UNIVERSIDAD TÉCNICA DE MACHALA

FACULTAD DE INGENEIRIA CIVIL

MAESTRIA EN SOFTWARE

TEMA: BIG DATA ANALYTICS EN LA GESTIÓN HOSPITALARIA PARA EL CUMPLIMIENTO DE INDICADORES DE GOBIERNO POR RESULTADOS

MODALIDAD:PROPUESTAS METODOLÓGICAS Y TECNOLOGÍAS AVANZADAS

Autor: BYRON JAVIER MOTOCHE MEDINA

Tutora: BERTHA EUGENIA MAZON OLIVO

Co-Tutor: EDUARDO ALEJANDRO TUSA JUMBO

MACHALA, 2021

PENSAMIENTO

“La información es la gasolina del siglo XXI, y la analítica de datos el motor de combustión”. Peter Sondergaard.

DEDICATORIA

A Dios, mi amada esposa por su apoyo constante que me brinda para alcanzar nuevos objetivos, profesionales y personales, mis padres y mis hermanas que me han guiado desde la infancia y siempre han sido fuente de mi inspiración.

A todas las personas que me han brindado su apoyado y han hecho que este trabajo de tesis se realice con éxito, en consideración a aquellos que me abrieron las puertas y compartieron sus conocimientos.

AGRADECIMIENTOS

Le agradezco a Dios por haberme concedido vivir hasta este día, haberme tutelado a lo largo de mi vida, por ser mi apoyo, mi luz y mi camino. Por haberme dado la fuerza para seguir adelante en aquellos momentos de debilidad. Le doy gracias a mi familia quienes han estado constantemente en los momentos que más he necesitado, y a los tutores quienes con su sabiduría me han sabido orientar en el desarrollo de este trabajo.

RESPONSABILIDAD DE AUTORÍA

Por medio de la presente declaro ante el comité Académico de la Maestría en Software de la universidad Técnica de Machala. Que el trabajo de titulación titulado "**BIG DATA ANALYTICS EN LA GESTIÓN HOSPITLARIA PARA EL CUMPLIMIENTO DE INDICADORES DE GOBIERNO POR RESULTADOS**", de mi propia autoría, no contiene material escrito por otra persona al no ser referenciado debidamente en el texto, parte de ella o en su totalidad no ha sido aceptada para el otorgamiento de cualquier otro diploma de una institución nacional o extrajera.

Firmado electrónicamente por:
BYRON JAVIER MOTOCHE MEDINA

Byron Javier Motoche Medina

C.I. 0703310755

Machala, 2021/03/12

REPORTE DE SIMILITUD TURNITI

BIG DATA ANALYTICS EN LA GESTIÓN HOSPITALARIA PARA EL CUMPLIMIENTO DE INDICADORES DE GOBIERNO POR RESULTADOS

INFORME DE ORIGINALIDAD

3%	2%	1%	0%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	Eunjoo Kim, Young-Bae Park, Kahye Choi, Young-Woo Lim et al. "Application of Machine Learning to Predict Weight Loss in Overweight, and Obese Patients on Korean Medicine Weight Management Program", Journal of Korean Medicine, 2020 Publicación	1%
2	mailweb.udlap.mx Fuente de Internet	1%
3	doku.pub Fuente de Internet	<1%
4	ingenieria.uaslp.mx Fuente de Internet	<1%
5	Heung-gu Son, Yunsun Kim, Sahm Kim. "Time Series Clustering of Electricity Demand for Industrial Areas on Smart Grid", Energies, 2020 Publicación	<1%

CERTIFICADO DEL TUTOR

Por medio de la presente apruebo que el trabajo de titulación, titulado **"BIG DATA ANALYTICS EN LA GESTIÓN HOSPITALARIA PARA EL CUMPLIMIENTO DE INDICADORES DE GOBIERNO POR RESULTADOS"**, del autor Byron Javier Motoche Medina, en opción al título de Master en Software, se ha presentado al Acto de Defensa.

Firmado electrónicamente por:
BERTHA EUGENIA MAZON OLIVO

Ing. Bertha Mazón Olivo, Mgs

C.I. 0603100512

Machala, 2021/03/12

CESIÓN DE DERECHOS DE AUTORÍA

Yo, Byron Javier Motoche Medina, en calidad de autor del presente trabajo titulado **"BIG DATA ANALYTICS EN LA GESTIÓN HOSPITALARIA PARA EL CUMPLIMIENTO DE INDICADORES DE GOBIERNO POR RESULTADOS"**, Autorizo a la UNIVERSIDAD TÉCNICA DE MACHALA la publicación y distribución en el Repositorio Digital Institucional.

El autor declara que el contenido que se publicará es de carácter académico y se enmarca en las disposiciones definidas por la Universidad Técnica de Machala.

Firmado electrónicamente por:
BYRON JAVIER MOTOCHE MEDINA

Byron Javier Motoche Medina

C.I. 0703310755

Machala, 2021/03/12

RESUMEN

El presente trabajo comprende el desarrollo de una propuesta enfocada a la implementación y buen uso de Big Data Analytics (BDA), con el fin de obtener mejores resultados en el procesamiento de información y toma de decisiones, tomando en consideración el uso de técnicas de análisis de datos, que proveen una visualización más descriptiva en los resultados, para una mejor comprensión y guía en la toma de decisiones. Además de destaca su utilidad para la presentación de resultados a las entidades gubernamentales que lo soliciten. Este proyecto se aplicó en un Hospital General de carácter público, por lo que se estableció un plan de acción para la ejecución de análisis de datos mediante la metodología CRISP-DM. Se realizó el proceso de Extracción Transformación y Carga (ETL), para cargar variables e indicadores de gestión hospitalaria en una arquitectura de Big Data.

Se aplicó la técnica de series temporales mediante cuatro modelos predictivos: NNETAR (Pronósticos de Series de Tiempo de Redes Neuronales), STLM (Descomposición estacional y de tendencias usando Loess con múltiples períodos estacionales), Holt-Winters y TBATS (Estacionalidad trigonométrica, Transformación Box-Cox, Errores ARMA, Componentes de tendencia y estacionales), para conocer qué sucedería en un futuro, tomando en consideración un período de 12 meses. Se trabajó con datos comprendidos desde enero 2014 a octubre 2020. Los modelos fueron comparados mediante las métricas RMSL y MAE con el fin de determinar cuál es el más aceptable para su uso. Se concluye que los modelos NNETAR y TBATS por sus resultados más cercanos a los reales, son los más adecuados.

PALABRAS CLAVES: BIG DATA, HADOOP, HOSPITAL, SALUD

ABSTRACT

This work includes the development of a proposal focused on the implementation and good use of Big Data Analytics (BDA), in order to obtain better results in information processing and decision-making, taking into consideration the use of techniques of data analysis that provide a more descriptive visualization of the results for a better understanding and guide for decision-making, in addition to highlighting its usefulness for the presentation of results to government entities that request it. This project was applied in a public Level II General Hospital, for which an action plan was established for the execution of data analysis using the CRISP-DM methodology. In addition, the Extraction Transformation and Loading (ETL) process was carried out, to load variables and hospital management indicators in a Big Data Warehouse.

The time series technique was applied using four predictive models: NNETAR (Neural Network Time Series Forecasts), STLM (Seasonal and trend decomposition using Loess with multiple seasonal periods), Holt-Winters and TBATS (Trigonometric seasonality, Box transformation -Cox, ARMA Errors, Trend and Seasonal Components), to know what would happen in the future, taking into consideration a period of 12 months. We worked with data from January 2014 to October 2020. The models were compared using the RMSL and MAE metrics in order to determine which is the most acceptable for use. It is concluded that the NNETAR and TBATS models, due to their results closer to the real ones, are the most appropriate.

KEY WORDS: BIG DATA, HADOOP, HOSPITAL, HEALTH

ÍNDICE GENERAL

LISTA DE ILUSTRACIONES Y TABLAS

ILUSTRACIONES

TABLAS

LISTA DE ABREVIATURAS Y SÍMBOLOS

HDFS.- Sistema de archivos distribuidos de Hadoop

RDMS.- Sistema de gestión de base de datos relacionales

BDA.- Big Data Analytics

ETL.- Extracción, Transformación y Carga

IGH.- Indicador de Gestión Hospitalaria

CRISP-DM.- Cross Industry Standard Process for Data Mining

SQL.- Structured Query Language

GPR.- Gobierno por Resultados

DLL.- Dynamic Link Library

API.- Interfaz de Programación de Aplicaciones

GLOSARIO

PREDICCIÓN: En un hecho que posiblemente ocurra en el futuro.

NODO: En términos de informática refiere, a la representación de un equipo cómputo o hardware.

CODIGO ABIERTO: En términos de software, significa que la aplicación o software es de uso libre, es decir no hay que pagar por la misma.

FRAMEWORK: Es la arquitectura en la que un sistema puede trabajar, en base a todos sus componentes, ya sean DLL, API.

PROCESAMIENTO DISTRIBUIDO: Es el conjunto de nodo que pueden procesar grandes cantidades de datos de forma sincronizada.

DLL: Es un archivo que contiene código y datos que pueden ser usados en otros programas.

API: Es un conjunto de definiciones y protocolos que se usan para desarrollar e integrar un software. Permite la comunicación entre varios dispositivos sin la necesidad de saber cómo fueron creados.

INTRODUCCIÓN

El objeto de este trabajo se delimita al estudio y/o recopilación de datos de varias fuentes bibliográficas científicas, que permitan dar a conocer de mejor forma, el criterio más adecuado en defensa a los beneficios que otorga el uso de la Big Data Analytics dentro del área hospitalaria. Más aún, teniendo en cuenta la problemática en el manejo y disposición de los datos para la toma de decisiones, creando un lazo entre lo que se requiere y lo que se espera por el personal que desea consumir la información. Además, se reconoce ofrecer un criterio de los beneficios de la Big Data y la mejor forma de aplicarlo dentro de un Hospital General.

Para el desarrollo de este trabajo, se toma en consideración al sector de la Salud, debido a que en la actualidad se ha visto expuesto a un gran tsunami de datos alojados en diferentes fuentes de información. Esto ha dado lugar a repositorios de archivos voluminosos que resultan difíciles de interpretar, comprender y aplicar para una respuesta rápida y oportuna en la toma de decisiones. Los beneficios y bondades que otorga las técnicas de análisis de datos en la gestión hospitalaria demandan una mayor atención en todas las áreas involucradas. Por ejemplo, entre los indicadores de gestión hospitalaria, se puede mencionar: tasa de mortalidad hospitalaria, porcentaje hospitalario de mortalidad neonatal, tasa de reingreso de pacientes. El presente trabajo plantea la siguiente pregunta de investigación: ¿Cómo analizar indicadores de gestión hospitalaria (IGH) que involucran grandes volúmenes de datos?

En pleno siglo XXI la ciencia de los datos o también llamado Big Data Analytics (BDA), está tomando mayor fuerza junto de la mano con las tecnologías de la información y comunicación. Esto ha aperturado en gran medida al desarrollo de las organizaciones e instituciones en diferentes ámbitos tanto públicas como privadas. Uno de los mayores aspectos a considerar de la Big Data Analytics son los aportes que provee a las organizaciones y esto apuntando al tratamiento de la información, que sin lugar a duda ha llevado a un nivel más alto la forma de poder darle significado a los datos para tomar una decisión [1]. Con la llegada de la industria 4.0 algunas empresas se han visto obligadas a cambiar sus procesos y formas de hacer las cosas en beneficio a un fin común, que es la competencia local y el conocimiento del qué hacer antes de tomar una decisión, es aquí donde entra Big Data Analytics, como ciencia icónica en la revolución de los datos [2] [3].

Ciertos países del primer mundo, están desarrollando técnicas y procedimientos que permitan de cierta forma, otorgar un mayor significado a los datos considerando que los datos sin una estructura adecuada no tienen sentido para su análisis y posterior entendimiento para tomar una decisión [4]. En América Latina la revolución de la industria ha llevado a que la información esté disponible, en circunstancias adecuadas que permitan el mejor desarrollo de las organizaciones; países como México, Brasil, Perú y entre otros han dado hincapié a la importancia de compactar y obtener mejores resultados mediante el análisis de la información con herramientas sofisticadas que dan mayor facilidad al uso de Big Data Analytics, como técnica para el análisis de la información [5].

Cabe destacar que la industria de la Salud no ha asimilado de forma rápida los cambios o beneficios que ha conllevado el uso de Big Data Analytics. Por lo tanto, se puede indicar que aún sigue en desarrollo. Posiblemente, la falta de interés se dé por motivos de información escasa, profesionales en el ámbito de la ciencia de datos, o recursos que más bien se destinan para otros proyectos [6].

El Gobierno Nacional del Ecuador ha implementado el sistema GPR (Gobierno por resultados) que su función principal es obtener datos relevantes de diferentes instituciones públicas del país, sin embargo, en los hospitales generales no existe un Sistema Integral Hospitalario que permita agrupar toda esa información y poder analizarla de forma rápida y oportuna, ya que se distribuye en documentos lógicos (Word, Excel), variedad de bases de datos (SQL Server, MySql) y físicos (matrices, escritos, informes), que se aglomeran con el pasar del tiempo y quedan en el olvido sin tener una constancia de una base de datos centralizada y de fácil acceso cuando la entidad hospitalaria pública lo requiera. Así mismo, se logre dar opción a la Dirección Médica, una fuente de datos para la toma de decisiones importantes para el bien común dentro de la entidad de salud [7].

OBJETIVO GENERAL Y ESPECÍFICOS

El objetivo general de esta investigación se enfoca en analizar indicadores de gestión hospitalaria (IGH) mediante técnicas que soporten grandes volúmenes de datos para el soporte de toma de decisiones institucionales.

Para el cumplimiento del objetivo general, se desarrollará los siguientes objetivos específicos:

1. Recopilar información científica acerca de Big Data Analytics y sus beneficios en el área de la Salud.
2. Analizar los requerimientos y fuentes de datos relacionados con los IGH.
3. Seleccionar la metodología y técnicas de BDA acorde a los IGH.
4. Implementar las técnicas de análisis, aplicados a los IGH, que soporten grandes volúmenes de datos.
5. Evaluar los modelos de análisis de datos mediante métricas como la exactitud o precisión.

ANTECEDENTES

Desde la antigüedad, la civilización ha estado inmersa en conocer el futuro o más bien que decisión tomar antes de ejecutar un plan de acción. En ello, están los mayas, egipcios y otras sociedades que han visto la utilidad de recopilar información desde cualquier fuente con el fin de conocer qué hacer. En la actualidad se puede decir que esa fuente se la denomina Big Data, que se la define con una ola de información que se puede explotar para darle una mejor forma y entender que es lo que nos quiere decir. Ecuador un país en vías de desarrollo, está accediendo a esta etapa en cuanto se refiere a la estructuración de la información. No obstante, se ven limitadas las empresas u organizaciones en invertir para poder obtener los beneficios del análisis de la Big Data [8].

Los hospitales públicos en el Ecuador se han visto envueltos en que la inversión tecnológica es medianamente baja, dando así lugar a un déficit en la forma en cómo se maneja la información de una manera centralizada y única. Esto motiva a recurrir a diferentes fuentes de información informales, que generan retrasos y falta de buena atención al cliente [9]. Se contempla mayormente esto, a que no ha existe un plan de mejoramiento o estudio que permita dar a conocer la importancia que es implementar un proyecto de análisis de información lo que muestra un abanico de posibilidades, mejoras y ahorros que conlleva hacerlo [10].

El desarrollo del proyecto se dará lugar en la provincia del Oro, ciudad Machala, específicamente en un Hospital General. Esta es una entidad médica que brinda servicios de salud pública. La situación actual de la entidad es normalmente eficiente, sin embargo, se encuentra dentro de un marco tecnológico medianamente escaso. Esto será punto de partida como análisis para establecer las debilidades del mismo y las ventajas que conlleva el estudio de este proyecto. Así mismo se ofrece soluciones óptimas a costos bajos que permitan dar impulso a los objetivos planteados en el desarrollo de esta tesis.

HIPÓTESIS

Para este proyecto se plantea la siguiente hipótesis:

- La aplicación de Big Data Analytics en IGH, proporciona soporte en la toma de decisiones en un Hospital General. La tabla 1 describe la definición conceptual, mientras que la tabla 2, detalla la definición operacional de las variables. (Ver Tabla 1-2).

Tabla 1.- Conceptualización de la hipótesis

Variables	Conceptos
Variable Independiente: Big Data Analytics en IGH	Big Data Analytics en IGH hace referencia al análisis de gran cantidad y diversidad de datos de los sistemas hospitalarios difíciles de ser tratados con métodos tradicionales.
Variable Dependiente: Soporte en la toma de decisiones en un Hospital General.	El soporte en la toma de decisiones en Hospital General se refiere a la Implementación y evaluación de técnicas, modelos de análisis, que soporten grandes volúmenes de datos en IGH.

Fuente: Autor

Tabla 2.- Operacionalización de las variables

Variables	Categorías	Indicadores	Técnicas
Variable Independiente Big Data Analytics en IGH.	1. Integración 2. Plataformas de Soporte. 3. Presentación de la Información.	1. Diseño de la arquitectura de Big Data. 2. Desarrollo del prototipo escalable. 3. Distribución del trabajo en nodos independiente y tolerante a fallos. 4. Localidad de	1. Recopilación de información mediante las técnicas RSL. 2. Extracción de información para la estructuración de la Big Data. 3. Análisis comparativo de

		los datos en cada nodo participante. 5. Aplicación de las técnicas ETL (Extracción, Transformación y Carga). 6. Recopilación de datos en fuentes confiables y verídicas, de acceso comprobable. 7. Uso de software especializado Hadoop y sus componentes. 8. Lenguaje de programación de alto nivel Python y R. 9. Herramientas web para la visualización web en Php y HighCharts. 10. Modelos de análisis de datos, series temporales, redes neuronales y predictivos.	modelos. 4. ETL. 5. Técnicas de análisis de datos
Variable Dependiente Soporte en la toma de decisiones en un Hospital General de nivel II	Satisfacción de usuario Eficacia	1. Visualización de indicadores en tiempo real. 2. Datos para la toma de decisiones 3. Búsqueda de información real y validada. 4. Visualización de pronósticos. 5. Disminución de materiales y costos.	1. Recopilación de información. 2. Análisis de modelos de datos.

Fuente: Autor

MÉTODOS DE INVESTIGACIÓN

Los métodos de investigación a emplearse en el presente estudio son los siguientes:

- Cualitativo, ya que se requiere obtener información de varias fuentes científicas, comprender la situación actual y ofrecer los mejores criterios de la investigación, además de ello utilizar técnicas de recopilación de datos.
- Basado en objetivos, esta metodología permite cumplir a cabalidad todos los objetivos planteados en el presente proyecto, con el fin de otorgar y mostrar su beneficio y viabilidad.
- Correlacional, se dará lugar a la relación entre nuestra variable de estudio y los resultados esperados, poniendo en evidencia los beneficios que otorga los lineamientos del proyecto y la solución al problema.

Una vez concluido el desarrollo del proyecto, se espera haber abarcado de forma completa la solución al problema actual, ofreciendo la posibilidad en que esta investigación pueda ser mayormente ampliada para su implementación en otras entidades públicas y sea referente de investigaciones futuras.

