proveerá de una etiqueta de clase y de una función de regresión que le otorgue valor predictivo. La predicción será que los puntos de un nuevo conjunto analizado por el modelo construido serán clasificados correctamente.

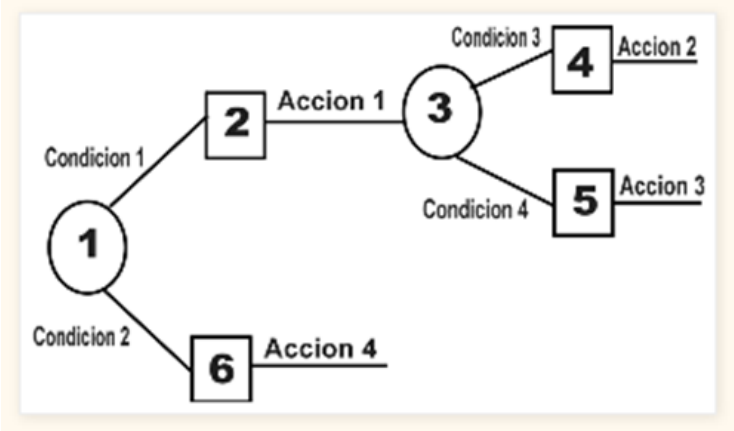


- ***Redes neuronales (NN).*** Las redes neuronales representan una estructura de aprendizaje automatizado inspirada en el funcionamiento del sistema nervioso de los animales. Está compuesto de una capa de entrada, con tantos nodos como número de campos y características que se están considerando; de una capa de salida, con un solo nodo que representa el campo predicho; y de una o más capas intermedias de nodos ocultos. Deberá establecerse una función de correlación entre los campos de entrada y de destino. Cuando hay más de una capa interpuesta de nodos ocultos puede

aprender mejor la función un modelo de red neuronal de retro propagación, que busca el ajuste de los valores intermedios desde el valor de salida.



- ***Árboles de decisión.*** Dado un conjunto de datos, se construyen diagramas de construcción lógica, similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que se presentan de forma sucesiva para la resolución de un problema. Al contrario que los modelos anteriores, resulta más fácil de usar y entender.



- ***De agregación o clustering***. Es un procedimiento que trata de agrupar de modo cercano a grupos de individuos con características semejantes. Entre los métodos más utilizados para establecer el agrupamiento figura el de los centroides, o kmeans y el de los vecinos más cercanos, o K-nn (K nearest neighbors). El algoritmo k-means trata de obtener la partición de un conjunto de n observaciones en k grupos, en el que cada observación pertenece al grupo más cercano a la media. Los elementos más próximos a la media son los centroides. En el conjunto inicial se pueden elegir aleatoriamente los centroides o bien las particiones. Se calcula la media de cada grupo y se repite el proceso hasta que la asignación de sus centroides no varía. El algoritmo k-nn es un método de clasificación supervisada, que se basa en establecer previamente ejemplos de entrenamiento ya clasificados. Un nuevo elemento que se desea clasificar se asignará a la clase a la que pertenezca el mayor número de vecinos más cercanos de un grupo de k elementos.

- ***Reglas de asociación.*** Se utilizan cuando una variable de destino o una medida similar no es importante, pero sí lo son las asociaciones entre los elementos de entrada. Por ejemplo, qué pueden tener en común las personas que además de comprar pañales y leche compran también cerveza. Esto sería un análisis de la cesta de la compra, que puede utilizarse para decisiones de marketing. El modelo se usa en muchas otras áreas, entre ellas la investigación en biología molecular.

4.4. Fases de un proyecto de construcción y mejora continua de un modelo de análisis predictivo

Toda vez que se haya determinado un objetivo para el análisis predictivo, un proyecto de construcción y mejora continua del modelo de análisis predictivo, de acuerdo con la metodología señalada por la consultora Forrester, debería progresar del siguiente modo:

- **Identificación de los datos necesarios y de su variedad de fuentes.** Los datos potencialmente valorables a veces están localizados en lugares de difícil acceso, tanto internos (silos de datos en aplicaciones de la propia empresa) como externos (medios sociales, datos del gobierno y otras fuentes de datos públicos). Las herramientas de visualización pueden ayudar a explorar los datos desde varias fuentes para determinar lo que puede ser relevante para el proyecto.

- ***Preparar los datos para el análisis***

  Calcular los campos agregados, limpiar los caracteres extraños, rellenar los datos que faltan, fusión de múltiples fuentes de datos, etc.

- ***Construir el modelo predictivo***

  Es la fase central del proceso. Mediante la herramienta elegida, se escogen, entre los mejores, uno o más algoritmos, dependiendo del tipo y la integridad de los datos y el tipo de predicción deseado. Los analistas ejecutan el análisis en un subconjunto de los datos llamados "datos de entrenamiento" y dejan aparte otro conjunto, el de "datos de prueba" que servirá para evaluar el modelo.

- ***Evaluar la eficacia y exactitud del modelo***

  El análisis predictivo no pretende llegar tanto a la exactitud como a un nivel alto de probabilidad. Para evaluar la capacidad de predicción del modelo, los analistas de datos utilizan el modelo para predecir el conjunto de "datos de prueba". Si el modelo predictivo puede predecir con el conjunto de datos de prueba, es un candidato para su implementación.

- ***Entregar copias del modelo para ponerlo en uso en el entorno operativo de sus compañeros de negocio***

  Hay poco valor en una predicción si no permite aprovechar una oportunidad de predicción o evitar un evento negativo. Los compañeros del negocio deben aprender a confiar en las predicciones de los modelos y los que crean los modelos tienen que aprender de sus socios en el negocio de lo que pueden ser las ideas más viables.

- ***Monitorizar y mejorar la eficacia del modelo***

  Los modelos predictivos son tan exactos como lo son los datos introducidos en ellos, y con el tiempo se pueden degradar o aumentar su eficacia. Para supervisar la eficacia de los modelos y el valor en curso, los datos recién acumulados se vuelven a ejecutar a través de los algoritmos. Si el modelo se vuelve menos preciso, los profesionales analistas tendrán que ajustar el modelo (por ejemplo, mediante el ajuste de los parámetros en los algoritmos) y / o solicitar datos adicionales.

4.5. Herramientas de análisis predictivo

Como se expone en este trabajo y según los estudios realizados, revelaban cómo una de las barreras para la adopción de esta tecnología era la escasez de personas con conocimientos avanzados para el aprovechamiento eficaz de su uso, especialmente en el campo de la analítica predictiva. Los fabricantes de software tomaron buena nota de ello y han procurado seguir ofreciendo continuas mejoras en sus productos, respondiendo así a un importante aumento de la demanda en todos los sectores.

Organizaciones de todas las industrias se están animando a dar valor al análisis predictivo. Con el crecimiento de la demanda, los proveedores de análisis predictivo están proveyendo herramientas que reducen la barrera y aumentan el atractivo a aquellos con menos habilidades estadísticas.

Fabricantes de software comercial de analítica predictiva:

- ***Líderes***

  En esta categoría coloca a IBM, SAS y SAP.

  IBM reúne un impresionante conjunto de prestaciones, poniendo la predicción en el centro. No importa cómo una organización desee empezar con el análisis predictivo, ya que cuenta con una opción para ellos. Su software ofrece prestaciones para construir modelos, realizar análisis y desplegar aplicaciones predictivas, tanto en las instalaciones como en la nube. El análisis predictivo de IBM puede tomar datos realmente grandes y proporcionar información decisiva. Por su parte, SAS continúa siendo una potencia en analítica. Con un enfoque estratégico desde 1973, SAS dispone de soluciones de análisis predictivo que ofrecen casi todas las características que un científico de

datos o usuario de negocios podría desear. También se mantiene al día respecto a las necesidades cambiantes de los usuarios. SAS Visual Analytics proporciona a los científicos de datos una solución “todo en uno” de herramientas de visualización y análisis predictivo. Las soluciones de SAS también se integran con las de código abierto R, Python y Hadoop.

- ***Actores Fuertes (Strong Performers)***

  Pertenecen a este grupo Alpine Data Labs, Alteryx, Angoss, Dell, FICO, KNIME, Oracle y RapidMiner. Los evaluadores consideran que todos ellos aportan cierto atractivo en sus productos, lo cual los mantiene como excelentes opciones para las empresas. Con mejores puntuaciones en estrategia, Alteryx, Angoss, FICO, Oracle, y RapidMiner serían también líderes.