Este trabajo ha sido estructurado de la siguiente forma:

- Capítulo 1. Presentación y desarrollo del marco teórico, conceptual y contextual de la investigación propuesta.
- Capítulo 2. Metodologías y estrategias en la recopilación de datos.
- Capítulo 3. Se muestra la propuesta, el desarrollo de la misma, y los resultados obtenidos.
- Capítulo 4. Detalla la discusión de los resultados obtenidos, mejoras y alcances. Además de las conclusiones y recomendaciones derivadas del estudio.

CAPÍTULO 1. MARCO TEÓRICO REFERENCIAL

1.1 ANTECENDENTE HISTÓRICO DE LA INVESTIGACIÓN

1.1.1 METODOLOGÍA APLICADA: REVISIÓN SISTEMÁTICA DE LA LITERATURA (RSL)

Este apartado considera exclusivamente la técnica utilizada para la recopilación, análisis y comprensión de los temas investigados, teniendo en cuenta los procedimientos a seguir según lo establecido por Barbara Kitchenham en cuanto a la revisión sistemática de la literatura en la ingeniera de software [11]. En su publicación destaca lo siguiente:

- Un RSL es un medio para evaluar y analizar todas las fuentes relevantes en cuanto a la pregunta de investigación.
- Un RSL tiene como objetivo presentar una evaluación justa utilizando metodologías confiables y auditables.
- Una investigación siempre debe de seguir los parámetros o estrategia de búsqueda predefinida tomando en consideración el respaldo de la hipótesis planteada.

1.1.1.1 DESARROLLO DE LA INVESTIGACIÓN

Tomando en consideración el contexto de la investigación dentro del campo de la ingeniería de software para la Big Data Analytics, se puede destacar gran material de investigación y aplicación [12-16]. También se puede observar en el área de Salud [17-20]. Existen conceptos fundamentales para la comprensión del tema a abordar en cuanto a estructuras, framework, arquitecturas de programación [21-24], además de procesos de extracción, transformación y carga de datos [25-26]. Al mismo tiempo se consideraron otros textos que son relevantes dentro del estudio de investigación.

1.1.1.2 TRABAJOS RELACIONADOS

Se abordó investigar las conclusiones y resultados de otros trabajos realizados por autores que vieron una oportunidad de implementación de Big Data Analytics en la Salud y otras áreas, donde destacan su utilidad, ventajas y los avances de la tecnología [13,18,27-29]. Este es el punto de partida para contrastar la hipótesis, más aún determinar la viabilidad de la investigación aplicarse en el Hospital General.

1.1.1.3 PREGUNTA DE INVESTIGACIÓN

Se establecieron las siguientes preguntas para la búsqueda de información acerca de Big Data Analytics y su aplicación en el área de Salud, además de sus herramientas y tecnología existente. La Tabla 3 detalla lo indicado:

Tabla 3. Preguntas de investigación RSL

Preguntas	Dimensiones
1- ¿Dentro del ámbito tecnológico, que beneficios ha traído la implementación de Big Data a las organizaciones?	Diseños, modelos de implementación, resultados ya evaluados, conceptos.
2.- ¿Cuáles son las herramientas o técnicas a usar para una buena implementación de Big Data?	Software, framework, técnicas ETL, buenas prácticas de implementación y diseño.
3.- ¿Cuál es el impacto en el área de Salud y que beneficios ha otorgado en otros países?	Experiencias y beneficios en el área de Salud.

Fuente: Autor

1.1.1.4 PROCESO DE BÚSQUEDA

Para el proceso de búsqueda se tomó en consideración varias bases de datos de referencias bibliográficas y citas, tales como:

- Scopus
- Dialnet
- WoS
- IEEE
- Springer
- Scielo
- SciencieDirect

1.1.1.5 CRITERIOS DE INCLUSIÓN Y EXCLUSIÓN

Para los criterios de inclusión se tomó en consideración, artículos y revistas, documentos que detallen la importancia de Big Data Analytics, sus beneficios, su impacto en la actualidad, y los cambios en el campo médico. Además, se circunscriben modelos de implementación, desarrollos y técnicas con resultados positivos, que puedan ser aplicados y evaluados en el desarrollo del presente trabajo. También se

consideraron investigaciones de fuentes secundarias como tesis y publicaciones universitarias. Todo sobre una fuente válida entre el año 2000 y 2020. Para el análisis del impacto de la Big Data Analytics en el campo medico se toma en consideración investigaciones desde el año 2015 al 2020.

Dentro de los criterios de exclusión no se toma en consideración publicaciones de revistas que no sean científicas o académicas, autores con investigación limitada, investigaciones anteriores al año 2000, y criterios expresados por fuentes terciarias.

1.1.1.6 CADENA DE BÚSQUEDA

Se tomó en consideración los vocablos o frases que expresan los criterios de búsqueda que se aplican en cada fuente o plataforma de búsqueda, tales como, título, índice, palabras claves, búsquedas avanzadas, metadatos, citación. Como es lógico se requirió de la siguiente ecuación de búsqueda tomando en cuenta los operadores AND y OR además del idioma (español-inglés):

Español:

- (Big Data OR Ciencias de los datos OR Tecnología de los datos OR Machine Learning OR Business Inteligence) AND (Concepto OR Definición OR Importancia OR “Ventajas y Desventajas” OR Criterios OR Evaluaciones)
- (Big Data OR Ciencias de los datos OR Tecnología de los datos OR Machine Learning OR Business Inteligence) AND (ETL OR “Técnicas de recopilación de información, extracción y carga”) AND (“Modelo de implementación”)
- (Big Data OR Ciencias de los datos OR Tecnología de los datos OR Machine Learning OR Business Inteligence) AND (Arquitectura OR Diseño OR Modelos) AND (herramientas OR aplicaciones OR framework)
- (Big Data OR Ciencias de los datos OR Tecnología de los datos OR Machine Learning OR Business Inteligence) AND (Salud OR Medicina OR Hospitales OR Medico OR ciencias médicas) AND (Beneficios OR Ventajas OR Desventajas)

Inglés:

- (Big Data OR Data Science OR Data Technology OR Machine Learning OR Business Intelligence) AND (Concept OR Definition OR Importance OR “Advantages and Disadvantages” OR Criteria OR Evaluations)
- (Big Data OR Data Science OR Data Technology OR Machine Learning OR Business Intelligence) AND (ETL OR “Information Collection, Extraction and Loading Techniques”) AND (“Deployment Model”)

- (Big Data OR Data science OR Data technology OR Machine learning OR Business intelligence) AND (Architecture OR Design OR Models) AND (tools OR applications OR framework)
- (Big Data OR Data Science OR Data Technology OR Machine Learning OR Business Intelligence) AND (Health OR Medicine OR Hospital OR Medical OR Medical Science) AND (Benefits OR Advantages OR Disadvantages)

1.1.1.7 SELECCIÓN DE ESTUDIOS

Para el desarrollo del RSL de este estudio en las diferentes fuentes ya expuesta, se realizaron los siguientes pasos:

- Realizar la búsqueda de información según la cadena de búsqueda ya definida.
- De los artículos y publicaciones resultantes decidir cuales incluir y excluir.
- El resultado de la investigación se vio definida en el contexto de la investigación es decir en los criterios a tomarse en cuenta para ser válidos según el criterio de la investigación.

La Tabla 4 muestra la cantidad de artículos, publicaciones, texto y trabajos encontrados tanto para el idioma inglés y español.

Tabla 4. Distribución de búsqueda por fuente

Fuentes	Artículos encontrados	Filtrados por titulo	Filtrado por resumen	Filtrado por palabras claves	% por fuente
Scopus	1250	25	6	8	**19%**
Dialnet	1458	48	18	11	**22%**
WoS	753	10	5	2	**11%**
IEEE	862	19	4	1	**13%**
Springer	589	35	11	7	**9%**
Scielo	1050	37	9	7	**16%**
Science Direct	753	7	2	1	**11%**
TOTAL	**6715**	**181**	**55**	**37**	**100%**

Fuente: Autor

1.1.1.8 RESULTADOS DE LA REVISIÓN

De acuerdo con la revisión realizada a las diferentes fuentes, y la selección de los artículos que servirán para el desarrollo de este trabajo en la Tabla 5 se exponen los siguientes resultados en base a las preguntas de investigación:

Tabla 5.- Resultado de la revisión RSL en fuentes bibliográficas

PREGUNTAS	TEMAS REFERENTES A LA INVESTIGACIÓN	PORCENTAJES	TOTAL
1- ¿Dentro del ámbito tecnológico, qué beneficios ha traído la implementación de Big Data a las organizaciones?	Económicos	10%	100%
	Estructurales	10%	
	Toma de decisiones	30%	
	Modelos de implementación	10%	
	Mejora de los procesos	40%	
2.- ¿Cuáles son las herramientas o técnicas a usar para una buena implementación de Big Data?	Spark	5%	100%
	Hadoop	40%	
	Lenguajes de programación de alto nivel	10%	
	Zookeper	20%	
	Hive	5%	
	ETL	20%	
3.- ¿Cuál es el impacto en el área de Salud y que beneficios ha otorgado en otros países?	Atención medica	10%	100%
	Predicciones de enfermedades	50%	
	Evaluaciones de alto riesgo	30%	
	Control hospitalario	10%	

Fuente: Autor

Del análisis respectivo, se considera lo siguiente:

1. Dentro del ámbito tecnológico la implementación de la Big Data Analytics genera mayor interés en la toma de decisiones con un 30% y una mejora de los procesos internos en las organizaciones con una representación del 40% del material investigado.
2. Las técnicas o herramientas con mayor evidencia científica con los procesos ETL se representa en un 20% y Hadoop en un 40%, cabe destacar que las demás

herramientas son útiles y pueden ser consideradas dentro del trabajo de investigación.

3. En cuanto al impacto en la Salud por parte de Big Data Analytics, se tomaron criterios relevantes con grandes utilidades para la predicción de enfermedades y evaluaciones de alto riesgo en pacientes.

Por lo expuesto, se puede determinar que Big Data Analytics es una práctica válida para el análisis de información. Si bien es cierto que se nombra a la Big Data dentro del campo médico; en los hospitales públicos no se hace mucho como para percibir su beneficio. Más bien solo se exponen teorías o percepciones, es por ello que el enfoque de este trabajo de tesis, apunta a resolver ese problema, dar uso a la Big Data Analytics dentro de los hospitales públicos.

1.2 ANTECEDENTES CONCEPTUALES Y REFERENCIALES

Para el desarrollo del marco conceptual, se detallará claramente las variables de investigación, por un lado, la Big Data Analytics como eje central, destacando su historia, el termino conceptual, características, ventajas y desventajas, aplicaciones, herramientas. Por otra parte, también se analizará el GPR (Gobierno por Resultado), por qué hay que usarlo, quien lo lidera y que detalles se deben de considerar en su utilidad. Dentro de esta sección también se tomará en cuenta los indicadores de gestión hospitalaria que se proponen para el desarrollo de esta tesis.

1.2.1 BIGDATA

1.2.1.1 HISTORIA

En el libro de la revolución de los datos escrito por Víctor Mayer [12], se detalla la aparición de la Big Data desde la era Paleolítica, donde las personas acumulaban la información en ciertas zonas dando uso a piedras, palos, paredes para registrar lo que han visto o experimentado. Sin embargo, no se hacía mucho por ella, es decir no le daban mucha utilidad, luego en la época de Babilonia se crea la primera biblioteca donde guardaban más de medio millón de registros de la humanidad, con la llegada de los romanos fue destruida. En el siglo II AC se crea una máquina en Grecia que servía para predecir las posiciones astrales de las estrellas. Luego en el año 1663 se estableció el primer análisis estadístico predictivo en contra de la peste bubónica, con el pasar de los años en 1865 se establece el término Business Intelligence a la acción realizada por el banquero Henry Furnese, el cual tomo ventaja mediante un análisis predictivo de sus

datos en las actividades comerciales que desempeñaba, esto le dio ventaja ante sus competidores. A partir del año 1989 ya empezó la evolución de la Big Data, la cual se detalla en la Tabla 6.

Tabla 6. Evolución de la Big Data en el tiempo

Año	Suceso
1989	Erik Larson habla por primera vez de la Big Data, en ese mismo año se empiezan a popularizar las herramientas del Business Intelligence.
1991	Nace el internet y la recopilación de información se vuelve más extensa.
1993	Se funda QlikTech, hoy en día Qlik, herramienta para el análisis estadístico y predictivo de la información.
1996	El precio del almacenamiento de datos se vuelve más accesible y los usuarios empiezan a usarlo.
1997	Google lanza su buscador, accediendo a una gran cantidad de información.
Desde el año 2000 a la actualidad	Con la revolución de la información y el avance de la tecnología, la Big Data fue tomando mayor forma, que puede ser accedida desde cualquier dispositivo. Las empresas empiezan a invertir para poder obtener datos relevantes de sus datos. Los términos de Machine Learning & IoT ya es un hecho, ya solo queda esperar lo que quedara para el futuro.

Nota: Tomando del libro Big Data. La revolución de datos. [12]

1.2.1.2 CONCEPTO

Gartner Inc., empresa con sede en Stamford, Connecticut, Estados Unidos, líder en consultoría e investigación de las tecnologías de información, define el significado de Big Data [13] "*La Big Data es un activo de información de gran volumen, velocidad y variedad que exigen formas rentables e innovadoras de procesamiento de información, para mejorar la comprensión y la toma de decisiones*". Este concepto abarca lo siguiente:

- Es un activo, referencia directamente a las empresas, si bien es cierto el mayor activo más importante en una organización es la información.
- Tiene volumen, velocidad y variedad, es completamente correcto, ya que la Big Data es enorme en cuanto a datos, la velocidad que se genera es increíble en la gran cantidad de dispositivos que existen actualmente y su variedad es incontable.

- Exige innovación en el procesamiento de información, cada vez más herramientas son utilizados para este fin.
- Comprensión y toma de decisión, estos dos términos comprenden al entendimiento de los datos luego de ser procesados, y con ello ser útiles para decidir qué hacer y cómo hacerlo.

Otros autores detallan otros conceptos tales como:

- El Big Data consta de datos tan grandes y complejos que es imposible manejarse con los métodos tradicionales de procesamiento [14].
- Es un término evolutivo que representa a la gran cantidad de datos estructurados, semi estructurados y no estructurados [15]
- Refiere a grandes cantidades de información, de varios tipos, que se producen a gran velocidad y proviene de diferentes fuentes, cuyo manejo y análisis requiere exclusivamente de potentes procesadores y algoritmos [16].

Con esta base, se puede deducir que la Big Data es la representación voluminosa de los datos que puede generarse de diferentes fuentes y por ende sirve para su posterior análisis para la toma de decisiones.

1.2.1.3 CARACTERÍSTICAS

Según un estudio realizado por la Escuela de Negocio de la Universidad de New York, destaca a la Big Data como la representación de las tres V (Volumen, Variedad y Velocidad), incluso Gartner apoya la misma teoría. ¿Pero qué refiere estas características?, se la expone a continuación [17]:

- Volumen:
    - Referencia a la gran cantidad de datos que se obtiene minuto a minuto a nivel mundial. Se estima que se produce millones de megabytes de datos que es casi imposible procesar, esto es uno de los retos de la Big Data, poder almacenar y procesar gran cantidad de información.
- Variedad:
    - Los datos no son homogéneos y se obtienen de diferentes fuentes, por lo que pueden ser estructurados y no estructurados. Uno de los retos de los procesadores de datos en Big Data es tener la particularidad de poder procesar la información de diferentes lugares, muchas veces hay que pre procesar los datos para crear repositorios homogéneos que permita un buen análisis de la información.

- Velocidad:
    - Se refiere a la rapidez con que los datos son creados, procesados y almacenados.

Es importante destacar que la Big data también posee otras características que debe cumplir como [18]:

- Veracidad. - De fuente confiables.
- Volatilidad. - Período de tiempo en lo que los datos son válidos.
- Viabilidad. - Posibilidad de que los datos aporten a las necesidades requeridas.
- Valor. - Información exclusiva e importante para la empresa
- Visualización. - Debe cumplir con los parámetros de presentación según las necesidades solicitadas.

1.2.1.4 VENTAJAS

La Big Data siendo un campo de exploración amplia se detalla cuatro puntos importantes a tomar en cuenta como sus ventajas, y estos son [19]:

- Velocidad para la toma de decisiones:
    - Al obtener datos precisos y exactos según el requerimiento expuesto, es más fácil tomar una decisión, sin entrar al detalle de los datos, es por ello que es veloz, claro está que, para tomar una decisión, los datos deben ser preparados y previamente analizados.
- Planes estratégicos:
    - Los datos que nos proporciona la Big Data permiten establecer predicciones y comportamientos, es así, que se puede realizar planes de negocios o estrategias que beneficien al usuario. Por ejemplo: Un plan de marketing promocional sobre un producto. La predicción de una enfermedad posible en cierta zona del mundo, etc.
- Vinculación con el usuario o cliente:
    - Conocer características personales de cada individuo permite tener una mayor vinculación de lo que necesita o posiblemente requiera a futuro.
- Feedback en tiempo real:
    - La Big Data no solo supone el análisis de información para tomar una decisión, sino más bien la interacción en tiempo real de los datos, para conocer qué está pasando en el entorno.

1.2.1.5 DESVENTAJAS

La Big Data como ciencia para el manejo de los datos, también tiene sus desventajas y estas son [20]:

- Ciberseguridad:
    - A mayor cantidad de información que se maneja, mayor exposición está a un ciberataque.
- Protección de datos:
    - Es importante siempre establecer los protocolos de uso y manipulación de los datos.
- Tecnofobia:
    - Muchos usuarios aun no aceptan el uso de su información para fines de exploración e interpretación, por lo que consideran como vulnerar su privacidad.

1.2.1.6 APLICACIONES EN EL AREA MÉDICA

Dentro del ámbito médico, la Big Data tiene muchas utilidades. Para el caso de este proyecto, se apunta al desarrollo de indicadores de gestión hospitalaria, además de obtener los datos solicitados por el GPR (Gobierno por resultado) que solicita el Ministerio de Salud Pública, a medida de ejemplo se puede destacar:

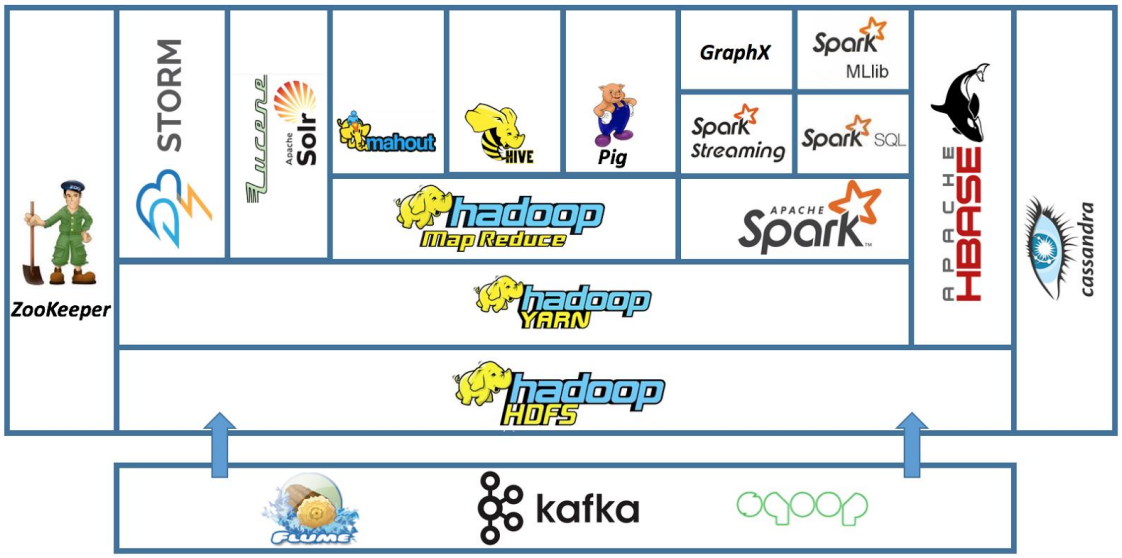
- Optimización sociosanitaria: El objetivo es poder tener la información del paciente y determina lo que necesita a tiempo.
- Sistemas de alertas: Con el análisis de los datos, se puede establecer alertar a pacientes que deben seguir un régimen de medicamente o tener revisiones médicas constantes.
- Predicción de necesidades para pacientes crónicos: En caso de ser necesario el doctor estará al tanto de los medicamentos o asistencia al paciente de forma oportuna.
- Optimización de recursos: Se puede establecer tendencias de consumos y gastos dentro del área de bodega, con el fin de mantener al día el stock.

El campo de la Big Data Analytics puede ayudar de cualquier forma al área médica, siempre y cuando se establezcan las necesidades analizar.

1.2.1.7 HERRAMIENTAS PARA EL ANÁLISIS EN BIG DATA

Hoy en día las herramientas para el análisis de los datos a gran volumen se han diversificado, cada una con funciones útiles según las necesidades del usuario o administrador de Big Data. En la Figura 1, se puede observar las herramientas que actualmente circulan en el mercado, destacándose por su funcionalidad y facilidad de acceso. Cabe destacar que así mismo como existen sin costo, también existen con pago de licencias, con mayores funcionalidades.

Figura 1.- Herramientas para Big Data en el mercado



Fuente: Fuente: Tomado de la página web Arquitectura de Big Data, referencia bibliográfica [52]

Dentro del contexto de esta investigación se tomará en cuenta las mejores herramientas de código abierto (open source) para el tratamiento de la Big Data, para ello se destacan [21]:

- Hadoop MapReduce
- Hadoop Yarn
- Apache Hive
- Lenguaje R
- Lenguaje Python
- Apache Zookeper
- Apache Spark

Hadoop MapReduce. Hadoop nace en el año 2008 como proyecto de código abierto siendo el resultado de varias investigaciones orientadas a crear una aplicación que

permita el manejo distributivo de la información en la web. Hadoop actualmente está soportado por Apache Software Foundation, empresa sin fines de lucro [22].

Hadoop es un framework que se destaca en el procesamiento de datos distribuidos, permite el escalamiento de una máquina a muchos servidores locales. La biblioteca de Hadoop está diseñada para detectar fallos en la capa de aplicación, lo cual propone alta disponibilidad. Las características con las que cuenta son:

- Capacidad para procesar grandes volúmenes de información.
- Poder de cómputo, pues al tener más nodos (equipos, servidores) tiene más capacidad de procesamiento.
- Tolerancia a fallos, debido a su procesamiento distribuido en varios nodos, se minimiza los fallos, ya que automáticamente se transfieren los procesos a otro nodo.
- Bajo costo, permite ahorrar en licencia y su implementación puede trabajar en equipos comerciales que permitan el almacenamiento de grandes cantidades de información.
- Escalabilidad, si existe la necesidad de procesar más información, fácilmente puede agregar más nodos.

Lenguaje R. Es un lenguaje libre de licencia abierta, que ejecuta cada línea de código en tiempo de diseño, esto quiere decir que no requiere de compilarse y es interpretativo. Su uso es ampliamente destacado en el área de la estadística y gráfica [23]. Es muy popular en el aprendizaje automático, minería de datos, investigación biomédica, etc. Entre sus características se destacan:

- Maneja efectivamente los datos.
- Contiene un grupo de operadores de cálculos.
- Mantiene integrada una herramienta versátil para el análisis de datos.
- Contiene opciones para realizar gráficos estadísticos.

En el campo de la Big Data, R es usado para representaciones de datos gráficos, creación de dashboards y generación de informes automáticos. Su utilización puede darse en las siguientes fases:

- Recopilación y preparación de los datos. Extracción de las fuentes de datos y eliminación de duplicidad.
- Análisis de los datos. Construcción de los modelos predictivos.

- Comunicación de los resultados. Generación de informe y exposición de los resultados.
- Aplicación de los resultados obtenidos. Toma de decisiones.

Lenguaje Python. Es un lenguaje de alto nivel de tipo interpretativo que se utiliza especialmente para la analítica de datos. Python soporta orientación a objetos, programación imperativa, es dinámico y multiplataforma. Dentro del campo de la Big Data es muy útil incluso para personas que no conocen de programación. Además, integra un gran número de librerías que permiten fácilmente el procesamiento de los datos [24].

Hadoop Yarn. Es una de las piezas fundamentales dentro del entorno de HADOOP, que permite soportar varios motores de ejecución incluso tomando en cuenta MAPREDUCE. Hadoop Yarn trabaja como un organizador y administrador dentro del entorno de aplicación. Tiene los siguientes componentes:

- Administrador de recursos. Se encarga de gestionar los recursos en el clúster.
- Administrador de nodos, responsable de administrar los contenedores en donde se ejecutan las aplicaciones en cada nodo.
- Aplicación master, administra y controla el ciclo de vida de las aplicaciones.

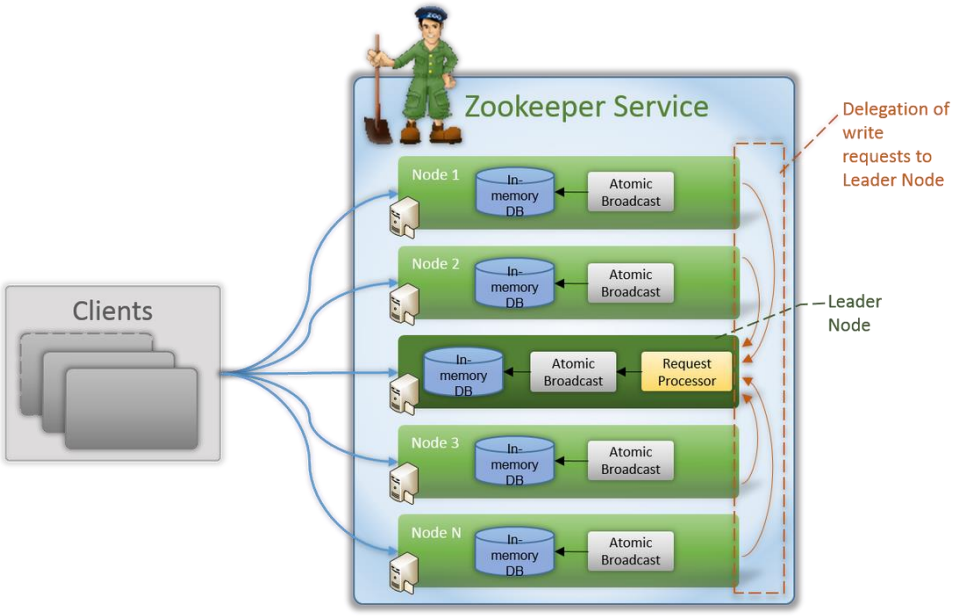
En relación a su diferencia ante MapReduce, se destaca por ser el responsable de administrar y controlar los recursos. En cambio, MapReduce es un framework de programación al que hay que indicarle que debe ejecutar.

Apache Hive. Es un software que permite la gestión de la información mediante su propio esquema de consultas (querys), que trabajan en la misma estructura Hadoop Distributed File System (HDFS). Se lo puede reconocer por su forma de manejo de sentencias SQL como Hibernate Query Language (HQL), que en realidad se representa como una variante. Su trabajo consiste en crear consultas HQL que trabajando en un entorno MapReduce (i.e., trabajo distributivo entre nodos) con el fin de obtener los datos necesarios. Se debe de considerar que al realizar procesamientos de datos y traducción al lenguaje Java su latencia aumenta, por lo que se puede considerar como una desventaja. Apache Hive, no debe ser considerado como un motor de base de datos, pero si una herramienta para el manejo de datos en diferentes nodos.

Apache Zookeper. Es un servicio de coordinación en aplicaciones distribuidas que basa su trabajo en la sincronización de un clúster. Puede ser percibido como un contenedor

centralizado, donde las aplicaciones distribuidas puede manejar los datos entre obtenerlos y colocarlos. Generalmente se lo usa, para que todo un entorno trabaje de forma sincronizada y eficiente. La Figura 2 ilustra su arquitectura.

Figura 2.- Arquitectura Zookeeper



Fuente: Tomado del libro Apache Zookeper Essentials por el autor Saurav Haloi [24]

Zookeper, se basa en un modelo cliente-servidor, donde un nodo líder administra y gestiona los procesos, mientras que los demás nodos están a cargo de los clientes. Además, gestiona de forma eficiente el control de fallos por lo que si un nodo cae inmediatamente Zookeper lo reemplaza.

Apache Spark, es un framework de código abierto, que permite la gestión de clústers. Su propósito es de forma general para el desarrollo de arquitecturas Big Data y se caracteriza por su velocidad en el procesamiento. Debido a su gran uso tiene soporte para otras herramientas tales como, Yarn, Mesos, Spark Standalone, Cassandra, Azuru, Kudu, entre otras. Spark, también se caracteriza en trabajar de forma pseudo distribuida en un solo equipo, el mismo que trabaja utilizando cada núcleo del procesador, esto a efectos de uso como pruebas de datos.

1.2.1.8 ARQUITECTURA DE BIG DATA

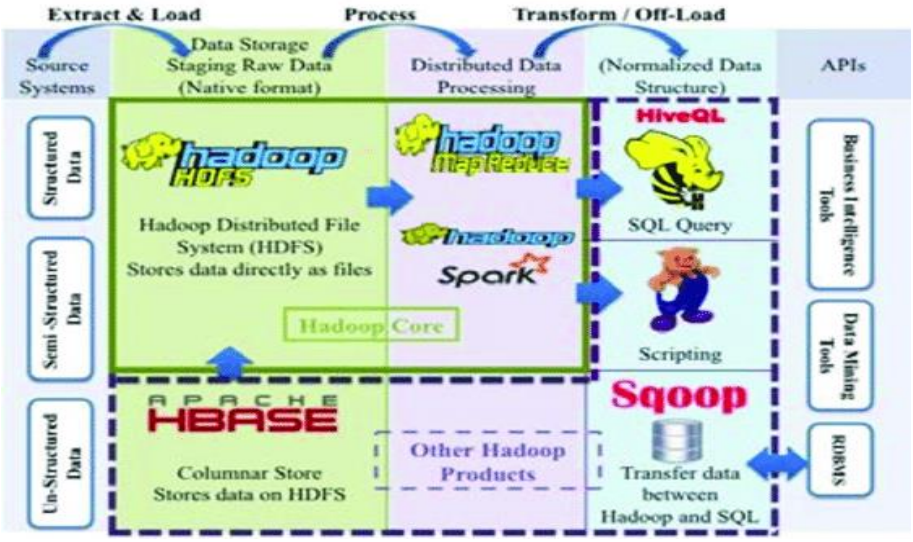
Según varias fuentes consultadas, la metodología de implementación de Big Data basado en una arquitectura dependerá de su utilidad y forma en como el administrador de Big Data la desarrolle, para efectos de comprensión se explica una de las

arquitecturas más usadas desde la recopilación de datos hasta su consumo. Cabe destacar que intervienen las técnicas ETL, para su desarrollo [25]. Para ello se toma en consideración las siguientes capas y/o ciclo de los datos:

- Capa de recursos de Big Data. Representa a toda la información estructurado y no estructurada que puede considerarse útil para el análisis de datos, la misma puede estar en diferentes fuentes. El origen de los datos dependerá de los siguientes factores:
    - Formato: Datos estructurados, semiestructurados y no estructurados.
    - Volumen: Cantidad de datos que se representan en la unidad de medida en bytes.
    - Punto de recopilación: Según sea fuentes primarias o secundarias.
    - Ubicación: Pueden ser datos interno o externo a la empresa.
- Capa de almacenamiento. Esta capa es exclusiva para la recopilación de información y encargarse en la conversión a un formato único que pueda ser almacenado en un sistema HDFS o en un sistema de base de datos relacionales (RDMS). Esto con el único fin de tener los datos listos para su procesamiento.
- Capa de análisis de datos. Esta capa se encarga de utilizar los datos almacenados desde la capa de almacenamiento, ciertas veces esta capa accede directamente a la información. Sin embargo, todo dependerá de cómo se establezca la planificación de lectura y procesamiento de datos. Para ello hay que tener en cuenta la herramienta a utilizar.
- Capa de consumo de datos. Esta capa hace uso de la información procesada para poder ser visualizada la aplicación que se requiera según la necesidad.

La Figura 3, detalla la arquitectura de Big Data y el procesamiento a seguir para el manejo de los datos, mediante herramientas de código abierto.

Figura 3.- Arquitectura Big Data & ETL



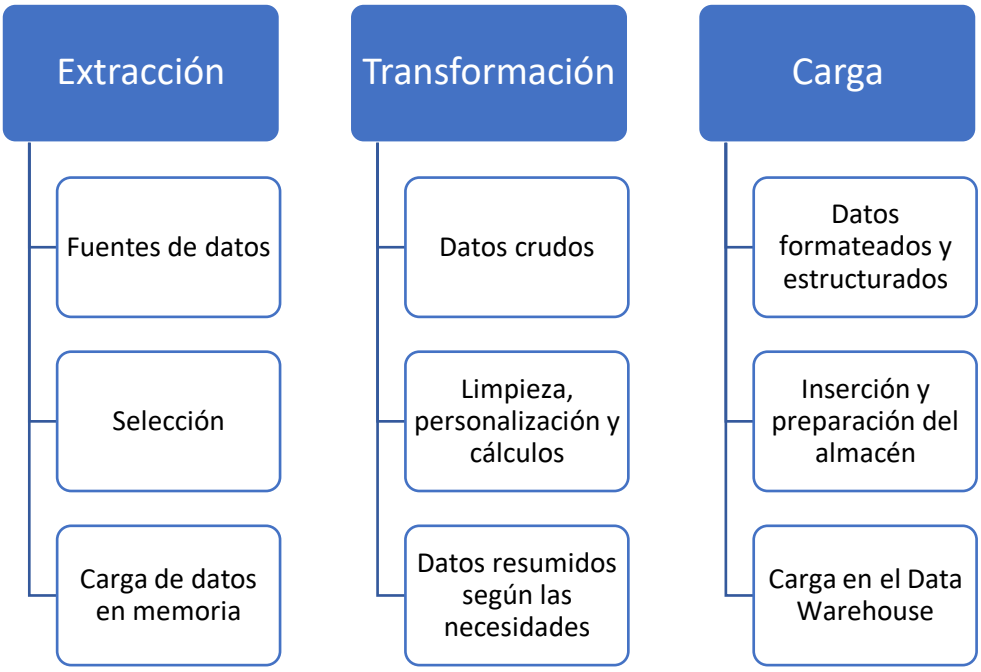
Fuente: Tomado de la web Research Gate by Hsiao-Kan-Lin [26]

El centro de cualquier almacén de datos, son los procesos de extracción, transformación y carga (ETL). Se puede determinar que el 70% de los recursos que se necesitan para la implementación y mantenimiento de una base de datos, son exclusivamente consumidos por este tipo de proceso [27].

En relación a la Figura 4, las técnicas ETL puede ser representada por:

- Extracción de los datos de una fuente apropiada o previamente analizada.
- Transportar los datos a un entorno donde será procesados.
- Transformar y calcular los datos.
- Depurar y limpiar los datos, con el fin de garantizar la estructura deseada y las reglas de negocio.
- Subir o cargar los datos limpios al respectivo Data Warehouse.

Figura 4. Procesos de un ETL



Fuente: Autor

Los procedimientos ETL, por su naturaleza se caracterizan generalmente por ser procesos de entrada, de operación y de salida. Cabe destacar que no existe un modelo estándar a seguir para realizar un procedimiento ETL, más bien queda a criterio a las necesidades y circunstancia del negocio, específicamente por temas de recursos.

1.2.1.9 TÉCNICAS DE ANÁLISIS DE DATOS (BIG DATA ANALYTICS)

Existen una variedad de técnicas para el análisis de datos que se pueden adaptar según las características de los datos recopilados, estas se clasifican en dos grandes ramas [28] :

- Según su objetivo: Existen diferentes formas de poder analizar la información desde el mínimo criterio al más complejo, con la finalidad de obtener una respuesta a una decisión a tomar, tales como:
    - Técnicas descriptivas: Son herramientas que permiten obtener datos según la realidad del negocio. Estas pueden ser, tasas de variación, tablas de frecuencias, árboles de decisión, etc.
    - Técnicas predictivas: Son técnicas que permiten obtener información a una respuesta del futuro, en base a ciertos criterios o variables que permitan ajustar datos para tener una mejor simulación y toma de decisión. Entre las cuales destaca, series temporales, regresión lineal, redes neuronales, machine learning, entre otros

- Técnicas prescriptivas: Son técnicas que permites obtener una recomendación a una situación dada, tomando en consideración reglas causa/efecto o algoritmos que puedan optimizar los procesos, por ejemplo, el modelo Montecarlo.

- Según su naturaleza: Para comprender de donde viene y cómo se comporta la información, hay que entender que existen datos según su naturaleza ya sea:
  - Por su volumen: Específicamente se destaca por el análisis de datos mediante machine learning, con los fines de aprendizaje automático y aplicación de redes neuronales.
  - Po su tipología: Se refiere claramente por el tipo de datos que se encuentra almacenado, y con ello la interacción de información en texto, imágenes, video, audio, etc. Dentro de esta etapa se toma en consideración el análisis semántico que parte del lenguaje natural de los datos, el análisis de sentimientos, el cual puede interpretar estados de ánimo según datos recopilados y el análisis multimedia que permite detectar patrones de comportamiento en archivos multimedia.

1.2.2 GPR (GOBIERNO POR RESULTADO) E INDICADORES HOSPITALARIOS

De conformidad al documento publicado por la Secretaria Nacional de Administración Pública [29], en el que se detalla al GPR, como un sistema integrado de herramientas, conceptos y metodología en cumplimientos de las mejores prácticas de gestión administrativa pública. Con su aplicación permite establecer todos los planes estratégicos, operativos, riesgo, proyectos y procesos institucionales en los diferentes niveles organizacionales.

En el ámbito de la Salud, es importante que se detalle mes a mes los índices de gestión y medición en relación al cumplimiento de los planes estratégicos que tiene el Ministerio de Salud Pública y el Gobierno Nacional, cumpliendo con todos los lineamientos y parámetros a detallar según las necesidades para cada Institución Pública. El Hospital en análisis al ser una institución médica pública debe reportar todos los indicadores solicitados por el GPR.