- ***Contendientes***

  Microsoft y Predixion Software entran en esta categoría. Ambos comienzan desde un nicho, pero tienen mucho espacio de funcionamiento para crecer y ser un valor único para las empresas. Microsoft se centra exclusivamente en lo servicios en la nube y Predixion Software permite a Excel capacidades de análisis predictivo también en la nube.

En cuanto a las soluciones de código abierto, se emite el siguiente comentario:

En la opinión mayoritaria de los programadores, la comunidad de software de código abierto (OSS) es un potente motor de análisis predictivo. El lenguaje de programación de código abierto **R**, para estadísticas y análisis predictivo, está omnipresente en los entornos universitarios Los desarrolladores de aplicaciones también tienen una gran cantidad de bibliotecas API disponible para preparar los datos y construir modelos predictivos utilizando Java, Python y Scala. Apache Mahout y WEKA tienen APIs de Java. Apache Spark MLlib incluye API para Java, Python y Scala. Los desarrolladores de Python pueden utilizar NumPy y SciPy para preparar los datos y construir modelos predictivos.

4.6. Principales aplicaciones de la analítica predictiva

El análisis predictivo puede aplicarse a muchas circunstancias y entornos, en los que se trata de resolver problemas de riesgos o de oportunidades que se puedan presentar en el futuro. Los más frecuentemente citados inciden en los siguientes campos:

- ***En la gestión de relación con los clientes (CRM), interesa conocer:***

    - A qué clientes se puede dirigir una campaña publicitaria con mayor opción de respuesta.
    - Quiénes pueden estar a punto de abandonar e interesa retenerlos.
    - Quiénes son buenos compradores, para premiar su fidelidad.
    - Cuáles son las preferencias de compra para proponer ofertas personalizadas, ventas cruzadas, etc.

- En qué circunstancias pueden consumir una mayor cantidad de productos, para poder mantener el abastecimiento.

- ***Relacionados con la asistencia sanitaria***

  - Conocer cuáles son los factores de riesgo de lo paciente para desarrollar problemas crónicos: asma, diabetes, enfermedades cardiovasculares, etc., con el fin de establecer un mejor control en la asistencia ambulatoria, evitando así la hospitalización.
  - Identificar patrones de riesgo de infección en pacientes monitorizados.
  - Detectar riesgos relacionados con el patrón genético, que permitan desarrollar medidas de prevención y tratamiento personalizados.
  - Conocer el riesgo de reingreso en el hospital antes de dar una alta anticipada.
  - Detectar fraudes y abusos en el cuidado de la salud.

- ***Análisis de cobros***
  - Conocer qué clientes pueden resultar morosos, con el fin de evitar el pago a crédito o de gestionar los cobros mediante agencias de recaudación externas.

- ***Detección del fraude***
  - Reconocer qué créditos son inapropiados, qué transacciones, cobros de prestaciones, reembolsos, reclamaciones son fraudulentas por suplantación o robo de identidad u otros motivos; se podrán tomar medidas a tiempo, evitando la continuación del proceso de forma automática y filtrándolo hacia una auditoría humana.

4.7. Analítica del Big Data en el área de la salud

En este apartado se incluyen algunos de los trabajos que más contribuyeron en su momento a dar a conocer las posibilidades que la tecnología Big Data ya había comenzado a ofrecer en el área de la salud. Así lo pusieron de manifiesto especial dos recocidos informes de referencia, como son el extenso informe que el McKinsey Global Institute (MGI) publicó en junio de 2011 bajo el título, Big Data: The Next Frontier for Innovation, Competition and Productivity[23].

4.7.1. Informe de Big Data: The Next Frontier for Innovation, Competition and Productivity

EL siguiente informe se cita a modo de documentación y modelo a seguir, en este documento se destaca el potencial transformador de Big Data en cinco grandes sectores, que los expertos de la consultora MGI han estudiado en profundidad: la atención sanitaria y el comercio minorista en EE.UU., la administración del sector público en la Unión Europea y a nivel global, la fabricación y los datos de geolocalización de personas y entidades.

En el caso de la atención sanitaria en EE.UU., se estima que en el plazo de diez años el valor creado mediante el uso de Big Data podría alcanzar los 300.000 millones de dólares al año, lo que supondría un ahorro anual de más de 200.000 millones en el gasto nacional en el sector. En el informe se distinguen quince recursos Big Data relacionados con el sector sanitario, que se distribuyen en diferentes categorías según ejerzan su impacto sobre la operativa

[23] Señalado en Bibliografía.

clínica, la investigación y desarrollo, el sistema de pagos y fijación de precios, el impulso de nuevos modelos de negocio o la salud pública.

La mayor parte de ellas pertenece a las dos primeras categorías, que son las que se comentan a continuación.

1) En la **operativa clínica**, Big Data podrá impulsar la creación de valor de cinco modos diferentes:

   - Mediante programas CER (Comparative Effectiveness Research) de investigación comparativa de la eficacia de los tratamientos aplicados a los pacientes. El análisis detallado de grandes conjuntos de datos como las características de los pacientes, de los resultados de los tratamientos y de los costos que han supuesto, puede identificar cuáles son los tratamientos más eficaces para cada caso, desde el punto de vista clínico y también bajo criterios de costo-eficacia. Implementando esta investigación, el sistema de salud podrá reducir la incidencia de tratamientos excesivos (intervenciones que perjudican más que benefician), o bien de tratamientos insuficientes (que se deberían haber prescrito, pero no lo fueron). Tanto en un caso como en el otro derivarían en peores resultados para los pacientes y mayores costos de atención sanitaria a largo plazo. Estos programas ya han venido aplicándose con éxito en el Reino Unido, Alemania, Canadá y Australia.

   - Con sistemas de soporte a las decisiones clínicas, para mejorar la calidad de las prescripciones. Estos sistemas registran las órdenes de tratamiento de los médicos y

realizan un análisis comparativo frente a las pautas recomendadas, alertando de potenciales errores o efectos adversos del medicamento. De este modo, se pueden reducir las reacciones adversas y la tasa de errores médicos, así como de reclamaciones de responsabilidad civil derivadas de los mismos. En un estudio realizado en una unidad de cuidados intensivos pediátrica en una gran área metropolitana de Estados Unidos, una herramienta de sistema de apoyo a las decisiones clínicas redujo las reacciones y eventos adversos por medicamentos en un 40% en tan sólo dos meses.

- Creando transparencia sobre los datos médicos. Los conjuntos de datos operacionales y de rendimiento del proveedor pueden ser analizados para crear mapas de procesos y cuadros de mando que permitan la transparencia informativa. El objetivo es identificar y analizar las fuentes de variabilidad y de pérdidas en los procesos clínicos con el fin de optimizar los mismos. Simplemente, la publicación de datos de costos, calidad y rendimiento, incluso sin una recompensa financiera tangible, a menudo crea la competencia que impulsa mejoras en el rendimiento.

- La monitorización de pacientes a distancia. La monitorización remota recoge datos de pacientes con enfermedades crónicas (diabetes, insuficiencia cardiaca, hipertensión, etc.). El análisis en tiempo real de los datos servirá para controlar la evolución, vigilar el cumplimiento (si los pacientes están realizando el tratamiento prescrito) y mejorar futuras opciones de tratamiento y de medicación. Los sistemas

incluyen dispositivos que monitorizan el estado del corazón, envían información sobre los niveles de glucosa en sangre, transmiten las instrucciones de los cuidadores, o incluso pueden incluir la tecnología "chip en una píldora", que deja constancia de que el paciente está tomando la medicación. Por lo general, el uso de datos de los sistemas de monitorización remota permite reducir la estancia del paciente en el hospital, disminuir las visitas al servicio de urgencias y mejorar la focalización de la atención de enfermería a domicilio y de las citas con el médico a nivel ambulatorio, reduciendo a largo plazo las complicaciones de la enfermedad que impliquen hospitalización.

- La analítica avanzada aplicada a reconocer perfiles de pacientes, permitirá identificar personas que se beneficiarían de un cuidado o de un cambio proactivo en el estilo de vida. Por ejemplo, al identificar pacientes que están en alto riesgo de desarrollar una diabetes, estos podrán beneficiarse de un programa de atención preventiva.

2) En **la investigación y desarrollo en la industria farmacéutica**, Big Data podrá también mejorar la productividad en los departamentos relacionados con la industria farmacéutica.

- Modelado predictivo. En el desarrollo de nuevos medicamentos, la agregación de datos masivos a la investigación en las fases preclínicas y en fases tempranas de ensayos clínicos permitirá la construcción de un modelo predictivo que anticipe los resultados finales en términos de eficacia, seguridad y efectos secundarios potenciales del

producto. Ello reducirá considerablemente el tiempo de investigación que habitualmente venía suponiendo largos y costosos ensayos clínicos.

- Herramientas y algoritmos estadísticos para mejorar el diseño de los ensayos clínicos. Estos recursos utilizan técnicas de minería de datos para optimizar la selección de pacientes candidatos a estudiar en el ensayo clínico, los sitios donde hay mayor concentración de casos, el diseño de protocolos más eficaces, la gama de indicaciones que se aplicarán al producto, el modelado comercial, las posibilidades de aprobación regulatoria o la selección de investigadores que han de dirigir el proyecto.

- Análisis de los datos de ensayos clínicos. Este procedimiento posibilitará la identificación de nuevas indicaciones del fármaco, así como el descubrimiento de efectos adversos. Una vez que el producto ya esté en el mercado, la recogida en tiempo real de los informes de reacciones adversas hará posible un fármaco vigilancia más ágil y eficiente al poder detectar los casos raros que han escapado a la insuficiente potencia estadística de un ensayo clínico.

- La medicina personalizada. Este recurso Big Data tiene como objetivo examinar las relaciones entre una determinada variación genética en las personas, la predisposición a padecer determinadas enfermedades y las respuestas específicas que esas personas puedan tener con ciertos medicamentos. Posteriormente se tendrá en cuenta la variabilidad genética de los individuos en el proceso de

desarrollo de fármacos. La medicina personalizada promete mejorar la atención de la salud de tres maneras:

a) Mediante la detección temprana y el diagnóstico antes de que un paciente desarrolle síntomas de la enfermedad.

b) Buscando terapias más eficaces teniendo en cuenta que pacientes con el mismo diagnóstico se pueden segmentar con arreglo a un perfil molecular diferente, es decir, no responden de la misma manera a la misma terapia, en parte debido a la variación genética.

c) Ajustando las dosificaciones del fármaco de acuerdo con el perfil molecular del paciente para reducir al mínimo los efectos secundarios y maximizar la respuesta beneficiosa deseada.

- Análisis de patrones de enfermedades. La identificación de patrones y tendencias de enfermedades permite modelar la demanda y los costes futuros, así como tomar decisiones estratégicas de inversión. Este análisis puede ayudar a las empresas farmacéuticas a optimizar el foco de sus actividades y a prever la asignación de recursos, incluyendo equipamiento y personal.

4.7.2. Informe Big Data in Healthcare Hype and Hope

Este estudio constituye un valioso aporte. Se basa en el análisis de proyectos y herramientas que ya están en funcionamiento, por muchas organizaciones siguiendo diferentes estrategias y vislumbrado cómo Big Data se está convirtiendo en una creciente fuerza de cambio en el panorama sanitario.

Dentro de su detallada exposición, se pueden considerar de especial interés los apartados que dedica a las cuatro principales dimensiones que caracterizan a Big Data aplicadas al sector sanitario y a las formas que propone para mejorar el cuidado de la salud.

4.7.2.1. Dimensiones de Big Data, aplicada a los datos del área de la salud

Volumen. Los datos de asistencia sanitaria están experimentando un crecimiento exponencial. En 2012 el volumen de los mismos era de 500 petabytes (1015 bytes) en todo el mundo. Se estima que en 2020 llegará a 25.000 petabytes. Este crecimiento provendrá por una parte de la digitalización de los datos ya existentes en las organizaciones, pero, sobre todo de los nuevos datos generados ya en modo digital: registros médicos personales, imágenes radiológicas, ensayos clínicos, secuenciación genómica, lecturas de sensores biométricos o imágenes en 3D.

Variedad. La diversidad de fuentes trae consigo también una variedad de tipos de datos (estructurados, semi-

estructurados y no estructurados), lo que confiere a esta dimensión un peso mayor que las demás en complejidad para el análisis de los mismos. Los centros de atención médica han estado generando principalmente datos no estructurados: registros médicos, notas manuscritas de médicos y enfermeras, registros de admisión y de altas hospitalarias, recetas en papel, imágenes radiográficas, de resonancia magnética, TAC, etc. Por otro lado, los datos estructurados y semiestructurados proceden de la contabilidad, la facturación electrónica, algunos datos actuariales, clínicos o de lectura de instrumentos de laboratorio.

Big Data podrá, además, capturar y almacenar nuevas formas de datos (estructurados y no estructurados) procedentes de otras fuentes, como los generados por sensores de dispositivos fitness, de las App de smartphones, de la investigación genómica o de los medios sociales. Combinar datos tradicionales con las nuevas formas de datos, tanto a nivel individual como poblacional, radica el potencial de Big Data en el área de salud, y ya se está comprobando cómo los conjuntos de datos procedentes de multitud de fuentes apoyan de modo rápido y fiable la investigación y el descubrimiento.

Velocidad. Aunque en muchas ocasiones no se requiere una gran velocidad en los procesos de captación, análisis y entrega de información, hay algunas situaciones médicas que exigen un flujo de datos constante en tiempo real:

monitorización de signos vitales en traumatismos, en la sala de operaciones durante la anestesia, en la UCI, etc.

Veracidad. Esta característica hace referencia a la incertidumbre de la calidad de algunos datos entrantes y del valor predictivo que se debe otorgar a los resultados del análisis en el que se hayan incluido esos datos. En la asistencia sanitaria es muy importante que el modelo predictivo resultante del análisis sea de calidad.

En algunos casos Big Data recurrirá a herramientas que filtren datos entrantes dudosos que podrían ser erróneos, para corregirlos o eliminarlos (textos manuscritos en recetas, anotaciones, etc.). En otras ocasiones información que poseía escasa relevancia estadística o no se detectó al trabajar con conjuntos de datos insuficientes, sí la va a tener o será detectable al agregar conjuntos de datos mayores.

Un ejemplo es someter a evaluación el modelo predictivo utilizado en los ensayos clínicos mediante la incorporación automatizada y en tiempo real de los datos de reacciones adversas a medida que son reportadas cuando el medicamento ya ha sido comercializado, donde se trabajará con una muestra prácticamente equivalente a la población total. Como ya se ha señalado, esto redundará en una mayor eficiencia del proceso de fármaco vigilancia.

4.7.2.2. Formas de utilizar Big Data para una mejor salud

A continuación, se señalan algunas formas de utilización de Big Data, referente principalmente a los retos en la atención médica, tarea no menor.

- Apoyo a la investigación genómica, La cual buscan identificar el perfil genético o biomolecular de las personas con el fin de encontrar variaciones que lo relacionen con predisposición a padecer determinadas enfermedades y consecuentemente, elegir las medidas más adecuadas de prevención y tratamiento. El traslado a la práctica clínica o al desarrollo de nuevos medicamentos basándose en el conocimiento adquirido en la investigación es lo que hace posible la denominada medicina personalizada.

- Transformar datos en información (y la información en datos), Dada la creciente cantidad y variedad de datos que se generan en el ámbito de la salud, y la alta proporción de no estructurados (cercana al 80% según algunas estimaciones), un aspecto clave para la mejor gestión de los mismos y su conversión en información utilizable es hacer que los no estructurados pasen a ser estructurados y así facilitar su manipulación por la máquina, ya algunas Herramientas utilizan un procesamiento de lenguaje natural (NLP) para convertir un registro médico narrativo en datos que sean comprendidos por la máquina. Así como también existen Software que realizan análisis predictivo para identificar

pacientes con riesgo de reingreso, para prevención de infecciones hospitalarias o de enfermedades crónicas.

- Apoyo al auto cuidado y colaboración ciudadana, Mediante la recogida de datos procedentes de sensores instalados en dispositivos “wearables” y smartphones, Big Data puede devolver información a los usuarios sobre su estado de salud. Aplicaciones para móviles, permiten el intercambio rápido y seguro de información entre pacientes y registros médicos. Plataformas, basada en la nube, recopilan datos en tiempo real desde los smartphones acerca del comportamiento, activo y pasivo, de los pacientes (movimiento, comunicación, uso del móvil, etc.) para ayudar a los médicos, enfermeras, familiares y pacientes a gestionar su salud, comenzando por las enfermedades crónicas. Con el consentimiento del paciente, la administración de los datos y análisis quedarán a disposición de los proveedores de salud y los investigadores a través de un panel de control.

- Apoyo a los proveedores de salud y a la mejora en la atención al paciente, Big Data se emplea también como herramienta para la inteligencia de negocio en la gestión sanitaria. Las plataformas actuales permiten explorar y medir en segundos una enorme cantidad de datos clínicos, de gestión financiera, redes sociales, etc., para medir el rendimiento que servirán para mejorar la atención al paciente y reducir costos.