1.2.2.1 INDICADORES HOSPITALARIOS

Los indicadores hospitalarios son datos resumidos que permiten una visualización más concreta de lo que se requiere conocer. En pocas palabras provee un resumen estadístico de un conjunto de datos en particular. Dentro del sector de la Salud, existen

varios indicadores que son válidos para la toma de decisiones e incluso pueden ser procesados por un previo análisis con Big Data. Para definir un indicador hay que tomar en consideración dos aspectos [30]:

- Indicador
    - Permite ver muestras y tendencias.
    - Es válido para ser comparado entre dos o más variables de análisis.
    - Permite el control y seguimiento de las tareas internas.
    - Pueden detectar desviaciones con el fin de mejorar la toma de decisión.
- Estándar
    - Rango de valores aceptados que se consideren resultados normales.
    - Siempre está ligado a la evolución de los resultados.
    - Son punto de referencia para validar si el desempeño es el correcto.

En un hospital son válidos los siguientes indicadores:

- Indicadores de gestión hospitalaria.
- Indicadores de atención al cliente.
- Indicadores de gestión de inventario.
- Indicadores de control de costos y presupuestos.
- Indicadores de control enfermedades.
- Indicadores de control de paciente, hospitalizados y en emergencia.

En relación al trabajo a realizarse, se integra el análisis de los requerimientos solicitados por el GPR y un dashboard de indicadores útiles para el área de gerencia y dirección hospitalaria, ya que no cuentan con esos recursos para la toma de decisiones, los datos se encuentran dispersos y son muy difíciles agruparlos.

1.3 ANTECEDENTES CONTEXTUALES

El siguiente trabajo de tesis se centra en un Hospital General, el cual está ubicado en la ciudad de Machala, provincia del Oro. Su misión está enfocada en la atención médica especializada en pacientes que recurren a ese centro hospitalario. Su visión se basa en la acreditación de las normas internacionales y nacionales de calidad de atención medica e infraestructura completa. Según el último informe [31] presentado del año 2019 en el mes de septiembre del año 2020, se destaca lo detallado en la Tabla 7.

Tabla 7. Servicios relevantes según Rendición de Cuentas del año 2019

Servicios relevantes	Promedio de personas por día
Atenciones médicas de consulta externa	293
Total de atención en emergencia	245
Total de Egresos hospitalarios	29
Total de altas diarias	28
Total de defunciones	1
Total de cirugías programadas	9

Fuente: http://htdeloro.gob.ec/rcuentas2019/Rcuentas2019.pdf

Estudios realizados en el área médica para la implementación de la Big Data se detallan: El trabajo propuesto en la Universidad Técnica de Machala en la que se destaca la importancia de la Big Data Analytics en el área de la Salud [32]. Específicamente en la predicción de brotes de enfermedades infecciosas, determina de forma específica el gran aporte que se hace, cuando se estudia desde una fuente de información centralizada, las características, patrones y comportamientos que puede tener una enfermedad y su modo de prevención. Además de este estudio se puede destacar otros indicadores a nivel hospitalario con el fin de mejorar la atención al paciente y cumplir con los parámetros indicados por el Ministerio de Salud Pública del Ecuador, entre otros. Big Data Analytics como tal, abre un abanico de opciones para el respectivo análisis y toma de decisiones.

En la actualidad las unidades hospitalarias cada día están acumulando una gran cantidad de información que amerita ser analizada. Datos como, por ejemplo, pacientes, exámenes, historias clínicas, imágenes médicas, inventario, partes diarios, camas disponibles etc.; lo cuales son importantes de analizar y que son muy útiles para la toma de decisiones. Si bien es cierto que los datos pueden ser procesados a menor escala, la unidad hospitalaria que requiere implementar el proyecto, demanda de forma urgente analizar varios indicadores que compense la recopilación de datos de varias fuentes y su necesidad a tiempo oportuno no solo para toma de decisiones sino también para la presentación de datos ante el Ministerio de Salud Pública. Esto también conlleva a establecer una mejor opción para los usuarios en cuanto al análisis de datos ya procesados y en línea.

1.3.1 PROPUESTA DE SOLUCIÓN Y CONTRIBUCIÓN

La propuesta de solución está orientada a:

- El uso de Big Data Analytics para el procesamiento de datos masivos que se generan diariamente en una entidad de salud pública, con el fin de poder dar análisis a los indicadores de gestión hospitalaria IGH, los cuales son solicitados por la alta gerencia y entidades gubernamentales para el control y manejo eficiente de los recursos del estado. Además, contempla la comparación de técnicas de análisis de datos sofisticada para entender y comprender qué tipo de modelo de datos es el más adecuado para su uso de forma general.

El desarrollo de la solución contribuirá de gran manera a una solución apta y viable, en razón de que, actualmente no existe una solución tecnológica que les permita tener los datos al día, en línea y completamente organizados, este proyecto será de utilidad para una posterior investigación de postgrado.

CAPÍTULO 2. METODOLOGÍA

2.1 TIPO DE ESTUDIO

El presente trabajo consta de dos fases, primero en el reconocimiento e identificación de la problemática actual en cuanto al manejo de la información y presentación de resultados al Ministerio de Salud Pública y GPR. La segunda parte presenta el diseño de la solución y los mejores criterios de implementación según la realidad actual. Para ello se toma en cuenta las opiniones de los involucrados y la infraestructura del hospital.

El desarrollo de este trabajo de investigación en cuanto al diseño del prototipo de implementación de una arquitectura de Big Data Analytics se basa en los siguientes métodos de investigación:

- Correlacional:
    - Se garantiza el análisis de las variables en estudio con el fin de obtener la relación que tiene la implementación de Big Data Analytics con la toma de decisiones dentro del centro hospitalario y los lineamientos de las normas estipuladas por el GPR (Gobierno por Resultado) a fin de cumplir con los resultados esperados del trabajo de investigación.
- Basada en objetivos:
    - Se toma en consideración el cumplimiento de cada objetivo específico para llegar al objetivo general de la tesis, con el fin de garantizar el desarrollo de la propuesta de implementación.

2.2 PARADIGMA

Según Thomas Kuhn, se define al paradigma de la investigación científica como la concepción general del objeto de estudio, los problemas y métodos a emplearse con el fin de comprender el caso de estudio y obtener los resultados esperados. En base a esa definición se toma en consideración el siguiente paradigma de investigación:

- Investigación cualitativa:
    - Con el fin de obtener los mejores criterios en cuanto al desarrollo, definiciones, características e implementaciones de arquitecturas de Big Data en el área médica, se tomó en consideración la obtención de datos de fuentes primarias y secundarias tanto de libros, textos y artículos científicos. Además, se obtuvo información del personal que trabaja

dentro del centro hospitalario y se toma de referencia el uso de la base de datos mediante técnicas ETL para la estructuración de los indicadores que ayuden a mejorar la visualización de los datos para la toma de decisiones.

- Investigación cuantitativa:
    - Este tipo de investigación se utilizó en este proyecto, para describir, explicar y predecir resultados mediante datos numéricos, uso de herramientas de análisis de datos, matemático y estadístico.

2.3 POBLACIÓN Y MUESTRA

En base a un estudio previo de la situación actual del Hospital, se tomó en consideración los datos más relevantes en cuanto a los indicadores de gestión hospitalaria analizar. Para ello se recopiló información de varias fuentes, obteniendo datos estructurados y semiestructurados. La población de datos identificados es:

- Documentos electrónicos
- Documentos físicos
- Matrices
- Base de datos

A nivel macro se puede considerar qué, si los datos estuvieran gestionados por una base de datos central con una buena infraestructura física, la misma llegaría a utilizar más de 20 GB de información que se han venido acumulando en diferentes fuentes de datos y que son de utilidad para el análisis y toma de decisiones. Esto en referencia a otras unidades hospitalarias que si cuentan con un proceso automatizado y pueden manejar la información de forma más rápida y online.

Este proyecto, se basa específicamente en el análisis de seis indicadores que se generan mensualmente y los cuales representan una acumulación de 1.121.875 registros desde el año 2014. Estos datos se tomaron en referencia como muestra para el desarrollo del proyecto. Es de vital importancia tener en cuenta que los resultados obtenidos a través del análisis de datos, será reflejo para futuros indicadores analizar. Entre ellos tenemos:

- Porcentaje de pacientes en espera de atención en consulta externa igual o menor a 15 días.
- Tasa hospitalaria de mortalidad materna.
- Porcentaje hospitalario de mortalidad neonatal.

- Número de pacientes en lista de espera quirúrgica.
- Porcentaje de ocupación de camas.
- Tasa de mortalidad hospitalaria.

Cada indicador tiene su origen en el análisis de otros indicadores secundarios, que permiten determinar el resultado requerido. Para una mejor compresión la Tabla 8 describe los indicadores utilizados para el análisis de datos:

Tabla 8.- Indicadores para el análisis de datos.

Indicadores primarios	Indicadores secundarios
PPEA.- Porcentaje de pacientes en espera de atención en consulta externa igual o menor a 15 días.	NPA.- Número de pacientes agendados para 15 días o menos en primeras consultas.
	TCE.- Total de pacientes que solicitan ser atendidos en consulta externa en primeras consultas.
THMM.- Tasa hospitalaria de mortalidad materna.	NDPO.- # Defunciones de pacientes obstétricas.
	TEPO.- Total egresos pacientes obstétricas.
PHMN.- Porcentaje hospitalario de mortalidad neonatal.	Sin indicadores secundarios.
NPLEQ.- Número de pacientes en lista de espera quirúrgica.	Sin indicadores secundarios.
POCP.- Porcentaje de ocupación de camas.	TPH.- Total de pacientes hospitalizados en el período.
	CDP.- Camas disponibles de ese período.
TDMH.- Tasa de mortalidad hospitalaria	TED.- Total de egresos hospitalarios por defunción.
	TEH.- Total de egresos hospitalarios.

Fuente: Autor

Un indicador primario es aquel que refleja el resultado promediado o condensado en base a los cálculos que se obtienen por los indicadores secundarios, en pocas palabras refleja el resumen de la información requerida. Un indicador secundario es aquel que se

alimenta de la información de fuentes primarias para que pueda ser útil para los cálculos requeridos por los usuarios. La unión de dos o más indicadores dan como resultado un indicador primario o de mayor importancia. Dentro de los materiales de análisis de información para el montaje del servidor y procesamiento de datos se tomó en cuenta lo siguiente:

- Sistema operativo Linux Ubuntu Server.
- Java.
- Hadoop y sus componentes.
- Lenguaje R.
- Python.

Para la visualización de datos se integró el uso de PHP con el framework Laravel y las librerías graficas en JavaScript de HighCharts. La solución de análisis de datos o BDA, se considera de gran utilidad para el desarrollo del proyecto las técnicas predictivas de datos, tales como series temporales y forescating en R, con el fin de determinar el mejor modelo que se adapte a las necesidades según los requerimientos de los usuarios y el Ministerio de Salud Pública. Cabe destacar que un modelo de pronóstico servirá como una guía ante las decisiones que se puedan a tomar de forma correcta.

2.4 MÉTODOS TEÓRICOS

El presente trabajo toma en consideración los siguientes métodos teóricos de investigación tales como:

- Método de recopilación de datos
    - Se aplican técnicas de recopilación de información con el fin de obtener información precisa y relevante para la estructuración de los datos. Esto conlleva a la exploración y revisión de las fuentes de datos actuales dentro del hospital. Cada registro, archivo y/o documento será verificado ante la autoridad o responsable de área.
- Método analítico
    - Ya que se requiere analizar todas las partes de un todo en cuanto a la Big Data Analytics y su aplicación en el área médica, técnicas de recopilación de datos (ETL) y las mejores prácticas en la implementación de este tipo de proyecto.
- Método sistemático
    - Debido a la naturaliza del proyecto, se requiere establecer los pasos adecuados a seguir para que los resultados esperados sean favorables.

- Método sintético
    - Para una mejor comprensión del lector, se resume el análisis de la información, creando una síntesis de cada tema estudiado, lo cual ofrece el mejor criterio en cada caso.

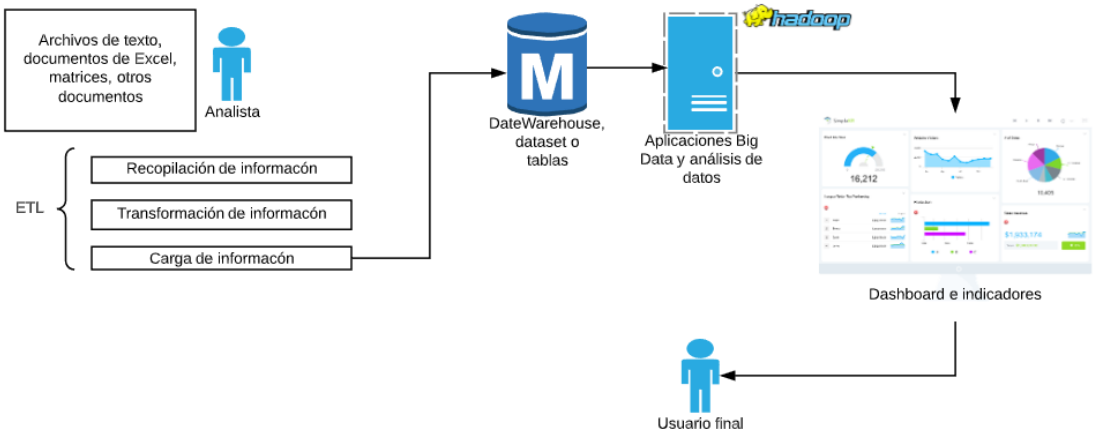
2.5 MÉTODOS EMPÍRICOS

Se toma en consideración los siguientes métodos para el desarrollo del objeto de estudio que involucra el impacto de la implementación propuesta y el criterio de otros autores con relevancia científica que permitirá tener un concepto más amplio de la solución y un criterio veraz y confiable de la investigación realizada.

- Para dar una mayor sustentabilidad de investigación en el desarrollo del proyecto, se incluyó en el Capítulo 1, la guía metodológica para la revisión sistemática de la literatura RSL de Bárbara Kitchenham, que otorga prioridad a las fuentes primarias y secundarias con validez científica desde las siguientes fuentes de datos como:
    - Scolpus
    - Dialnet
    - WoS
    - Springer
    - Scielo
    - Science Direct
- Fuentes de datos, debido a la naturaleza del proyecto, se requiere aplicar técnicas de recopilación, transformación y carga de información (ETL) a la arquitectura del diseño de implementación con el fin de obtener los resultados esperados. Se tomará en consideración las siguientes fuentes de datos:
    - Archivos de texto.
    - Documentos de Excel
    - Matrices
    - Otros documentos relevantes dentro de la investigación.
    - Registro de agendamiento de pacientes en MySql
    - Gestión de Bienes e inventario en SQL Server.

  Con la información obtenida se procedió a realizar el tratamiento, transformación y carga en Hadoop. La Figura 5 muestra la topología de recopilación de información aplicada a la base de datos siguiendo los lineamientos ETL.

Figura 5. Esquema de aplicación ETL para el procesamiento de información



Fuente: Autor

- Luego del tratamiento de los datos se pueden deducir métodos predictivos y descriptivos para la toma de decisión, información válida para el conocimiento de los usuarios finales. Para los siguientes procesos de análisis, se aplicó la metodología Cross-Industry Process For Data Mining (CRISP-DM).

2.6 TÉCNICAS ESTADÍSTICAS

Para los criterios dados y datos obtenidos se tiene en consideración lo siguiente:

- Técnicas predictivas de datos en BDA (Big Data Analytics):
    - Series temporales. Análisis de los datos en un periodo de tiempo.
    - Holt-Winters, SMLT, BATS, NNETAR. Se tomará en consideración este método predictivo ya que existirá un previo análisis de series temporales, lo cual será de mucha utilidad para ver el comportamiento de los datos.

A manera de resumen, este capítulo describe la forma en cómo se desarrolla el proyecto y las consideraciones sustanciales para que pueda ser calificado como viable. Se tomó en consideración el uso de técnicas para el análisis de datos BDA, las herramientas que se usan para Big Data y metodologías de la investigación que permitan el desarrollo correcto de este trabajo. El siguiente capítulo detallará el proceso a seguir para el uso de la Big Data y el desarrollo de técnicas de análisis de datos en lenguaje R.

CAPÍTULO 3. RESULTADOS OBTENIDOS

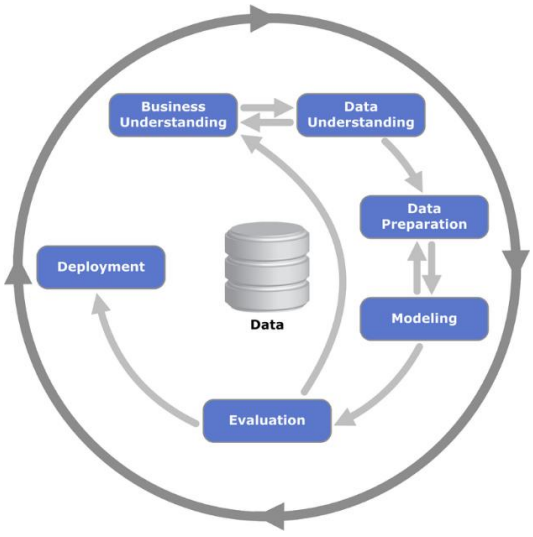
3.1 FUNDAMENTACIÓN TEÓRICA DE LA PROPUESTA

La unidad hospitalaria objeto de análisis y puesta en marcha de esta propuesta, ha evidenciado la necesidad de poder tener de alguna forma, la información estructurada y comprensible. Si bien es cierto que los colaboradores internos han realizado enormes aportes en poder unir la información y procesarla en hojas de cálculo como Excel para una mejor comprensión, hoy en día, se muestra que existe más información y compleja en recopilar. Por lo tanto, el uso de la Big Data Analytics permite automatizar el proceso de recopilación, transformación y visualización de los datos dándole el valor requerido dentro del Hospital o por lo menos el que se quiere obtener según las necesidades de los usuarios.

3.1.1 METODOLOGÍA DE IMPLEMENTACIÓN

Se aplicó a este proyecto la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) que para la minería de datos es un eje fundamental para la necesidad a cubrir dentro del desarrollo de esta propuesta, para ello se tomó en consideración las siguientes fases (ver Figura 6):

Figura 6.- Metodología de Minería de datos CRISP-DM



Fuente: Tomado de HealthDataMiner [32]

- Comprensión del negocio: Dentro de esta fase se establece las necesidades urgentes y prioritarias a cubrir. A nivel de minería de datos, se establece un plan

de acción a cumplir para la recopilación y transformación de los datos. Las variables detectadas dentro de este apartado son los indicadores tales como:

- Porcentaje de pacientes en espera de atención en consulta externa igual o menor a 15 días.
  - Número de pacientes agendado para 15 días.
  - Total de pacientes que solicitan ser atendidos.
- Tasa hospitalaria de mortalidad materna.
  - Número de defunciones de pacientes obstétricas.
  - Total egresos pacientes obstétricas.
- Porcentaje hospitalario de mortalidad neonatal.
- Número de pacientes en lista de espera quirúrgica.
- Porcentaje de ocupación de camas.
  - Total de pacientes hospitalizados en el periodo.
  - Camas disponibles de ese periodo
- Tasa de mortalidad hospitalaria.
  - Total de egresos hospitalarios por defunción.
  - Total de egresos hospitalarios.