- Aumentar el conocimiento, Existen diversas herramientas que se utilizan para aumentar el conocimiento y resolver una gran variedad de problemas basándose en los datos. Existen software que utilizan Big Data para identificar medicamentos falsificados, protegiendo así la salud del paciente, y para que las empresas farmacéuticas puedan rastrear la distribución de medicamentos y prevenir el fraude. Así como también existen otros que recogen los datos de los pacientes con asma y les proporciona retroalimentación que les ayuda a controlar mejor su enfermedad. Un dispositivo móvil mediante una aplicación para IOS / Android se conecta con inhaladores para el asma que disponen de un sensor que controla la hora y el lugar de los hechos en el momento de tomar la dosis y recoge datos de los posibles factores ambientales desencadenantes, así como de los síntomas. Sickweather LLC escanea las redes sociales (Facebook, Twitter) para rastrear brotes de enfermedades.

- Agregación de datos para construir un ecosistema mejor, La agregación de muchos más datos dispares procedentes de fuentes externas al conjunto inicial disponible constituye una interesante aplicación de Big Data que permite realizar nuevos tipos de análisis y facilitar respuestas a las grandes preguntas que realizan los investigadores. Distintas empresas han habilitado plataformas seguras de interconexión sanitaria global basada en la nube, que recoge y almacena vía inalámbrica todos los datos biométricos que le son

enviados desde un Smartphone y otros dispositivos sanitarios (medidores de glucosa, tensiómetros, balanzas, etc.). Como ejemplo la Empresa Nike que han sabido utilizar los datos que han obtenido durante tiempo con un fin diferente al de la inicial. Su objetivo fue ser una herramienta para controlar nuestro estado de salud, bien mientras dormimos o mientras realizamos alguna actividad física. Por esto Nike, que posee una importante cantidad de datos acerca del estado de salud del pueblo americano, ha decidido abrir una aseguradora de salud y han empezado a invertir en hospitales para introducirse en el sector.

Los estudios basados en ese gran fondo de datos tendrán mayor relevancia estadística y valor predictivo.

Dentro de este apartado, se destaca el papel que ya está jugando el sistema Watson de IBM como el primero de una nueva clase de plataformas analíticas y de soporte a las decisiones que utiliza un análisis en profundidad de contenidos, un razonamiento basado en la evidencia y un procesamiento del lenguaje natural para apoyar el diagnóstico y la toma de decisiones clínicas de modo más rápido y preciso.

Entre sus principales características se señalan las siguientes:

- Dotado con 21 subsistemas de computación, 16 terabytes de memoria RAM y del sistema de procesamiento del lenguaje natural más avanzado en el mundo, Watson puede almacenar una enorme cantidad de datos, desde los de las historias clínicas de los pacientes hasta todas las publicaciones de revistas médicas, incluyendo las que recogen los últimos avances en el diagnóstico y tratamiento, siendo capaz de leer toda esa información a razón de 200 millones de páginas en tres segundos y recordar cada palabra.

- Watson revisa los datos de la historia del paciente, antecedentes familiares, síntomas y resultado de pruebas, formula hipótesis que contrasta con la información almacenada y devuelve un listado de enfermedades clasificadas por nivel de confianza, que ayudan al médico en el diagnóstico y tratamiento.

- Finalmente, uno de los ejemplos más claros es su utilidad en el campo de la oncología. Watson no es un médico, pero tiene la capacidad de interrelacionar y sacar conclusiones de todos los millones de datos que se producen en el mundo en el campo de la medicina. "Esta información es imposible que sea absorbida por un humano, ni el mejor médico del mundo podría. Watson puede. Por eso es que esta tecnología se convertirá en el mejor asistente del humano,

no lo reemplazará, lo ayudará a hacer su trabajo mucho mejor", aclara Rodrigo Seguel, gerente de tecnología de IBM Chile.[24]

4.8. Big data en nuestro sistema de salud

A la vista de lo analizado, indagado y estudiado en este trabajo de investigación, parece indudable que las posibilidades que ofrece la adopción de la tecnología Big Data al área de la salud son muchas y ventajosas.

Desde el punto de vista de la incorporación y adopción de las Tecnologías de la Información (TI), con especial énfasis en el trabajo colaborativo y virtuoso entre los mundos público y privado, articulados desde los proyectos de “Sistemas de Información de la Red Asistencial” (SIDRA) y otros afines, que hoy permiten a la empresa Rayen Salud situarse como un referente destacado en el ámbito de la Informática Clínica en Chile y con amplia proyección a nivel Latinoamericano.

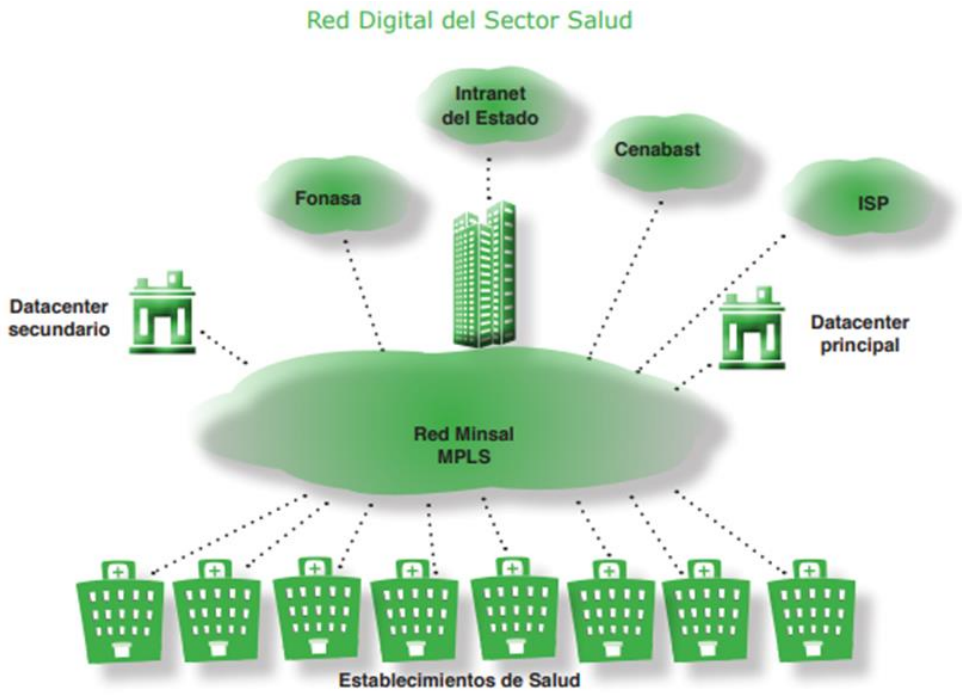
[24] Artículo citado en bibliografía.

Este proceso ha ido dando los pasos necesarios para la incorporación de los Sistemas Informáticos de Registros Clínicos Electrónicos, que aborden los procesos transversales del registro de las atenciones a las personas. Desde la generación de una Red Privada de Internet MINSAL, pasando por la provisión de infraestructura a los establecimientos para sus redes de datos locales y eléctricas, dotación de equipamiento computacional y capacitación a funcionarios técnicos sobre la administración de Proyectos Complejos y otras temáticas tecnológicas, como también capacitación o alfabetización digital a miles de funcionarios.

Así, nuestro país es observado con atención por otros de Latinoamérica y el Mundo, debido al éxito obtenido en materia de incorporación y adopción de Tecnologías de la Información (TI) en la Salud Pública, con un alcance en el uso de Ficha Clínica Electrónica del 80% en la Atención Primaria de Salud (APS) y del 45% en los Hospitales, declarado por las Autoridades Sanitarias del país, permitiendo importantes resultados de rentabilidad social y económica; ahorro y contención de costos; trazabilidad de las atenciones en salud y mejoras considerables en la seguridad de la información de los pacientes.

El estado actual en materia de infraestructura y conectividad de la Red Digital del Sector Público de Salud, conecta a todas las instituciones del área (hospitales, consultorios, centros de urgencia, seremis, oficinas del Minsal y organismos autónomos) y que tiene la estructura que se presenta en la figura siguiente[25]. Se trata de una red privada virtual (VPN) de cobertura nacional, que conecta a 1.500 establecimientos de atención primaria y secundaria de salud probablemente es la red más grande de la región, que cuenta además con una intranet sectorial, un portal web y un correo institucional de 40.000 casillas. Esta red provee comunicación de voz y datos, así como también de servicios de videoconferencia e internet.

[25] Articulo PDF señalado en bibliografía como: MAPA DE RUTA, Plan Estratégico de Tecnologías de Información.



No obstante, a lo señalado anteriormente, solo está relacionado con el sector público y no se vislumbra una cohesión en el cual se implemente un Plan Nacional de Uso Secundario de Información, que permita gestionar un Big Data Sanitario, el que permita desarrollar estudios epidemiológicos en tiempo real e incorporar los elementos de la gestión clínica prospectiva. Sin omitir, por otro lado, los principales obstáculos que será necesario ir superando, retos que debe afrontar nuestro sistema sanitario para aprovechar las ventajas que promete la tecnología Big Data, cuyos resultados, en mi opinión, no serán ni inmediatos ni siempre beneficiosos.

A modo de ejemplo, mencionaré algunos de ellos

- Extraer conocimiento de fuentes heterogéneas y complejas, a veces no estructuradas.

- Comprender notas clínicas no estructuras en su contexto correcto.

- Gestionar adecuadamente gran cantidad de datos de imagen clínica y extraer información útil para generar biomarcadores.

- Analizar los múltiples niveles de complejidad que van desde los datos genómicos hasta los sociales.

- Capturar los datos de comportamiento de los pacientes, a través de distintos sensores, con sus implicaciones sociales y de comunicación.

- Evitar los problemas de privacidad que pueden generar riesgos para los individuos.

Entre tanto, “habrá que ir desarrollando nuevos marcos y sistemas de referencia que faciliten trabajar en las tres grandes fuentes de Big Data sanitaria: la Historia Clínica e Imagen Médica, las redes sociales y sensores.

Por otro lado, la tecnología que utilizan los datos masivos será la gran aliada de la medicina del futuro o Medicina de las 4P: **personalizada**, **predictiva**, **preventiva** y **participativa**. El papel que esta nueva tecnología podrá desempeñar en cada uno de estos aspectos de la gestión clínica. Para sacar su máximo partido, en la sanidad del futuro sería preciso capturar, almacenar y analizar todos los datos disponibles sobre ensayos clínicos, historiales médicos, secuenciación de ADN de pacientes, información procedente de redes sociales. Se debería disponer, por tanto, de una enorme base de datos compartida entre todos los hospitales, clínicas y consultorios del área de la salud.

Además, se advierte que en la actualidad es necesario superar algunas trabas, entre las que figura la barrera tecnológica. La tecnología Big Data tiene que consolidarse todavía en el ámbito sanitario, por lo que se cree necesario que

aumenten las inversiones públicas y privadas en este tipo de soluciones. No obstante, el factor más importante sea el humano, es decir, científicos de datos que sepan sacar provecho de la tecnología. Es crucial contar con la presencia de analistas de datos expertos en el ámbito de la salud para que, a través del uso de tecnologías Big Data, puedan dar el soporte adecuado a los médicos en la toma de decisiones relativas a sus pacientes."

Los expertos consideran que Big Data está contribuyendo a reinventar el sistema sanitario. Uno de los cambios observados hace referencia a la transformación del tradicional seguimiento periódico del enfermo en consulta a un proceso continuo en el que, a través de una plataforma digital o de una app, los profesionales pueden controlar minuto a minuto la evolución del paciente y realizar mediciones constantes"

La bajada de costes en el área de la genómica y la proliferación de los wearables son dos de los aspectos que van a visualizar un volumen de datos médicos nunca visto hasta ahora.

La bajada de costos mundial ha masificado el acceso a esta tecnología y abre un nuevo abanico de posibilidades para el tratamiento personalizado y el análisis de datos médicos.

Algunas de las principales aplicaciones del estudio del genoma hoy en día son:

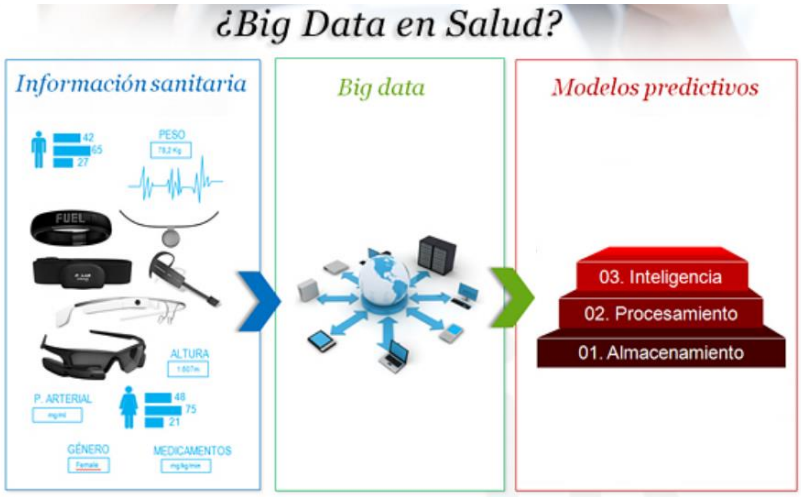
- El uso de modelos predictivos para identificar pacientes de alto riesgo, por ejemplo, de diabetes de tipo 1.

- Clasificación de subtipos de enfermedades para seleccionar tratamientos clínicos dirigidos y más precisos, por ejemplo, en cáncer.

- Proveer mejor información para el cribado de candidatos en los test clínicos de fármacos y tratamientos.

Por otro lado los avances en este campo de las wearables está dejando el desarrollo de nuevos dispositivos que se implantan en el cuerpo del usuario, bajo esta filosofía se están desarrollando sensores que controlan la cantidad de glucosa de un paciente con diabetes de tal forma que un dispensador electrónico inyecte de forma automática la cantidad de insulina precisa, este mismo sensor tomaría datos relevantes sobre la salud del paciente y se enviarían vía internet a los ordenadores, tablets o dispositivos del equipo médico.

Asimismo, lo demuestra el desarrollar de **Smart protocols**, un sistema de apoyo a la decisión clínica para facilitar el seguimiento ante diferentes patologías, el que será piloteado en el Hospital Padre Hurtado en la Región Metropolitana y en la Región de la Araucanía en el 2017. Las patologías tienen protocolos de atención y este sistema será una forma de apoyo y sugerencia para el médico. Si se detecta que un paciente tiene cáncer, sugerirá un tratamiento basado en protocolos internacionales y nacionales.



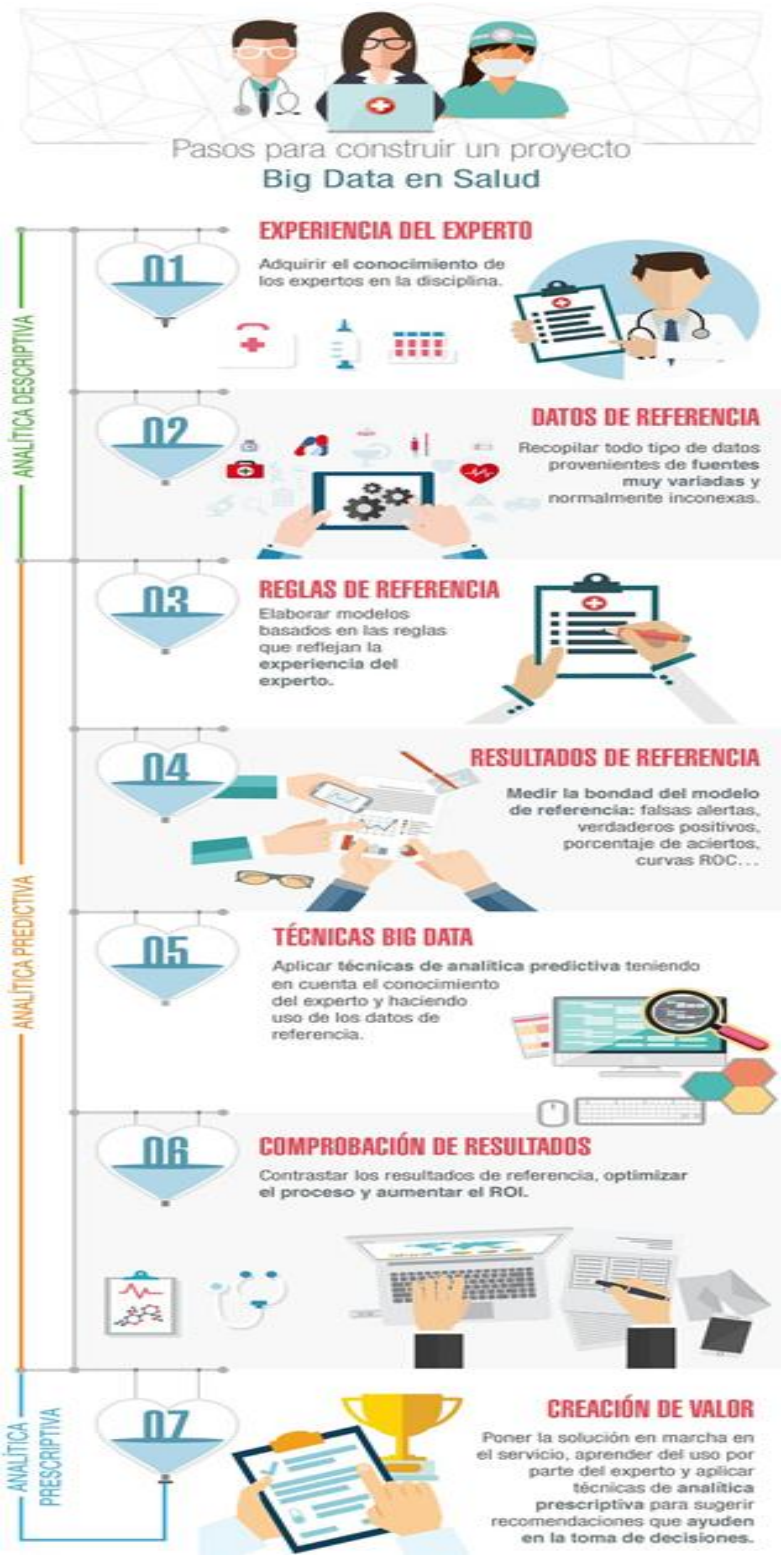
En cuanto a los avances en el sector privado, cabe señalar, que para las clínicas y hospitales del sector privado la ficha clínica es un estándar, por lo tanto, el desafío pasa por unificar los registros clínicos, de manera que el paciente se mueva por los centros con su historial.