- Preparación de los datos: Al tener identificadas las variables analizar según el proceso BDA en el modelamiento de datos, el objetivo de esta fase es tener ya los datos listos para su puesta en marcha a la transformación, con ello se dará uso de Hadoop para su preparación. Cabe destacar que esta parte los datos son clasificados y estructurados para su mejor comprensión. Por el simple hecho de dar análisis por series temporales y modelo de pronóstico, los datos tendrán un orden cronológico dentro del lenguaje R.
- Modelado: Esta fase contempla la selección de los modelos que tienen mejor relación a los resultados requeridos. Para tomar en consideración un modelo que va acorde a las necesidades a cubrir, deberá de tener mayor relación a los datos que se van presentando en el tiempo según cada indicador evaluado.
- Evaluación del modelo: Al disponer de los modelos adecuados para el análisis, se toma en consideración aquellos modelos que tienen mejores referencias o resultados a los que se requieren. Esta sección se desarrollará en el capítulo 4 donde se definirá el modelo que menor error genere en relación a los datos analizados.
- Despliegue: Esta fase determina toda la información que se expondrá a los usuarios quienes consumirán los datos expuestos para la toma de decisión, dentro del desarrollo de la propuesta, se establece que los usuarios consumirán los resultados en un ambiente web.

3.2 ARQUITECTURA DE LA BIG DATA

Según los estudios y fuentes científicas consultadas, no existe un modelo de arquitectura ideal o perfecta que pueda aplicarse a todo proyecto. Por el contrario, se entiende y comprende que depende mucho de los recursos y el alcance a seguir. Es por ello que se ha tomado en consideración un compendio de la mejor forma de crear una arquitectura de Big Data para este trabajo de investigación, con el fin de obtener buenos resultados y el cumplir de los objetivos planteados.

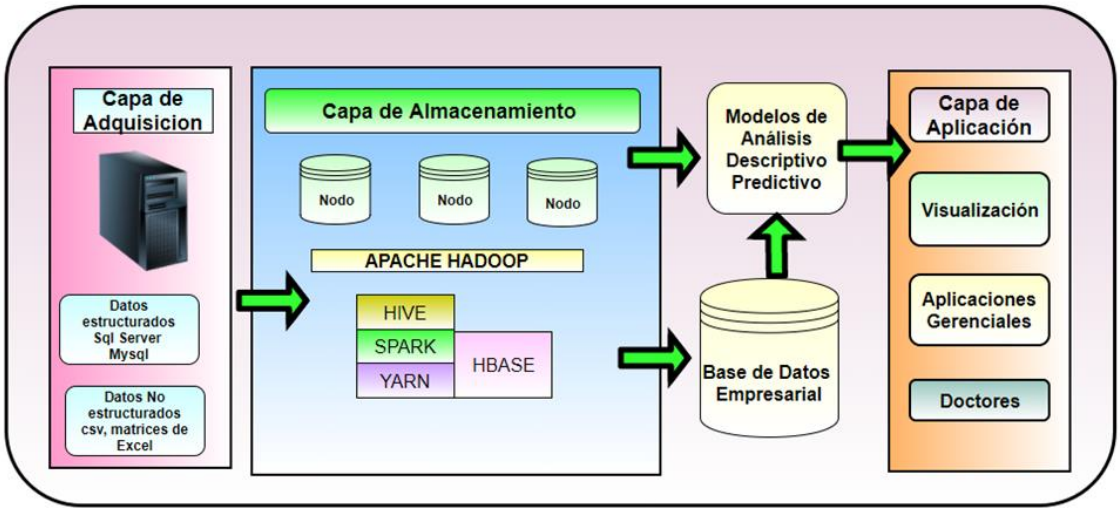
Para darle un mayor énfasis a la arquitectura dentro de un ambiente productivo cumplirá los siguientes aspectos:

- Extracción de información de fuentes confiables y verídicas.
- Generación de dashboard según la necesidad del usuario en cumplimiento a los indicadores ya detallado en el apartado anterior.
- Uso de herramientas adecuadas para el procesamiento de información.

La Figura 7 muestra de forma global la estructuración del modelo de arquitectura de Big Data a cumplir, los mismos que son mantenidos por tres capas fundamentales:

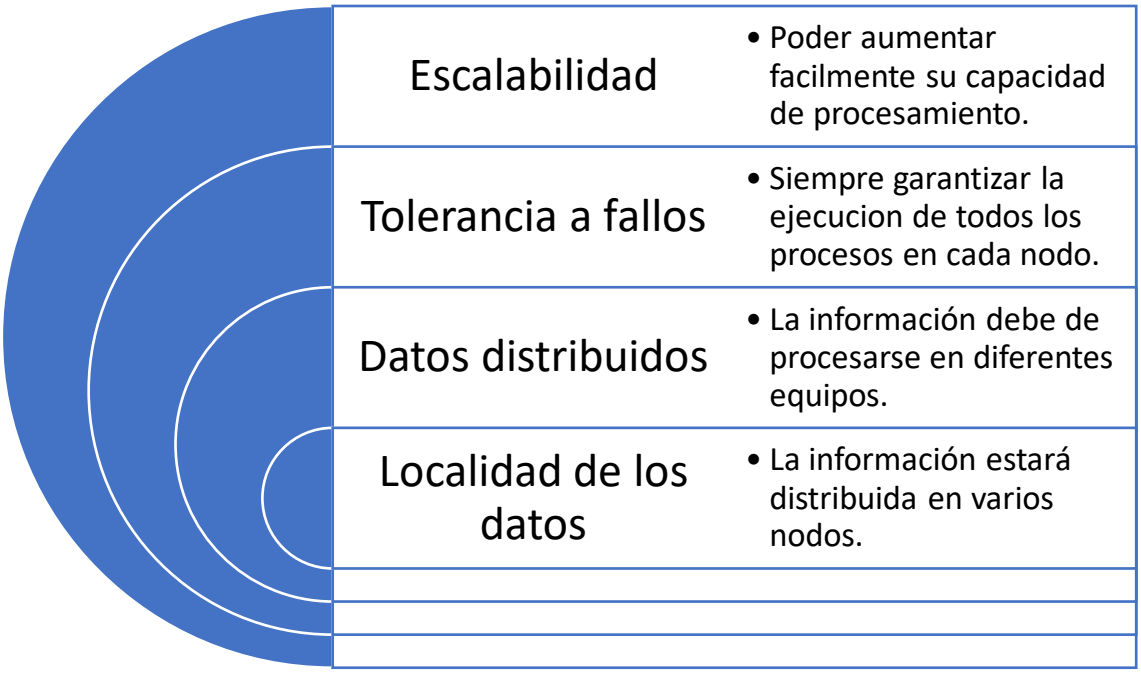
- Capa de aplicación: Se determina la recopilación de datos, su estructuración y seguimiento según la metodología CRISP-DM.
- Capa de almacenamiento: Detalla la estructuración de los nodos a trabajar y aplicaciones a usar, con el fin de tener los datos almacenados para su procesamiento.
    - Sección de modelamiento: Los datos para tener una mejor comprensión y sea útiles para los usuarios, deben ser una metodología de análisis de datos, para ello se establece los modelos más adecuados aplicarse.
- Capa de aplicación: Se encargará de proveer la información ya lista para su consumo mediante un acceso web que será usado por médicos y usuarios de altos mandos.

Figura 7.- Arquitectura de Big Data



Fuente: Autor

Entre las características a cumplir dentro del desarrollo del proyecto se toma en consideración el siguiente esquema:



3.2.1 RECURSOS PARA ELPROTOTIPO DE IMPLEMENTACIÓN

La unidad hospitalaria, cuenta con los siguientes recursos para la simulación de implementación:

- Un servidor de octava generación para la instalación de la fuente principal de procesamiento.
    - Este servidor esta virtualizado en compartición del sistema actual hospitalario.
- 5 PC básicas para establecer los nodos distribuidos.
- 1 equipo para la simulación del servidor web y presentación del dashboard.
- Equipos de red para la conexión y distribución de los datos.
- Software para la instalación de los sistemas de Big Data.

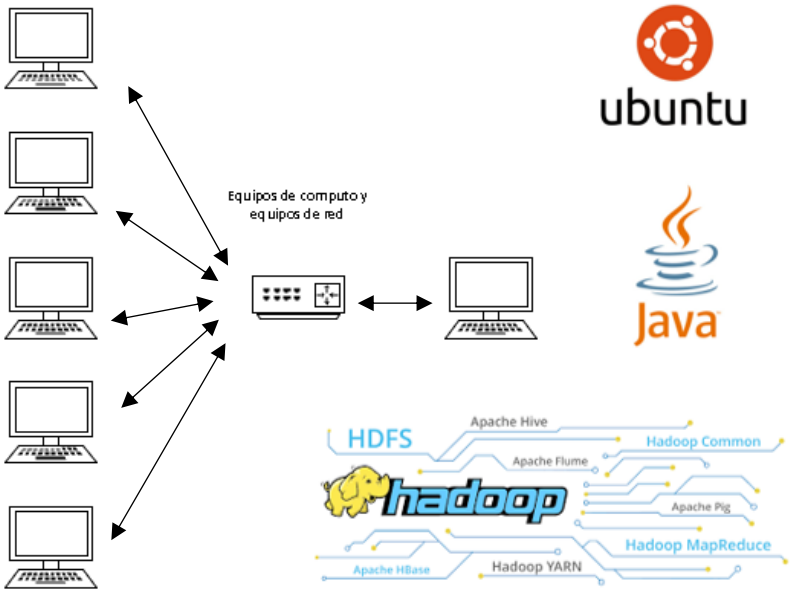
3.2.1.1 DISEÑO DE LOS RECURSOS

En el diseño se tomará en cuenta el modo distribuido o multinodo el cual destaca el uso de un computador master y computadores esclavos para la ejecución de procesos. La gran ventaja de esta configuración es el modelo de procesamiento el cual permitirá gestionar mayores cantidades de datos, ya que cada nodo aportará sustancialmente con procesamiento y memoria RAM.

Su desventaja puede recaer en la estructura de red, para ello se tomará en consideración equipos fiables para tales características en un caso se requiera procesar suficiente información. Para el objetivo del proyecto no es necesario implementar un equipo potente de red. El nodo maestro será el que distribuya todo el trabajo a los demás nodos.

La Figura 8, es una representación visual de cómo estará estructurado el proyecto, la cual destaca los recursos actuales y la forma en cómo se esquematizará el trabajo, con el fin de poder simular la transformación de datos y comprobar la utilidad de la Big Data en el campo médico.

Figura 8.- Esquema de implementación de los recursos



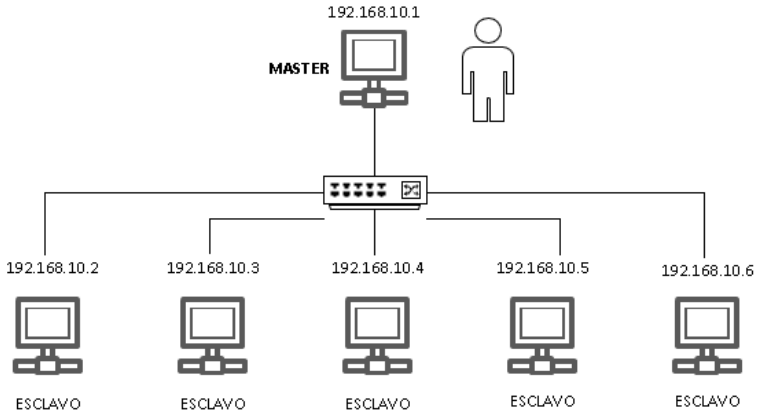
Fuente: Autor

3.2.1.2 ESTRUCTURA DE RED

El esquema de red para el modo distribuido cumple con la topología de estrella de tipo C. Todos los equipos se conectarán mediante un switch y estarán dentro del segmento de red:

192.168.10/24, tal como se muestra en la Figura 9.

Figura 9.- Esquema de configuración de red para Big Data multinodo



Fuente: Autor

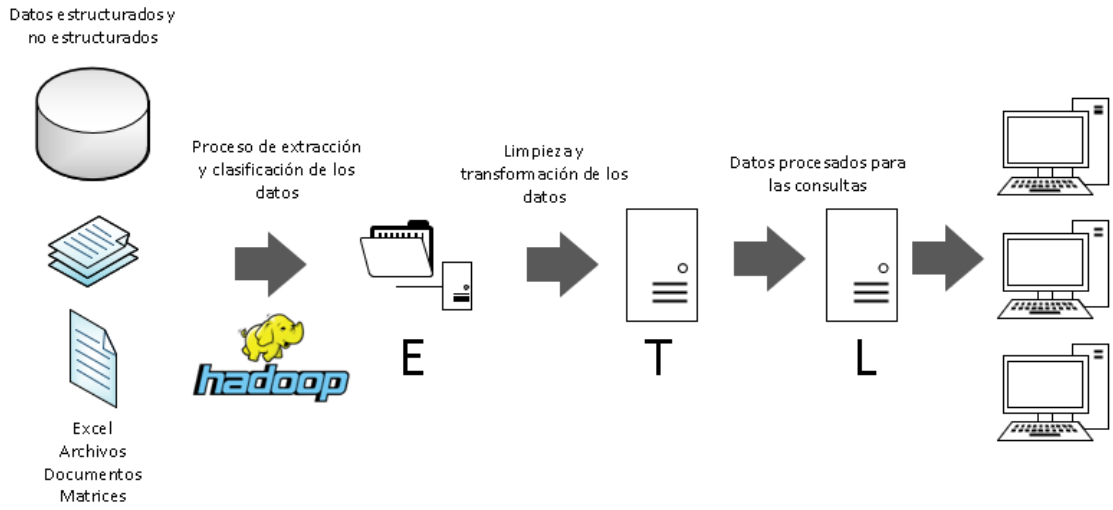
3.2.2 DISEÑO DE LA ARQUITECTURA

Para el diseño se toma en cuenta los recursos disponibles y el alcance del proyecto además de la implementación del modo distribuido, los cuales destacan exclusivamente con las técnicas ETL para la integración de la Big Data, por ello se determina tres fases del proyecto, las cuales se detallan a continuación:

- Fase 1: Extracción de la información
    - Datos estructurados
    - Datos no estructurados
- Fase 2: Procesamiento de la información
    - Hadoop y su compendio de herramientas
- Fase 3: Carga o presentación de la información
    - Dashboard e indicadores de resultados

En base al estudio realizado y conclusión de las fuentes científicas consultadas se tomará el diseño de la arquitectura que se muestra en la Figura 10.

Figura 10.- Diseño en la arquitectura de BIG DATA



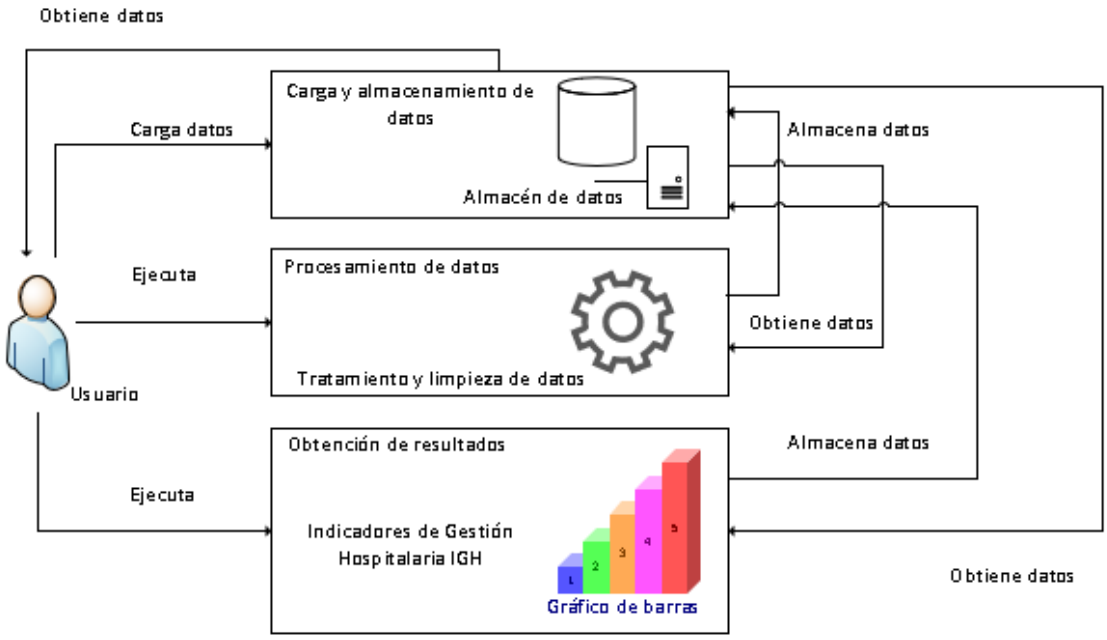
Fuente: Autor

Para ello Hadoop es el protagonista para el proceso de transformación de la información, con ello se espera obtener los resultados esperados. En cuanto a la presentación de los datos se establece el uso de indicadores mediante herramientas web como lo son PHP y HighCharts de JavaScript.

3.2.2.1 INTERACCIÓN DE LOS DATOS Y USUARIO

Es importante destacar cuales son los componentes que intervienen y como el usuario va a interactuar con los datos, que generalmente son tareas que conlleva la carga de datos por parte del usuario o administrador del sistema, el tratamiento de los datos o limpieza y la obtención de los resultados. Como se puede ver en la Figura 11, la estructura de Big Data como tal, cumple con los requerimientos necesarios para el objetivo de este proyecto, dando así lugar a su forma de implementación y comprensión del tema.

Figura 11.- Interacción del usuario y sistema en Big Data

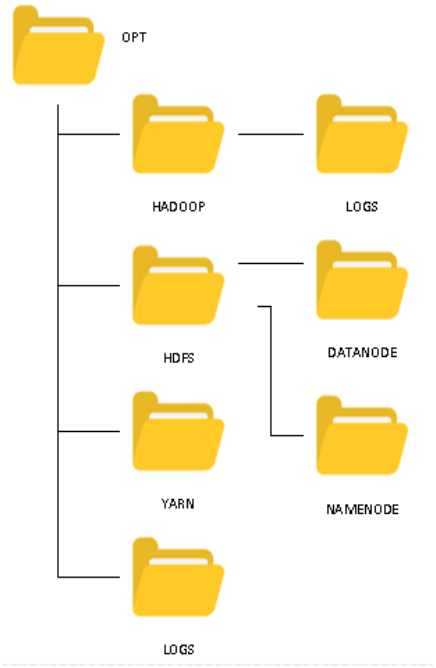


Fuente: Autor

3.2.3 IMPLEMENTACIÓN DE LA ARQUITECTURA

En la implementación de la arquitectura se toma en cuenta las tres fases que son la carga o extracción, la transformación y la presentación de los datos. El objetivo de este apartado es crear una estructura de archivos HDSF de tipo HADOOP, del cual el sistema recurrirá para seguir con las demás tareas ya sea el tratamiento de datos o presentación de los mismos. Para este fin, se diferencia los datos, unos los que son cargados por el usuario y sin procesar, y otros, los que están procesados. El mismo sistema debe ser capaz de establecer el esquema ideal de separación entre los dos tipos de datos. La estructura de los ficheros de forma general tendrá la siguiente presentación (ver Figura 12):

Figura 12.- Estructura general de las carpetas para Big Data

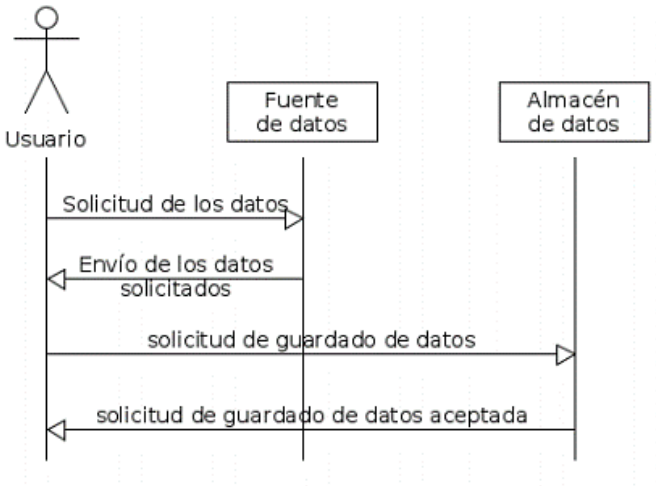


Fuente: Autor

3.2.3.1 FASE 1: EXTRACCIÓN DE LA INFORMACIÓN

Comprender como el usuario va a interactuar en esta parte, es importante, ya que es la manera en cómo la arquitectura se está proponiendo dentro del proyecto. La Figura 13 muestra el diagrama de secuencia respectivo.