Así como comenta Federico Becker, de Everis, que se implementara en la Fundación Arturo López Pérez, centro referencial en el tratamiento de pacientes con cáncer, el software **ehCOS** que resuelve el registro clínico electrónico en atención primaria, de especialidades y en hospitales y clínicas. Este sistema facilita el intercambio de información entre los diferentes prestadores de salud, permitiendo la visualización unificada del registro clínico de cada paciente, integrando servicios como laboratorio, imágenes y banco de sangre", comenta. Por otro lado, mediante la modalidad de atención Mobile Health, la firma Citrix busca que los médicos accedan a la historia clínica de sus pacientes desde cualquier dispositivo. Esta modalidad incluye la unificación de los registros en fichas médicas únicas, que permiten a los usuarios ahorrar tiempo en exámenes y horas de atención. Las soluciones Citrix XenApp y XenDesktop permiten tener aplicaciones como servicio para equipos médicos que trabajen en varias instalaciones, mediante dispositivos como Smartphone, Tablet, computadores. Esto asegura a los médicos un tiempo de inactividad nulo, porque obtienen acceso en tiempo real a sus escritorios y reduce los costos[26].

En Chile la Tecnología y los Servicios Informáticos asociados están disponibles para seguir avanzando en los desafíos planteados como país al 2020, en un contexto de mejora continua y sin retroceder al papel y el lápiz en aquellos procesos consolidados del sector sanitario. Ello asegurará avanzar con mayor certeza y eficiencia en la gestión y atención de las personas en salud.

[26] señalado Álvaro Guerra, Citrix Field Sales Manager Chile.

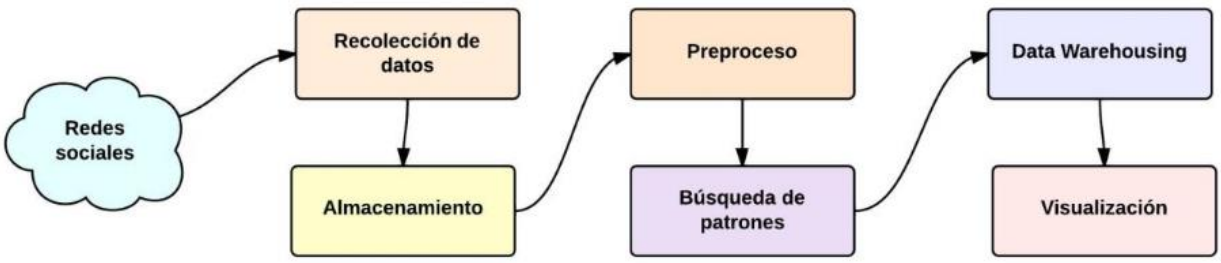
4.9. Pasos para construir un proyecto Big Data en Salud



5. CASOS DE ÉXITO

Este apartado presenta algunos casos únicos de utilización de Big Data que pueden servir para ilustrar las posibilidades que esta tecnología ofrece al sector de la salud.

Antes de exponer los casos de éxito, primero señalaremos y definiremos el diagrama de un caso muy utilizado, un ejemplo sencillo en donde la tarea consistirá en la descarga a través de las redes sociales de información relacionada con una empresa. Una vez recolectada la información se almacena en Hadoop para realizar procesos de análisis para determinar cuáles son las palabras y patrones más comentados cada día para poder realizar un análisis de sentimiento.



Diagrama, flujo de información del caso de uso

En la figura, se muestra un diagrama de una capa de recolección de datos que se conecta a una o varias redes sociales y descarga la información. La información descargada se almacena en alguna base de datos o sistema de ficheros para poder realizar un pre proceso de la información. Este pre proceso consiste en la purificación del texto para poder realizar mejor los algoritmos de búsqueda de patrones. Los resultados de la búsqueda se almacenan en una herramienta de data warehousing para poder ser visualizada.

- **El proyecto Artemis para la unidad de prematuros de Hospital For Sick Children de Toronto.**

Se implementó por primera vez en el Hospital para Niños Enfermos de Toronto en agosto de 2009 y ha estado funcionando continuamente desde entonces. Utilizando una plataforma de Big Data de IBM (IBM InfoSphere Streams) desarrolló una aplicación que recoge y analiza en tiempo real hasta dieciséis flujos continuos de datos de los dispositivos de monitorización de signos vitales de bebés prematuros (ritmo cardíaco, frecuencia respiratoria, temperatura, etc.) de modo que vigile la aparición de cambios apenas perceptibles (que la máquina ha aprendido a reconocer basándose en el análisis de datos históricos) que suelen acontecer 24 horas antes de que aparezcan síntomas de infección grave. La detección de esos cambios enciende la alerta que permite a los médicos comenzar el tratamiento de la infección o bien los advierte de que el tratamiento que se está haciendo no es efectivo.

- **La aplicación Flu Trends de Google**

Probablemente, el caso más referenciado que puso de manifiesto el poder predictivo que en minería de datos tiene el análisis con Big Data fue la aplicación **Flu Trends**, desarrollada por los ingenieros de Google, que permitió determinar casi en tiempo real por dónde se estaban propagando brotes de gripe y predecir su trayectoria posterior. Su eficacia quedó probada a raíz de la epidemia causada por el virus H1N1 entre 2009 y 2010.

La originalidad que presentaba el método de Google fue que no estaba ligado a datos del diagnóstico de gripe procedentes del estudio médico, sino que se basaba “en los Big Data, los datos masivos: la capacidad de la sociedad de aprovechar la información de formas novedosas para obtener percepciones útiles o bienes y servicios de valor significativo”. Su trabajo había consistido en hallar correlaciones entre la frecuencia de ciertas búsquedas de información y la propagación de la gripe

a lo largo del tiempo y del espacio. Una enorme cantidad de modelos matemáticos compararon sus predicciones frente a los datos oficiales de años anteriores, lo que sirvió para identificar una combinación de cuarenta y cinco términos de búsqueda, como "gripe", "catarro", "fiebre" o "dolor de cabeza", que, en su conjunto, ofrecía una fuerte correlación con los datos oficiales. Google puso su aplicación a disposición de los responsables sanitarios en más de dieciséis países para ayudarles a anticipar y combatir más eficazmente los brotes de gripe.

- ***Predicción de respuesta al tratamiento en linfoma de Hodgkin***

El linfoma de Hodgkin (LH) es un tipo de cáncer cuyo porcentaje de curación es elevado, pero la respuesta al tratamiento es difícil de predecir y muy cambiante según los pacientes. La valoración de progresión de la enfermedad se realiza mediante imagen radiológica (TAC y PET), y para evaluar el pronóstico de respuesta al tratamiento se ha estado utilizando un conjunto de 35 variables biológicas. El trabajo ha consistido en someter a un nuevo estudio mediante técnicas de analítica predictiva, el pronóstico de respuesta al tratamiento de 263 pacientes que habían sido diagnosticados de LH, utilizando ahora un gran conjunto de datos de valoración de progresión y de variables biológicas de pronóstico de respuesta, con el fin de descubrir cuáles de estas variables determinarán con mayor probabilidad qué paciente iba a responder y quién no. Mediante la aplicación de algoritmos de aprendizaje automático y de optimización global cooperativa se logró un modelo predictivo que seleccionaba, entre las 35 variables evaluadas, únicamente a tres que eran altamente discriminatorias de progresión o de remisión de la enfermedad: la ferritina sérica, la alina transaminasa, y la fosfatasa alcalina. Con ello se simplificaba considerablemente el estudio para conocer anticipadamente si el tratamiento iba a ser efectivo, al tiempo que se mantenía el reto de mejorar el tratamiento para el grupo de no respondedores.

6. CONCLUSIONES

En el transcurso y desarrollo de este trabajo se pretendió entregar una visión de la tecnología que está avanzando a pasos agigantados y que notros como un área de la salud nacional, no estamos al costado ni mucho menos ajenos a los datos que esta misma nos entrega, me refiero al mundo del Big Data la que surge como respuesta a la creciente necesidad de un número cada vez mayor de organizaciones de disponer de nuevas herramientas capaces de procesar de forma eficiente un enorme volumen de datos de diversa tipología procedentes de fuentes muy heterogéneas. A diferencia de otras soluciones tradicionales, Big Data permite gestionar grandes volúmenes de datos de todo tipo, especialmente no estructurados, a una velocidad muy superior y con un consumo de recursos mucho menor. Otra de las grandes ventajas que ofrece esta tecnología es la posibilidad de utilizar técnicas de analítica avanzada, como la analítica predictiva, con el objetivo de operar sobre datos masivos y construir modelos predictivos de calidad que sirvan de soporte a la toma de decisiones.

En Chile se avanzado en el proyecto del gobierno citado para el 2020, que dice relación al plan de construir un repositorio nacional de datos en salud, que permitirá llevar a cabo procesos analíticos a nivel país, permitiendo la toma de decisiones a nivel local y central en materias de gestión, políticas públicas de salud y en definitiva, un mejor entendimiento de la salud de la ciudadanía y de cómo las prestaciones que reciben inciden en ésta, por lo menos en el sector público, se visualiza con éxito lo obrado en materia de implementación y adopción de la Ficha Clínica Electrónica, como núcleo de una serie de sistemas que han permitido la optimización de la gestión, el uso de la información clínica, el apoyo al diagnóstico y mejor atención de los pacientes, como base del aporte de las TI a las Redes Integradas de Servicios de Salud.

Gracias al avance tecnológico, el concepto de Big Data está siendo relevante en muchas industrias, en salud, el concepto de Big Data es un fenómeno nuevo y en algunos países como Alemania, Dinamarca y Estados Unidos es una realidad y el registro clínico electrónico es utilizado para el uso secundario de los datos de la ficha médica para predecir riesgos a nivel individual, con el objeto de tomar decisiones con una mirada prospectiva y establecer políticas sanitarias basadas en evidencia.

Para que la propuesta planteada consolide los resultados esperados, es necesario alcanzar un alto nivel de compromiso de todas aquellas áreas, afrontar con solvencia la gestión del cambio, es decir, que se logré un trabajo conjunto y una evaluación constante de todos los procesos que se deban diseñar. También es importante monitorear constantemente el desempeño de los indicadores que se deben implementar en el sistema informático, ya que, gracias a esto, se puede cuantificar y diagnosticar más fácilmente las posibles áreas de mejora dentro de los procesos.

La tecnología está disponible y profundizar los excelentes resultados de la década en materia de TI en Salud es posible, generando un trabajo sinérgico entre los distintos actores involucrados, desde el mundo público, privado y de la Academia, todo en post y beneficios de los principales actores, nosotros mismo los pacientes.

7. GLOSARIO

- ***Analítica de Datos***

  Se conoce como Analítica de Datos, o Data Analytics o Data & Analytics a todas aquellas tareas orientadas a la exploración de los datos, con la intención de encontrar patrones o conocimiento útil, que permita optimizar o rentabilizar un proceso de negocio

- **Balanced ScoreCard**

  El Balanced ScoreCard (BSC / Cuadro de Mando Integral) es una herramienta que permite enlazar estrategias y objetivos clave con desempeño y resultados a través de cuatro áreas críticas en cualquier empresa: desempeño financiero, conocimiento del cliente, procesos internos de negocio y aprendizaje y crecimiento

- **Big Data**

  Big data (en español, grandes datos o grandes volúmenes de datos) es un término evolutivo que describe cualquier cantidad voluminosa de datos estructurados, semiestructurados y no estructurados que tienen el potencial de ser extraídos para obtener información.

- **Biomédica**

  La ingeniería biomédica es el resultado de la aplicación de los principios y técnicas de la ingeniería al campo de la medicina. Se dedica fundamentalmente al diseño y construcción de productos sanitarios y tecnologías sanitarias.

- **Business Intelligence**

  La inteligencia de negocios o Business Intelligence (BI) es el conjunto de procesos, aplicaciones y tecnologías que facilitan la obtención rápida y sencilla de datos provenientes de los sistemas de gestión empresarial para su análisis e interpretación, de manera que puedan ser aprovechados para la toma de decisiones

- **Cloud**

  Conocida también como servicios en la nube, informática en la nube, nube de cómputo, nube de conceptos o simplemente "la nube", es un paradigma que permite ofrecer servicios de computación a través de una red, que usualmente es Internet.

- **CSV**

  Es un formato de archivo de texto que se puede utilizar para intercambiar datos de una hoja de cálculo entre aplicaciones. Cada línea de un archivo CSV de texto representa una fila de una hoja de cálculo.

- **Data mart**

  Una data mart es una versión especial de almacén de datos. Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones.

- **Data Mining**

  La minería de datos es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen entre ambas.

- **Data warehouses**

  Un Data Warehouse es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta.

- **Dimensión de datos**

  Una dimensión representa un solo conjunto de objetos o sucesos del mundo real. Cada dimensión que identifique para el modelo de datos se implanta como una tabla de dimensiones. Las dimensiones son los calificadores que dan sentido a la tabla de hechos.

- **ETL**

  Extract, Transform and Load, ETL es el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un almacén de datos, reformatearlos, limpiarlos y cargarlos en otra base de datos, data mart o bodega de datos.

- **Genómica**

  Se denomina genómica al conjunto de ciencias y técnicas dedicadas al estudio integral del funcionamiento, el contenido, la evolución y el origen de los genomas. Es una de las áreas más vanguardistas de la Biología.

- **Genoma**

  El genoma es el conjunto de genes contenidos en los cromosomas, lo que puede interpretarse como la totalidad del material genético que posee un organismo o una especie en particular.

- **GUI (Graphical User Interface)**

  Es un tipo de interfaz de usuario que utiliza un conjunto de imágenes y objetos gráficos para representar la información y acciones disponibles en la interfaz. Habitualmente las acciones se realizan mediante manipulación directa para facilitar la interacción del usuario con la computadora.

- **Hadoop**

  Es un Framework de software que soporta aplicaciones distribuidas bajo una licencia libre. Permite a las aplicaciones trabajar con miles de nodos y petabytes de datos. Hadoop se inspiró en los documentos Google para MapReduce y Google File System (GFS). Hadoop es un proyecto de alto nivel.

- **HBase**

  Es una base de datos distribuida no relacional de código abierto modelada a partir de Google Big Table y escrita en Java.

- **HDFS**

  es un sistema de ficheros distribuido, escalable y portátil escrito en Java y creado especialmente para trabajar con ficheros de gran tamaño. Una de sus principales características es un tamaño de bloque muy superior al habitual (64 MB) para no perder tiempo en los accesos de lectura.

- **HTML**

  HTML, sigla en inglés de HyperText Markup Language (lenguaje de marcas de hipertexto), hace referencia al lenguaje de marcado para la elaboración de páginas web. Es un estándar que sirve de referencia del software que conecta con la elaboración de páginas web en sus diferentes versiones.

- **Know-how**

  Es un neologismo anglosajón que hace referencia a una forma de transferencia de tecnología.