Figura 13.- Interacción de usuarios en la fase inicial



Fuente: Autor

Dentro de esta fase se tomó en consideración todos los archivos y documentos que los usuarios usan para el registro de los datos, además de la información registrada en la base de datos interna. Los resultados se enumeran en la Tabla 9:

Tabla 9.- Archivos, matrices y datos extraídos dentro del Hospital para la preparación de la Big Data

Matrices en archivo de Excel	Registro de pacientes Registro de familiares de pacientes Atenciones médicas por emergencia Atenciones médicas por consulta externa Historia clínica de pacientes Evolución del paciente Registro de nacimientos neonatal Registro de fallecimiento neonatal Registro de defunciones en pacientes Registro de consumo hospitalario en cirugías de alto impacto Registro de consumo hospitalario de cirugías de menor impacto Registro de exámenes por especialidad Movimientos de inventarios en área quirúrgica Movimientos de inventarios en área de farmacia
Matrices en documento físico	Capacitaciones al personal medico Registro de programas de implementación médica Registro de atención al usuario Registro de Score Mama Despacho de productos Registro de información básica del paciente
Base de datos	Tabla clínica de pacientes Tablas de cirugías y consumos en quirófano Tablas de consumo en laboratorio Tablas de imágenes médicas Tablas de enfermería Tabla de proveedores Tabla de compras Tabla de ventas

	Tabla de clientes Tabla de usuarios

Fuente: Autor

A manera de resumen, la Figura 14 presenta la cantidad de datos extraídos para el análisis. Así mismo, la Figura 15, exhibe los usuarios que participaron para la recopilación de los datos.

Figura 14.- Resumen de datos extraídos

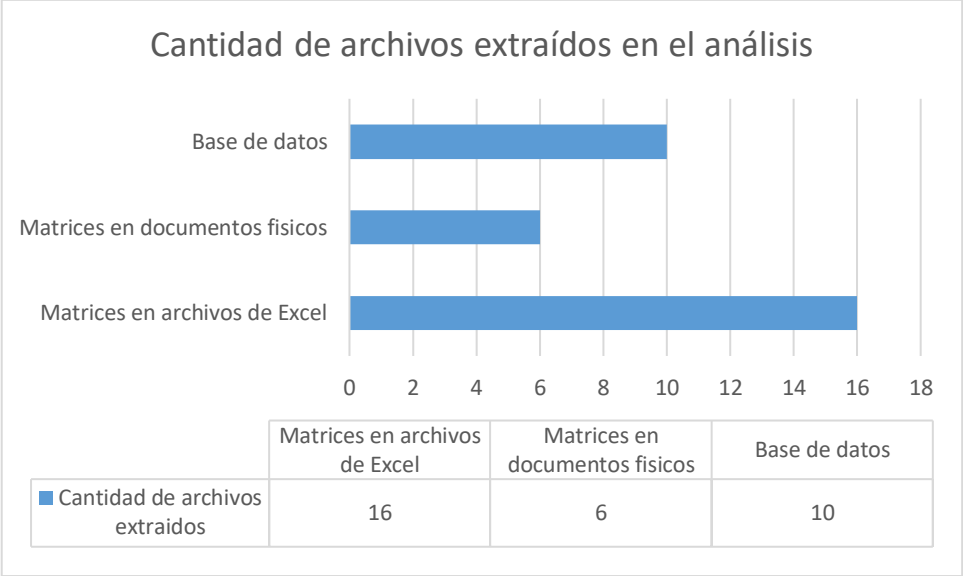
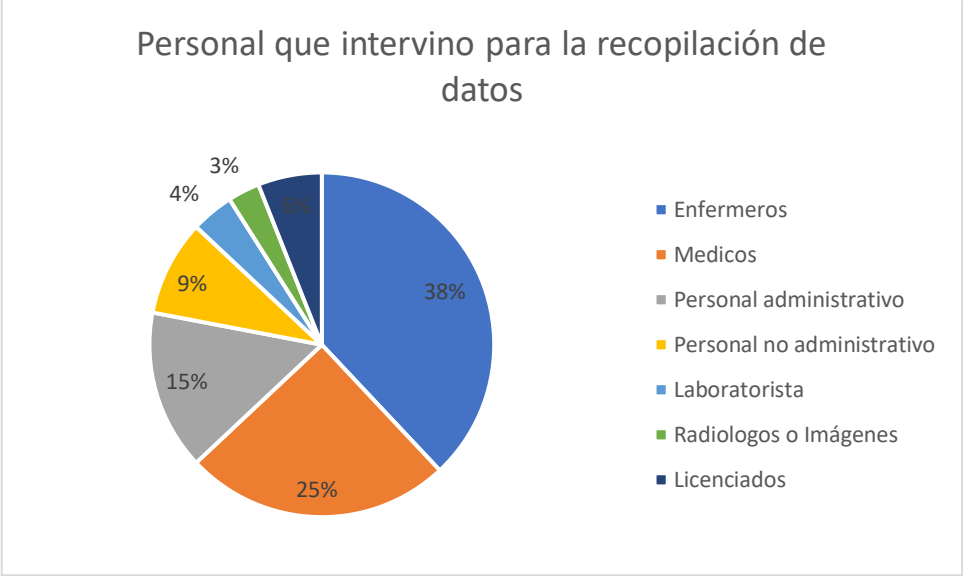


Figura 15.- Usuarios que participan para la recopilación de datos

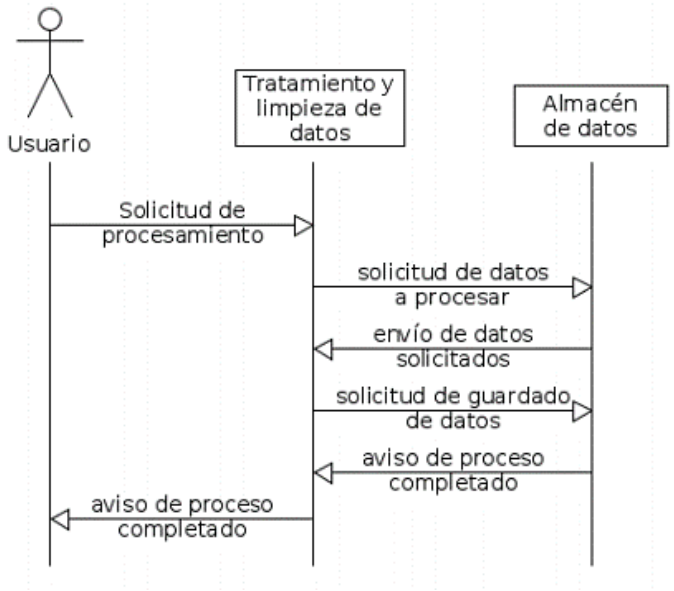


Toda la información obtenida pasará a una estructura CSV para la preparación y transformación de los datos. Cabe destacar que todo lo recopilado será necesariamente útil para los indicadores de gestión hospitalaria IGH.

3.2.3.2 FASE 2: PROCESAMIENTO DE LA INFORMACIÓN

Esta fase también conocida como tratamiento y limpieza, en la cual su labor principal es dar forma a los datos que se encuentran no estructurados a estructurados, con el fin de poder dotar valor a la información. El esquema de interacción se representa en la Figura 16:

Figura 16.- Interacción del usuario en la fase de tratamiento y limpieza



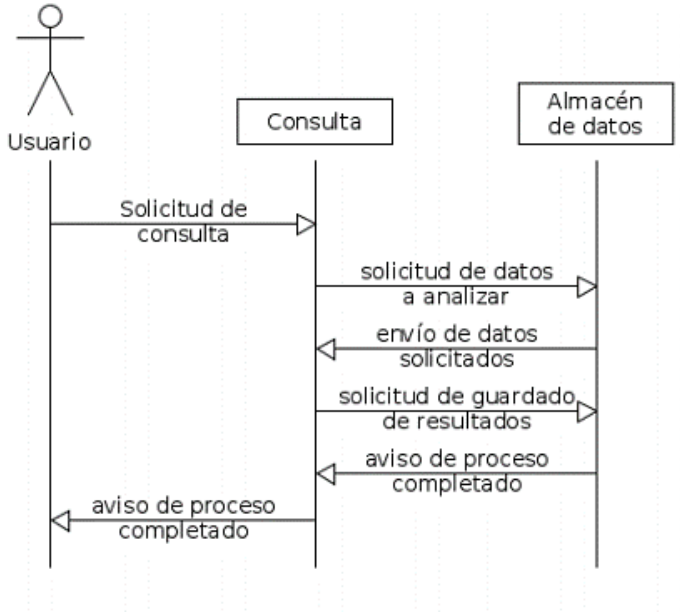
Fuente: Autor

Exclusivamente lo que el usuario realiza, es la solicitud de procesamiento de datos en el sistema, a su vez se obtiene la información alojada en el almacén de datos, se procesa, se estructura y retorna el aviso del proceso completado.

3.2.3.3 FASE 3: CARGA DE LOS DATOS O PRESENTACIÓN

Dentro de este apartado se toma en cuanta dos secciones, la forma en cómo se establece la consulta de los datos y la obtención de los resultados. La Figura 17 muestra el esquema de interacción:

Figura 17.- Interacción del usuario para realizar consulta sobre los datos

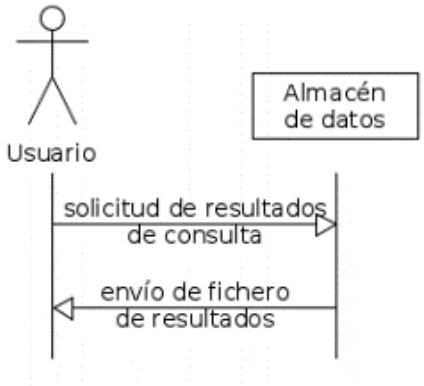


Fuente: Autor

Para el proceso de consulta el usuario realiza la solicitud al sistema, luego se procede a dar análisis y proceso en el almacén de datos. Finalmente, se procede a retornar el aviso del proceso completado, es como de forma global debe de trabajar el sistema dentro de la arquitectura propuesta.

La segunda parte consiste en obtener los datos ya procesados mediante una consulta ya generada. Lo único que debe de hacer el usuario es obtener un archivo para la presentación de información en la herramienta que se requiera. Para el caso de este proyecto en una plataforma web, en la figura 18 se muestra un dashboard los resultados ya procesados.

Figura 18.- Interacción del usuario para la obtención de los resultados



Fuente: Autor

Claro está que la alimentación del dashboard exclusivamente será mediante una base de datos la cual tendrá la información ya procesada. Es cuestión de que el administrador del sistema, alimente la información que se requiera, o se genere un proceso automático de carga. Dentro de este proyecto se contempla una carga manual a una base de datos MYSQL para la presentación de los resultados. En la sección de Anexos, se encuentra el proceso de instalación de un sistema Big Data y preparación para la lectura de datos.

3.3 APLICACIÓN DE LAS TÉCNICAS BIG DATA ANALYTICS (BDA)

Dentro del desarrollo de las técnicas de Big Data, primero se realizan los gráficos en serie temporales de cada indicador, donde se detalla el movimiento en el tiempo y su tendencia, además de otros parámetros, luego se aplica los modelos de pronósticos para su análisis. Para una mejor comprensión se detallarán los códigos de variables de cada indicador para su identificación en la Tabla 10.

Tabla 10.- Identificación de las variables por cada indicador

Código	Nombres
PPEA	Porcentaje de pacientes en espera de atención en consulta externa igual o menor a 15 días.
THMM	Tasa hospitalaria de mortalidad materna.
PHMN	Porcentaje hospitalario de mortalidad neonatal.
NPLEQ	Número de pacientes en lista de espera quirúrgica.
POCP	Porcentaje de ocupación de camas.
TDMH	Tasa de mortalidad hospitalaria.

Fuente: Autor

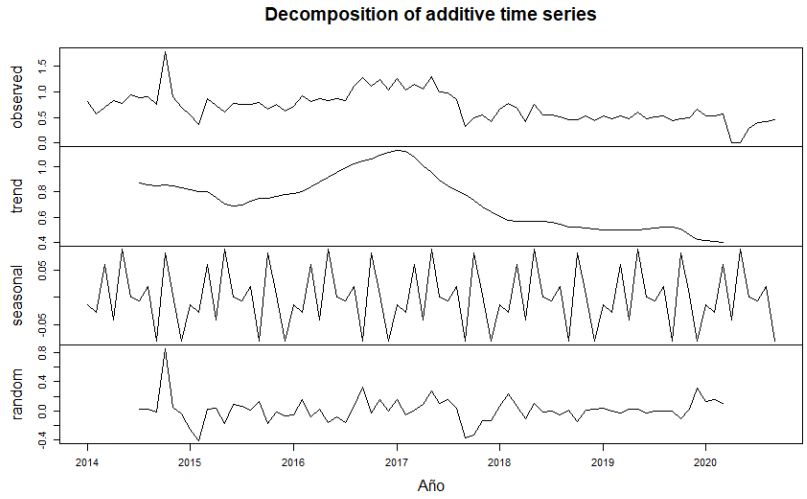
3.3.1 SERIES TEMPORALES

Todos los indicadores a analizar tienen el mismo procedimiento de realización. Para mayor información se puede consultar el Anexo 2. Al considerarse un análisis más detallado, se grafica lo observado, la tendencia, el efecto estacional y el residuo, esto permite comprender lo siguiente:

- Observado, es el resultado de la tendencia + efecto estacional + residuo.
- Tendencia, representación del movimiento a largo plazo.
- Efecto estacional, comprende las fluctuaciones periódicas.
- Residuo o error, son variaciones a corto plazo, muchas veces impredecible.

PPEA

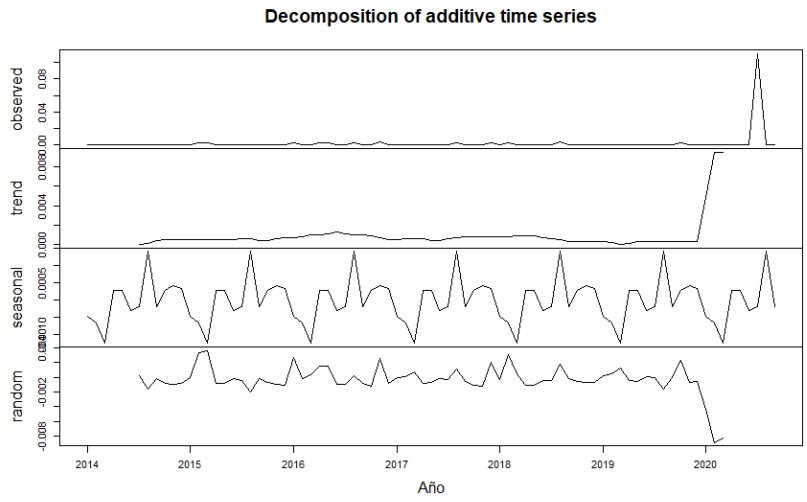
Figura 19.- Serie temporal PPEA



En la Figura 19, se puede observar que a medida que pasa el tiempo la tendencia de los pacientes que esperan ser atendido va bajando, aunque existen fluctuaciones de las cuales falta por cubrir el número de atenciones médicas según la cantidad de pacientes.

THMM

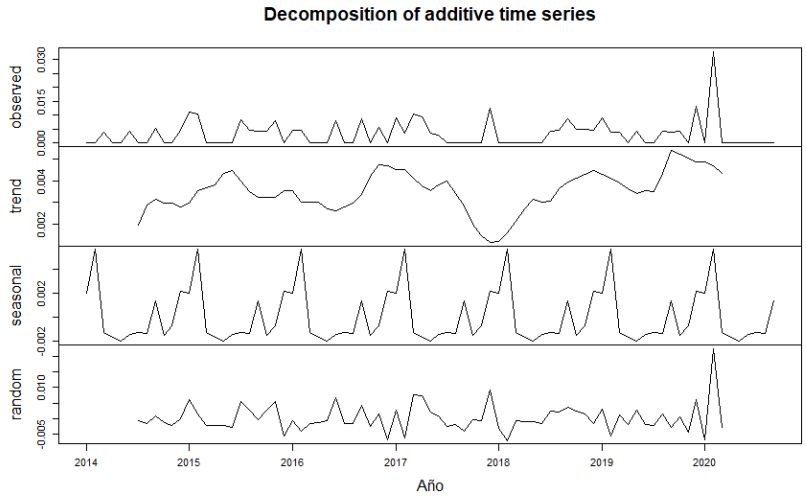
Figura 20. Serie temporal THMM



Según el análisis de la Figura 20 puede destacarse que la tasa de mortalidad es baja, aunque en el último periodo tuvo un repunte.

PHMN

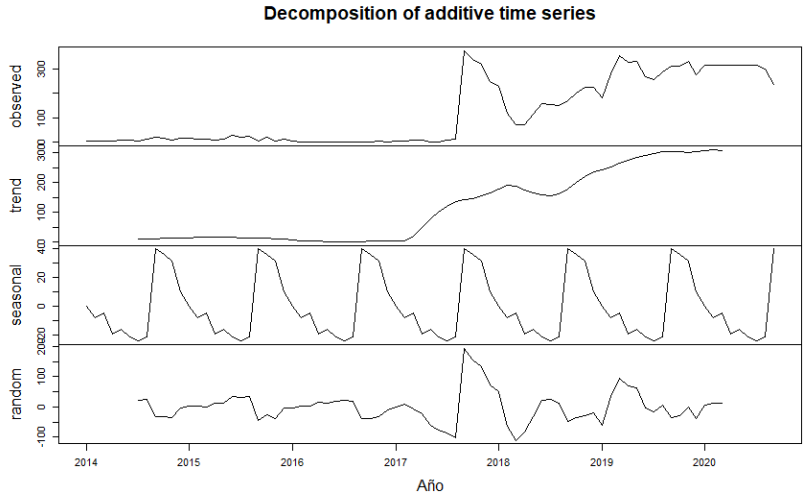
Figura 21.- Serie temporal PHMN



En relación a la Figura 21, en estas circunstancias el porcentaje de mortalidad neonatal, puede ser preocupante, aunque deberían de darse un mayor análisis de las causas de muertes en bebes, posiblemente puede ser normal, según los parámetros médicos y hospitalarios.

NPELQ

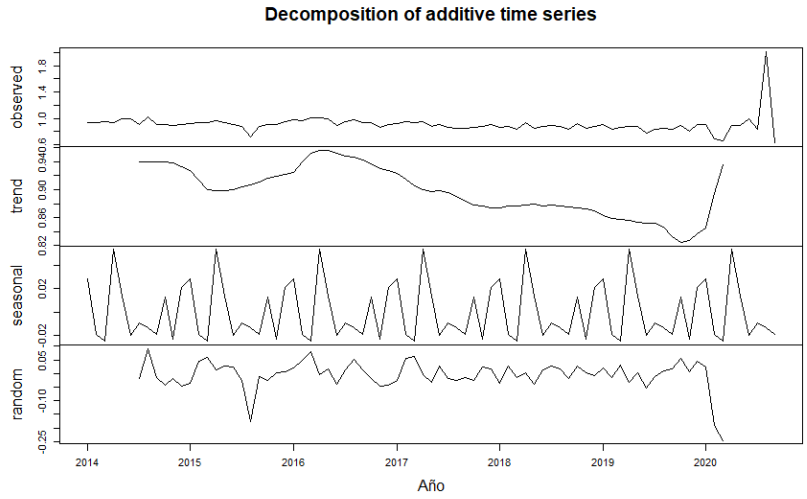
Figura 22.- Serie temporal NPELQ



Según la Figura 22, es notable que a medida que pasa el tiempo la tendencia y los efectos estacionarios para la cantidad de pacientes que esperan por alguna operación, va en crecimiento. Se requiere un cambio de infraestructura para la atención de más pacientes.

POCP

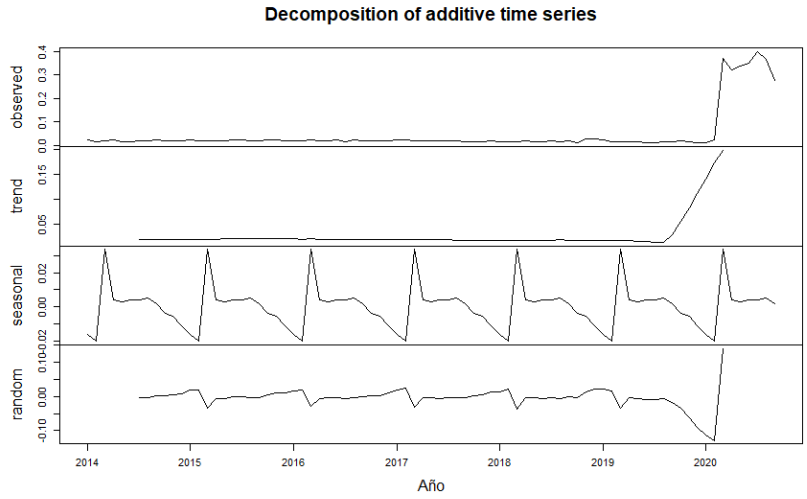
Figura 23.- Serie temporal POCP



Es importante destacar que un porcentaje mayor al 80% en ocupación camas, es muy bueno para una unidad hospitalaria. No obstante, en la Figura 22, también hay que tomar en consideración la cantidad de pacientes que dejan de ser atendidos. Se debería de aplicar alguna medida como derivación a otras unidades.

TDMH

Figura 24.- Serie temporal TDMH



La Figura 24, en efecto se puede concluir que la tasa de mortalidad hospitalaria es baja, aunque en el último periodo repuntó, posiblemente por la causa de muertes por COVID.

Una vez concluido con el análisis de los resultados de cada indicador, utilizando series temporales descompuestas, se procederá aplicar las técnicas predictivas.

3.3.2 MODELOS PREDICTIVOS

Dentro de este apartado se presentan cuatro modelos predictivos que se adaptan a los datos reales o tiene mejor respuesta. Estos son: NNETAR (Pronósticos de Series de Tiempo de Redes Neuronales), STML (Series de tiempo con múltiple estacionalidad), Holt-Winters, TBATS (Modelo de espacio en estado suavizado exponencial). La representación de los resultados en R se encuentra en el Anexo 3.

PPEA

Figura 25.- Modelo predictivo PPEA

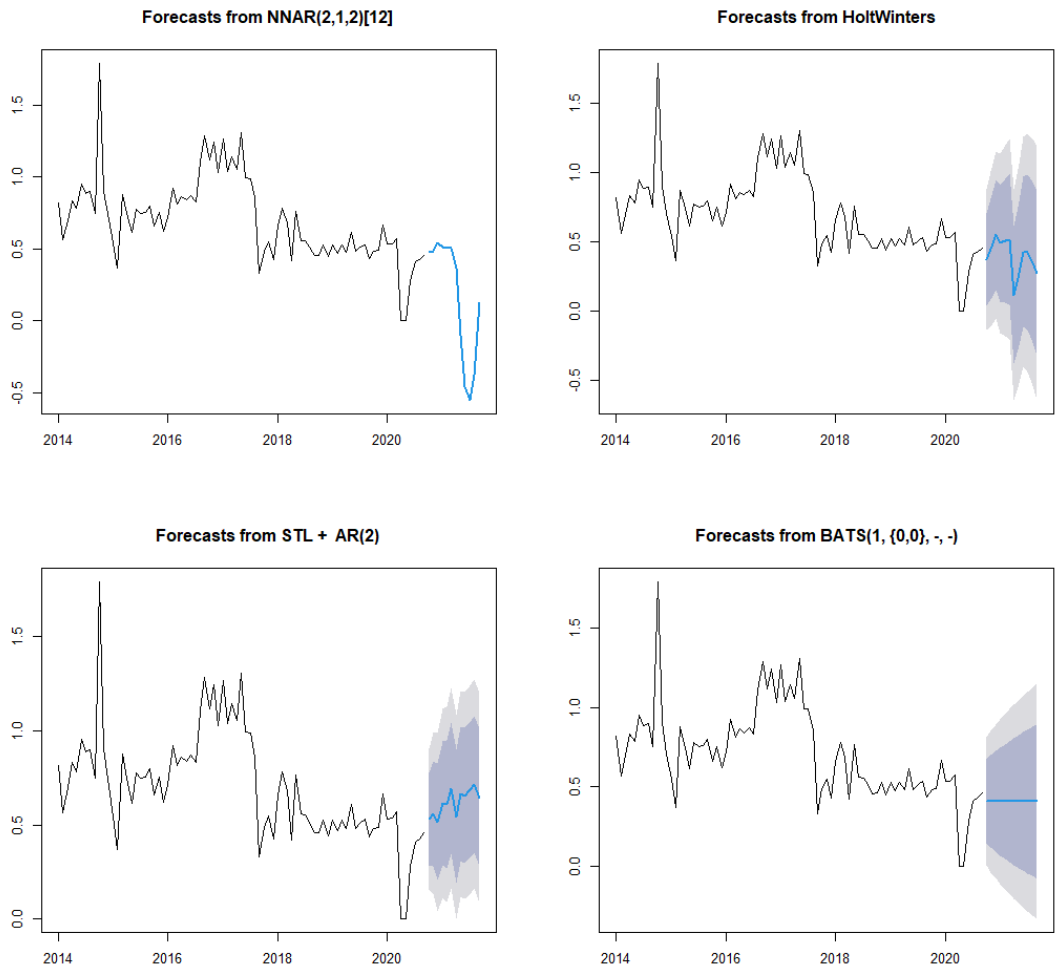


Figura 25: Al dar un vistazo a los resultados gráficos de la predicción de este indicador, se puede observar una mayor relevancia entre los modelos Holt Winters, SMLT y NNAR,

ya que tienen una mayor relación a la estacionalidad o secuencia de datos. Los demás modelos no tienen mucha relación a la realidad.

THMM

Figura 26.- Modelo predictivo THMM

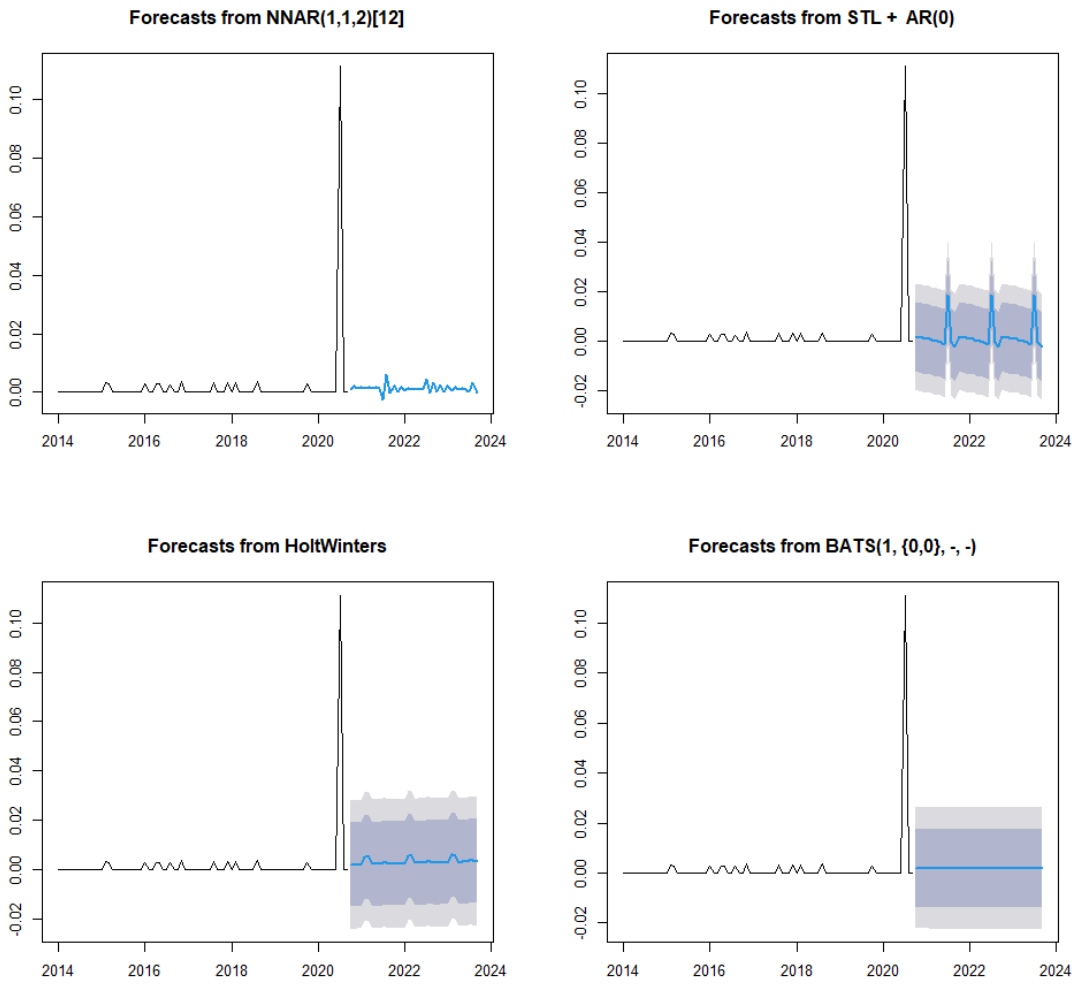


Figura 26: Este indicador expresa un resultado estable, aunque en el último periodo tuvo un pico. Más aún, los modelos Holt Winter, SMTL, NNAR, tienen mayor relación. El modelo TBATS, trata de ajustarse a los valores reales, lo cual también puede ser considerado para el análisis.

PHMN

Figura 27.- Modelo predictivo PHMN

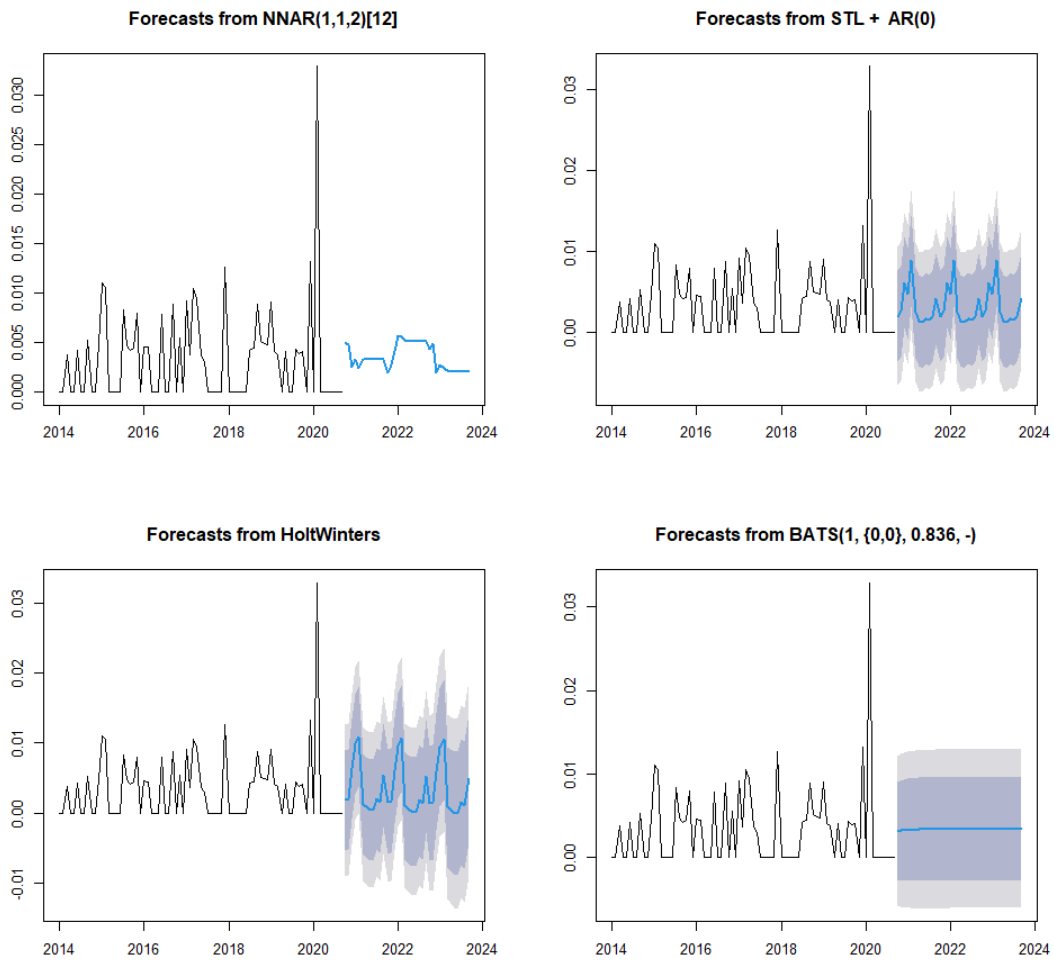


Figura 27: Entre altas y bajas se muestra este indicador, por lo que su resultado a futuro se relaciona al mismo comportamiento que tuvo en el pasado, tiene mayor relación con el modelo Holt Winters.

NPELQ

Figura 28.- Modelo predictivo NPELQ

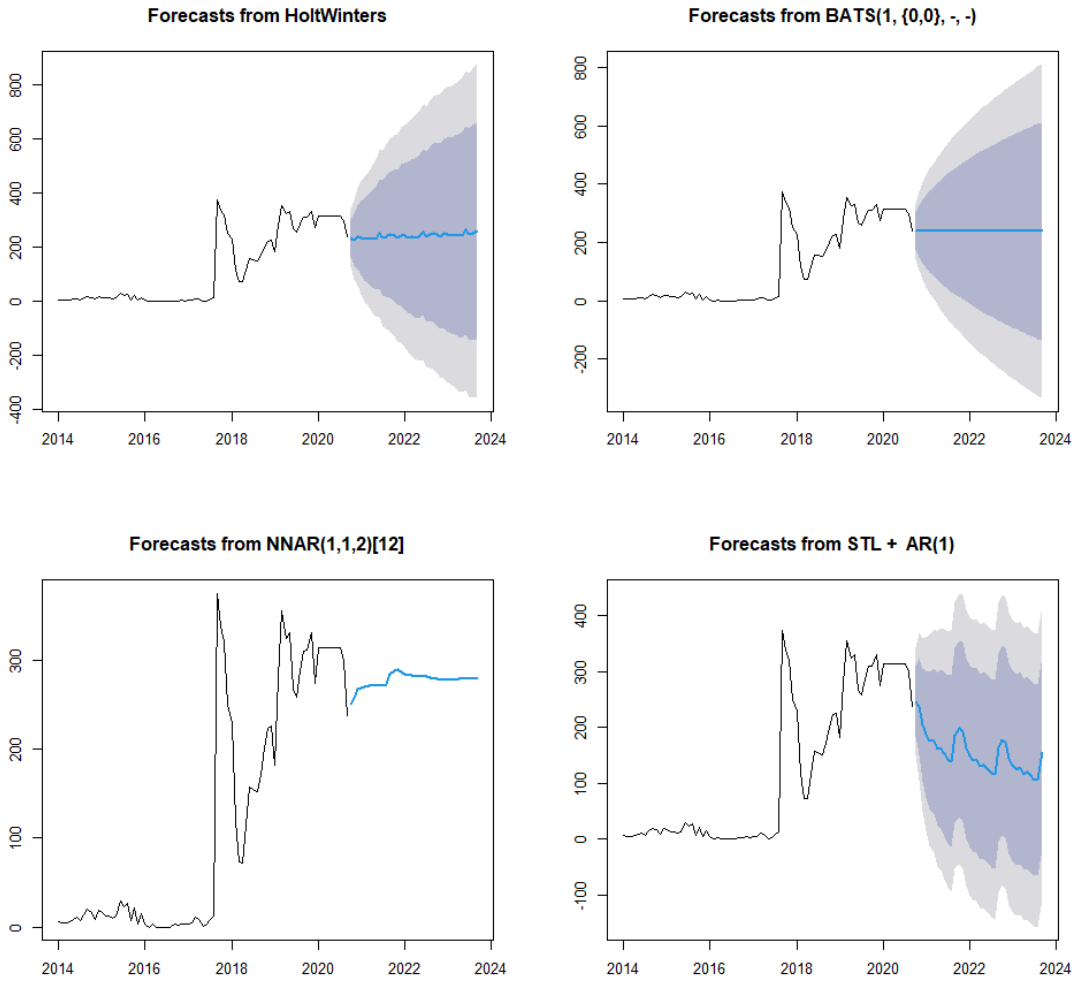


Figura 28: Se expone una forma irregular en todos los modelos, en cuanto al comportamiento de este indicador, más aún existe cierta relación entre Holt Winter y NNAR, el evaluador o análisis de información deberá de considerar los resultados pertinentes.