- **Machine learning**

  El aprendizaje automático o aprendizaje de máquinas es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender.

- **MDM (Master Data Management)**

  MDM (gestión de datos maestros), es un método que permite a una organización relacionar todos sus datos críticos con un solo archivo llamado archivo maestro, de forma que se obtiene un punto de referencia común para los datos más importantes.

- **Modelado predictivo**

  Los modelos predictivos a menudo se desarrollan utilizando un proceso llamado "aprendizaje supervisado", en el que se selecciona un conjunto de resultados predeterminados, se identifican variables que pueden contribuir a predecirlos y se aplican algoritmos de análisis estadístico a un conjunto de datos de prueba, son la clave para poder, mediante un esfuerzo analítico, detectar oportunidades de inversión, conocer la previsión de ventas o la cuota de mercado, identificar los segmentos de consumidores más rentables o los mercados de destino con mayor potencial.

- **NLP (procesamiento de lenguaje natural)**

  Es el campo que combina las tecnologías de la ciencia computacional (como la inteligencia artificial, el aprendizaje automático o la inferencia estadística) con la lingüística aplicada, con el objetivo de hacer posible la comprensión y el procesamiento asistidos por ordenador.

- **NoSQL**

  Es un término que describe las bases de datos no relacionales de alto desempeño. Las bases de datos NoSQL utilizan varios modelos de datos, incluidos los de documentos, gráficos, claves-valores y columnas. Las bases de datos NoSQL son famosas por la facilidad de desarrollo, el desempeño escalable, la alta disponibilidad y la resiliencia.

- **OLAP**

  Es el acrónimo en inglés de procesamiento analítico en línea. Es una solución utilizada en el campo de la llamada Inteligencia de negocios cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

- **Open source**

  Es un modelo de desarrollo de software basado en la colaboración abierta. Se enfoca más en los beneficios prácticos que en cuestiones éticas o de libertad que tanto se destacan en el software libre.

- **Petabytes**

  Un petabyte es una unidad de almacenamiento de información cuyo símbolo es PB, y equivale a 10^{15} bytes = 1 000 000 000 000 000 de bytes.

- **SQL**

  Es un lenguaje específico del dominio que da acceso a un sistema de gestión de bases de datos relacionales que permite especificar diversos tipos de operaciones en ellos.

- **Streaming**

  La retransmisión es la distribución digital de contenido multimedia a través de una red de computadoras, de manera que el usuario utiliza el producto a la vez que se descarga.

- **Tele asistencia**

  Es un servicio de asistencia 24 horas que permite a las personas estar permanentemente conectadas con un equipo de profesionales socio sanitarios y recibir, con sólo pulsar un botón, ayuda inmediata en situaciones de emergencia o inseguridad, y atención continua para su día a día.

- **Terabytes**

  Un Terabyte es una unidad de medida informática cuyo símbolo es el TB, y es equivalente a 2 a los 40 bytes. Comúnmente se acepta que un terabyte es equivalente a 1,000gb lo cual implica 1,000,000 de Mb o 1,000,000,000 de Kb.

- **TI (Tecnología de la Información)**

  La tecnología de la información es la aplicación de ordenadores y equipos de telecomunicación para almacenar, recuperar, transmitir y manipular datos, con frecuencia utilizado en el contexto de los negocios u otras empresas.

- **TIC**

  El término Tecnologías de Información y Comunicación (TIC) tiene dos acepciones. Por un lado, a menudo, se usa el término 'tecnologías de la información' para referirse a cualquier forma de hacer cómputo; por el otro, como nombre de un programa de licenciatura que se refiere a la preparación que tienen estudiantes

- **Tuplas**

  La tupla es un tipo de dato secuencial. Sirve para agrupar, como si fueran un único valor, varios valores. El tipo de datos que representa a las tuplas se llama tuple, y es inmutable: una tupla no puede ser modificada una vez que ha sido creada.

- **VPN**

  Una red privada virtual, en inglés: Virtual Private Network es una tecnología de red de computadoras que permite una extensión segura de la red de área local sobre una red pública o no controlada como Internet.

- **Watson**

  Watson es un sistema informático de inteligencia artificial que es capaz de responder a preguntas formuladas en lenguaje natural, desarrollado por la corporación estadounidense IBM.

- **Wearables**

  Wearable hace referencia al conjunto de aparatos y dispositivos electrónicos que se incorporan en alguna parte de nuestro cuerpo interactuando de forma continua con el usuario y con otros dispositivos con la finalidad de realizar alguna función concreta, relojes inteligentes o smartwatchs, zapatillas de deportes con GPS incorporado y pulseras que controlan nuestro estado de salud.

- **XML**

  El formato estándar “Extensible Markup Language (XML), tiene varias características que lo hacen conveniente, entre las que podemos destacar: Es un estándar abierto, flexible y ampliamente utilizado para almacenar, publicar e intercambiar cualquier tipo de información.

8. BIBLIOGRAFIAS

- Barranco Fragoso, R. (18 de Junio de 2012). *IBM DeveloperWorks*. Obtenido de ¿Que es Big Data?: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

- Cossio Lopez, H. (05 de Enero de 2017). *El Mostrador*. Obtenido de La era cognitiva de Watson: la frontera con las máquinas que la humanidad ya cruzó.: http://www.elmostrador.cl/cultura/2017/01/05/la-era-cognitiva-de-watson-la-frontera-con-las-maquinas-que-la-humanidad-ya-cruzo/

- de Saint Pierre, D., Borgoño, V., Fuentes, L., Mauro, A., Lan Sung Hsieh, H., Fernández, M., & Holuigue, C. (Abril de 2013). *www.minsal.cl.* Obtenido de MAPA DE RUTA, Plan Estratégico de Tecnologías de Información (e-Salud) 2011-2020: http://www.salud-e.cl/wp-content/uploads/2013/08/Mapa-de-ruta-completo.pdf

- Feldman, B., M. Martin, E., & Skotnes , T. (Octubre de 2012). *drbonnie360.com.* Obtenido de Big Data in Healthcare Hype and Hope: https://www.west-info.eu/files/big-data-in-healthcare.pdf

- Heudecker, N., & Kart, L. (03 de Septiembre de 2015). *Gartner*. Obtenido de Survey Analysis: Practical Challenges Mount as Big Data Moves to Mainstream: https://www.gartner.com/doc/3123817/survey-analysis-practical-challenges-mount

- Jamack, P. (04 de Febrero de 2013). *IBM DeveloperWoks*. Obtenido de Analítica de inteligencia de negocios de big data: https://www.ibm.com/developerworks/ssa/library/ba-big-data-bi/

- Joyanes Aguilar, L. (2014). *Big Data: Analisis de grandes volumenes de datos en las organizaciones.* Barcelona: Marcombo.

- Laney, D. (14 de Enero de 2012). *Gartner Blog Network*. Obtenido de Gartner Blog Network: https://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/

- Mayer-Schonberger, V., & Neil Cukier, K. (2013). *Big data. La revolución de los datos masivos.* Madrid: Turner.

- *McKinsey Global Institute.* (Junio de 2011). Obtenido de Big data: The next frontier, for innovation, competition and productivity.: https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi_big_data_full_report.ashx

- Rodriguez, J. (10 de Octubre de 2012). *Informatica++.* Obtenido de Octubre, el mes de los grandes datos: http://informatica.blogs.uoc.edu/2012/10/10/octubre-el-mes-de-los-grandes-datos-2/

- Sanchez, J. (01 de Septiembre de 2014). *Accenture.* Obtenido de Las empresas consideran Big Data fundamental para su transformación digital: https://www.accenture.com/es-es/company-big-data-fundamental-transformacion-digital
- Schroeck, M., Shockle, R., Smart, D., Romero-Morales, P., & Tufano, P. (2012). *Analytics: el uso de big data en el mundo real, Cómo las empresas más innovadoras extraen valor de datos inciertos.* Madrid: IBM Institute for Business Val222ue y Escuela de Negocios Saïd en la Universidad de Oxford.

- Vazquez, D. (01 de Septiembre de 2017). *America Retail*. Obtenido de El impacto de las redes sociales en el sector salud de Latinoamérica: http://www.america-retail.com/marketing-digital/marketing-digital-el-impacto-de-las-redes-sociales-en-el-sector-salud-de-latinoamerica/

- Zaforas, M. (1 de Septiembre de 2016). *Paradigma Digital*. Obtenido de Paradigma Digital: https://www.paradigmadigital.com/dev/puede-aportar-big-data-al-mundo-la-medicina/#comments