POCP

Figura 29.- Modelo predictivo POCP

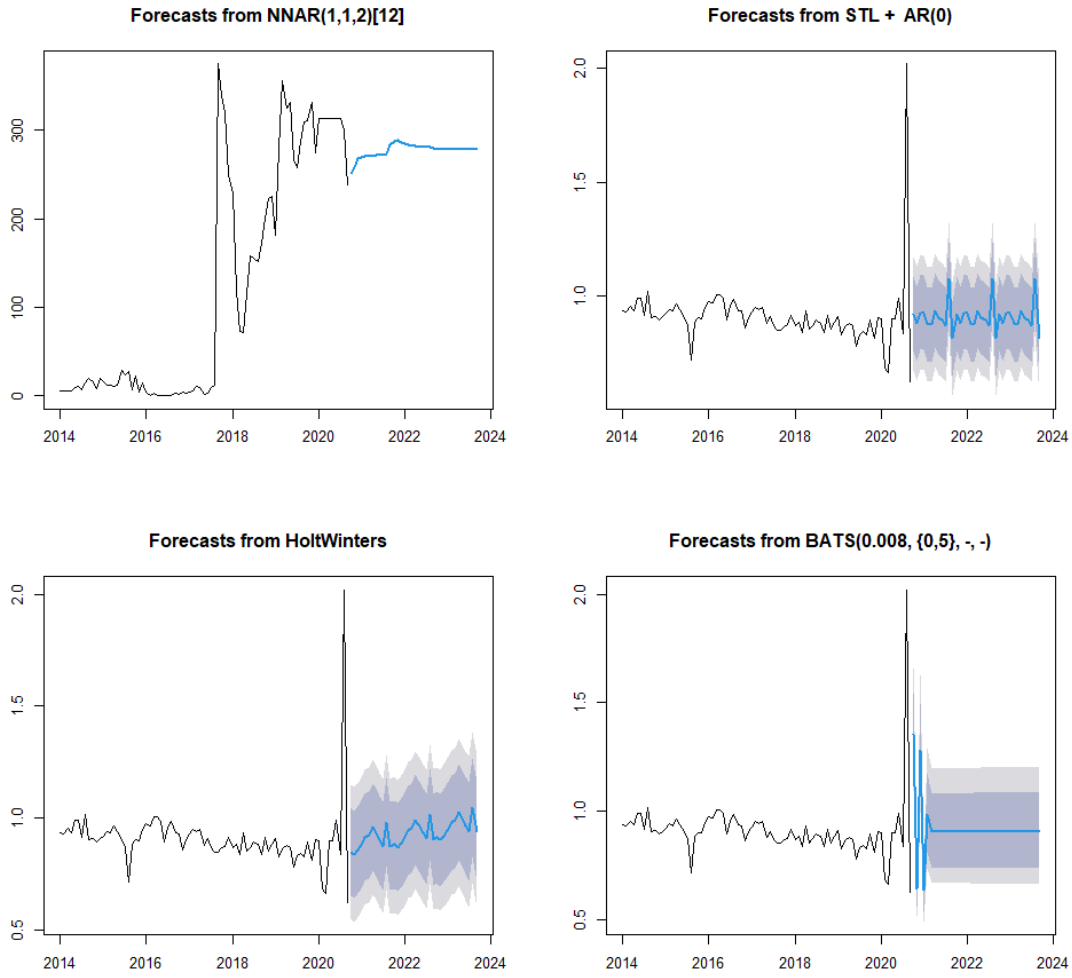


Figura 29: Para este indicador, hay una relación de predicción por estacionalidad, por lo que se puede considerar su uso según el que se requiera, se recomienda lo modelos NNAR o Holt Winter, que tienen mejor relación con los demás indicadores, los resultados de relación pueden observarse en el Anexo 3.

TDMH

Figura 30.- Modelo predictivo TDMH

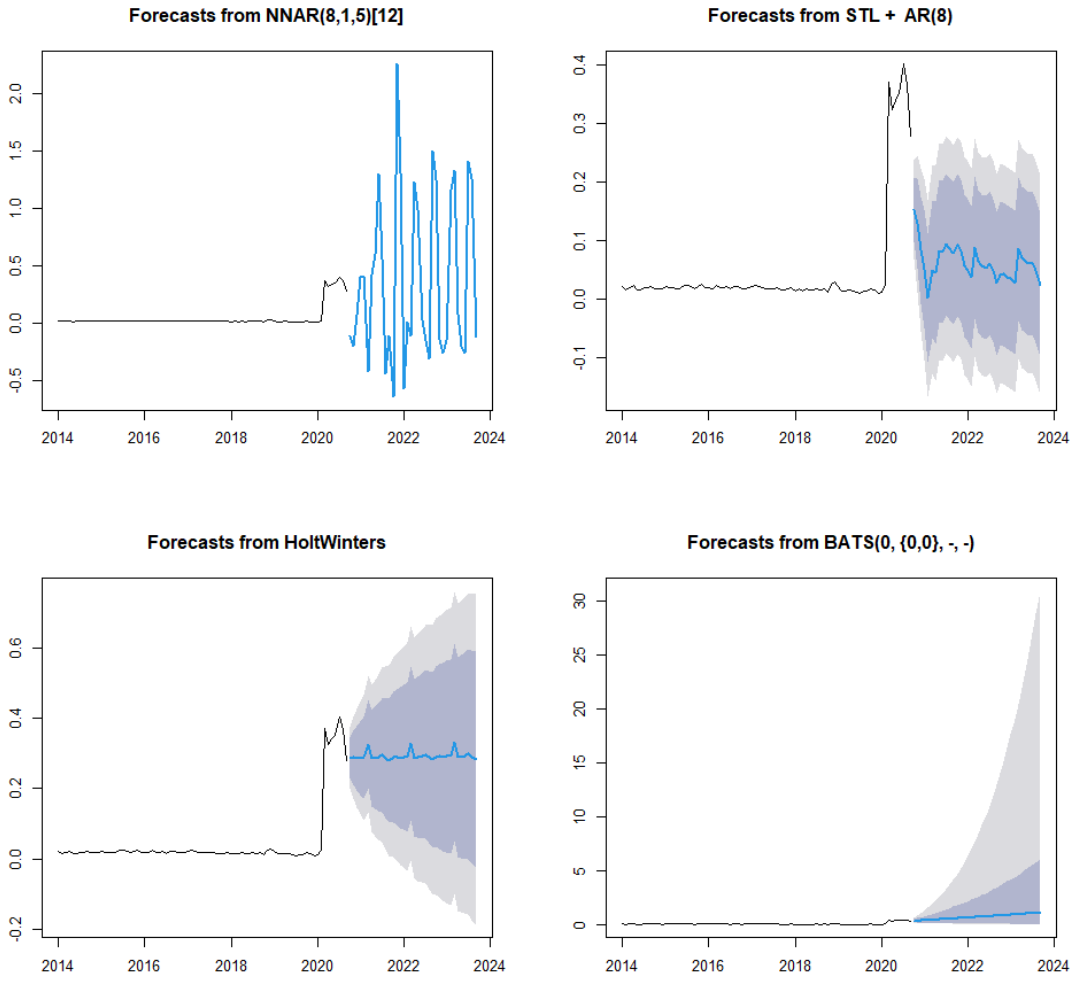


Figura 30: Los resultados de este indicador, da a conocer un comportamiento lineal muy cerca de 0, por lo que el más factible en el modelo a usar es el TBATS, ya que se acopla a la estacionalidad de los valores anteriores.

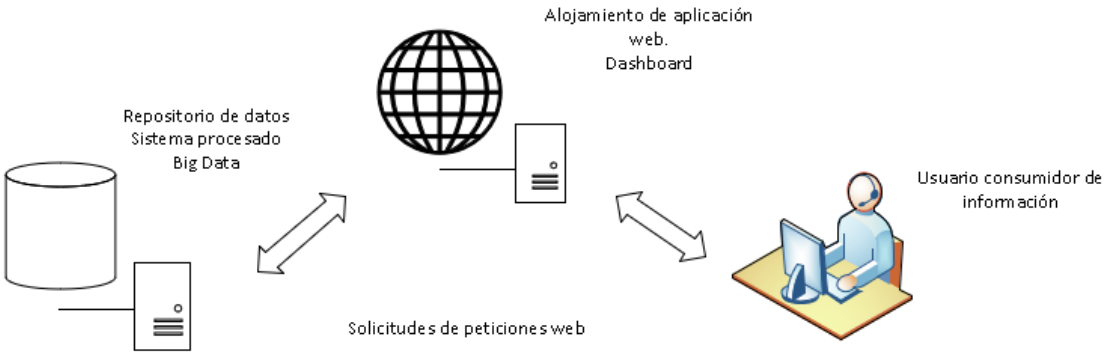
De acuerdo a los resultados presentados en el Anexo 3, se puede determinar mediante una revisión de medida de error, cual es el que tiene menor valor, para ello se toma en consideración los resultados RMSE y MAE de cada modelo en el desarrollo del próximo capítulo.

3.4 ARQUITECTURA WEB

Dentro de la propuesta del proyecto, se determinó el diseño de un prototipo web con el fin de mostrar los IGH (Indicadores de Gestión Hospitalaria), con acceso a los usuarios que utilizarán la información ya procesada por el sistema de Big Data. Cabe destacar

que hay ciertos indicadores que son de importancia para el análisis y toma de decisiones dentro de la unidad hospitalaria. La Figura 31 describe el esquema:

Figura 31.- Esquema de consumo de indicadores en plataforma web



Fuente: Autor

El prototipo propuesto es mediante la estructura cliente-servidor, además de trabajar en dos capas.

3.4.1 REQUERIMIENTOS PARA EL DESARROLLO DE LA PLATAFORMA DE CONSULTA WEB

Para el desarrollo de la plataforma web, se toma en consideración en implementar el diseño Modelo, Vista, Controlador (MVC), además debe de tener las siguientes funciones:

1. Acceso mediante usuario y contraseña
2. Acceso a menú de reportería de indicadores
    a. Pacientes en espera por atención de consulta externa
    b. Tasa de mortalidad materna
    c. Tasa de mortalidad neonatal
    d. Paciente en espera para sala quirúrgica
    e. Tasa de ocupación camas
    f. Tasa de mortalidad hospitalaria

A nivel de hardware se requiere:

1. Equipo PC con las siguientes características:
    a. Procesador core i5 de 9na generación
    b. Memoria de 8gb de ram
    c. Disco duro SSD de 1TB

A nivel de software se requiere:

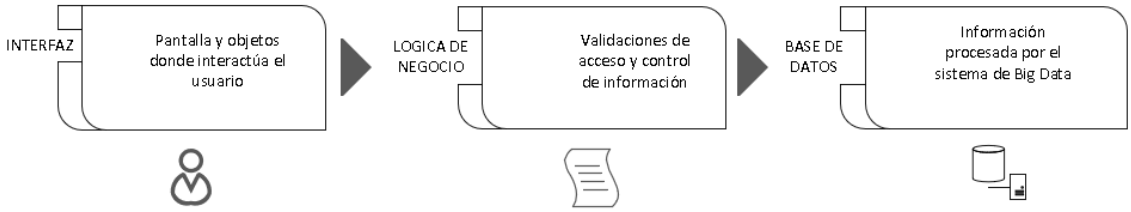
1. Aplicación web server: Laragon incluye:
    a. PHP
    b. MYSQL
2. Framework Laravel
3. Composer
4. Plantilla AdminLte, versión free

Debido a la naturaleza de la plataforma de consulta, no se requieren exigencias en cuando a capacidades o mayor infraestructura de hardware.

3.4.2 APLICACIÓN DE LA METODOLOGÍA WATCH PARA EL DESARROLLO DE SOFTWARE

Se destaca por ser un método en la ingeniería de software que se utiliza para poder planear, organizar y desarrollar una aplicación en los grupos de desarrollo de software Con ello se establece una aplicación de forma idónea para el proyecto en desarrollo. Los componentes que intervienen en esta metodología se ilustran en la Figura 32:

Figura 32.- Componentes de aplicación de consulta web



Fuente: Autor

El detalle de los componentes se explica a continuación:

1. Interfaz: Permite la interacción del usuario con la lógica de negocio del software.
2. Lógica de negocio: Establece las reglas y parámetros, validaciones y controles para el acceso de datos y uso de los mismos.
3. Base de datos: Alojamiento de la información en donde el usuario puede obtener para el análisis previo a la toma de una decisión.

3.4.2.1 INGENIERÍA DE REQUISITOS

El objetivo de esta parte es poder determinar los pasos y funciones a realizar para los actores involucrados en el desarrollo de los procesos, con el fin de establecer parámetros para la toma de requisitos. La Tabla 11 enumera los resultados.

Tabla 11.- Obtención de requisitos para el desarrollo web

Pasos	Actividades	Técnicas	Productos
Descubrimiento del requisito	• Identificación de la necesidad • Identificación de los usuarios a intervenir • Recolección de datos	• Casos de uso UML • Entrevista • Reuniones con usuarios	• Diagrama de casos de uso • Requisitos documentados
Análisis de los requisitos	• Clasificación de los requisitos • Negociación de los requisitos	• Técnicas de negociación o pactos.	• Documento de requisitos

Fuente: Autor

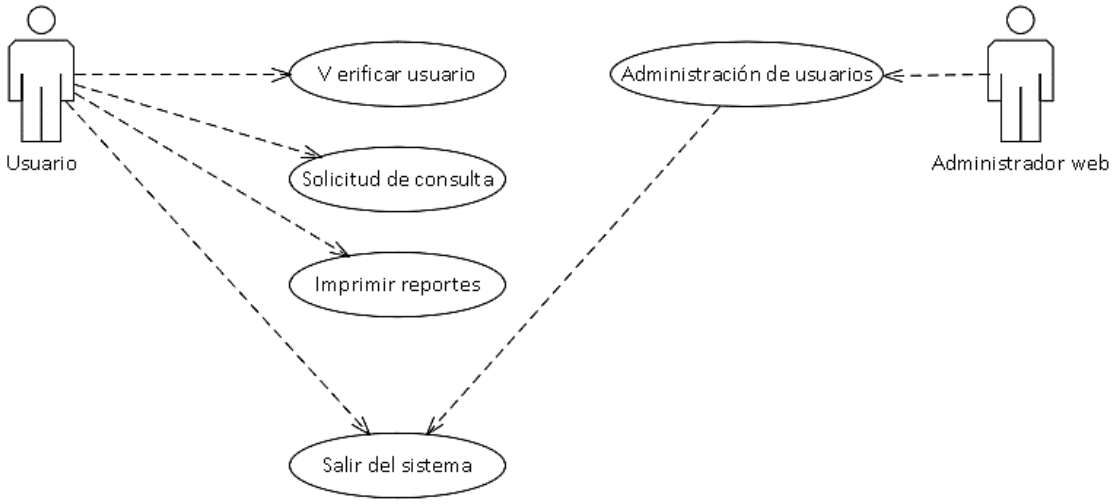
3.4.2.2 CASOS DE USO

Se establecen los siguientes casos de uso:

1. Caso de uso usuario
    a. Verificar el usuario
    b. Solicitud de consulta
    c. Imprimir reportes
    d. Salir del sistema
2. Caso de uso administrador
    a. Administración de usuarios

La Figura 33, detalla el esquema en como los usuarios interactúan con el sistema.

Figura 33.- Casos de uso de la plataforma web



Fuente: Autor

DESCRIPCIÓN DE LOS CASOS DE USO

Caso de Uso	Verificar usuario
Objetivos	Permitir verificar antes del ingreso al sistema, con el fin de evitar accesos no autorizados
Actores	Usuarios registrados
Precondiciones	Tener acceso al sistema
Pasos	Solicitud realizada por el usuario

Caso de Uso	Solicitud de consulta
Objetivos	Mostrar la información referente a los indicadores de gestión hospitalaria IGH, que fueron ya procesados por el sistema de Big Data.
Actores	Usuarios registrados
Precondiciones	Tener acceso al sistema
Pasos	Solicitud realizada por el usuario

Caso de Uso	Imprimir reportes
Objetivos	Poder obtener información impresa y sea entregada a gerencia a buen uso del usuario.

Actores	Usuarios registrados
Precondiciones	Tener acceso al sistema
Pasos	Solicitud realizada por el usuario

Caso de Uso	Administración de usuarios
Objetivos	Gestionar el uso de la plataforma web, mediante el registro, actualización, eliminación de usuarios.
Actores	Administrador web
Precondiciones	Tener acceso al sistema como administrador
Pasos	Solicitud realizada por el administrador

Caso de Uso	Salir del sistema
Objetivos	Eliminar la sesión de acceso del sistema, para que un usuario no autorizado pueda ingresar
Actores	Usuarios registrados
Precondiciones	Tener acceso al sistema
Pasos	Solicitud realizada por el usuario

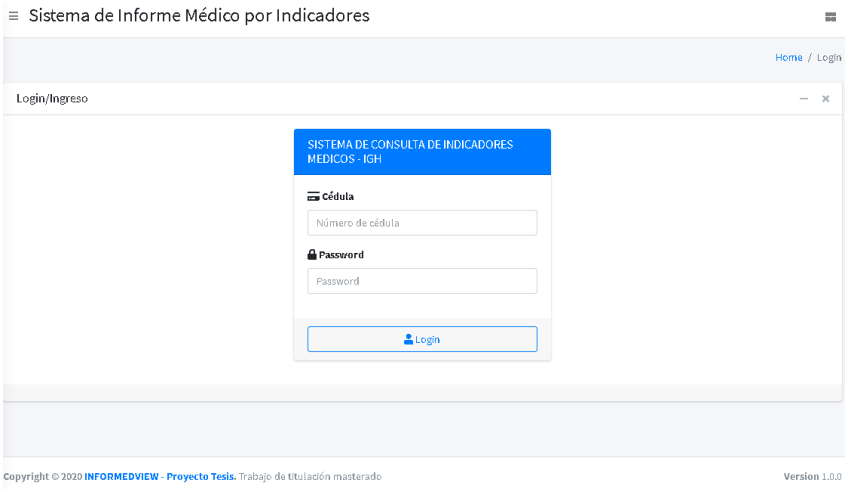
3.4.2.3 DISEÑO WEB

El objetivo de la aplicación web por ser un sistema de consulta, contendrá:

- Login para usuarios
- Menú de opciones
    - IGH
        - Indicadores
    - Administración de usuarios
- Menú Salir

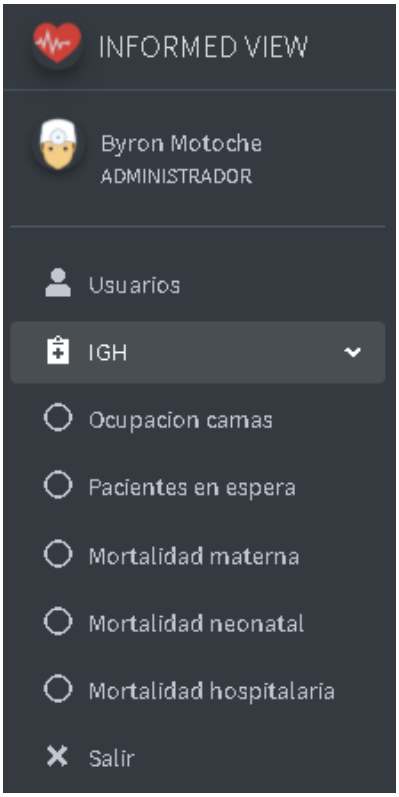
3.4.2.3.1 LOGIN DE USUARIOS

Está sección permite la identificación de los usuarios registrados en el sistema y quienes solo pueden hacer consultas de los datos IGH.



3.4.2.3.2 MENU DE OPCIONES

El menú de opciones dará movilidad a los usuarios a las diferentes pantallas que compone el sistema de consulta. Cada usuario tendrá su respectivo permiso de acceso.

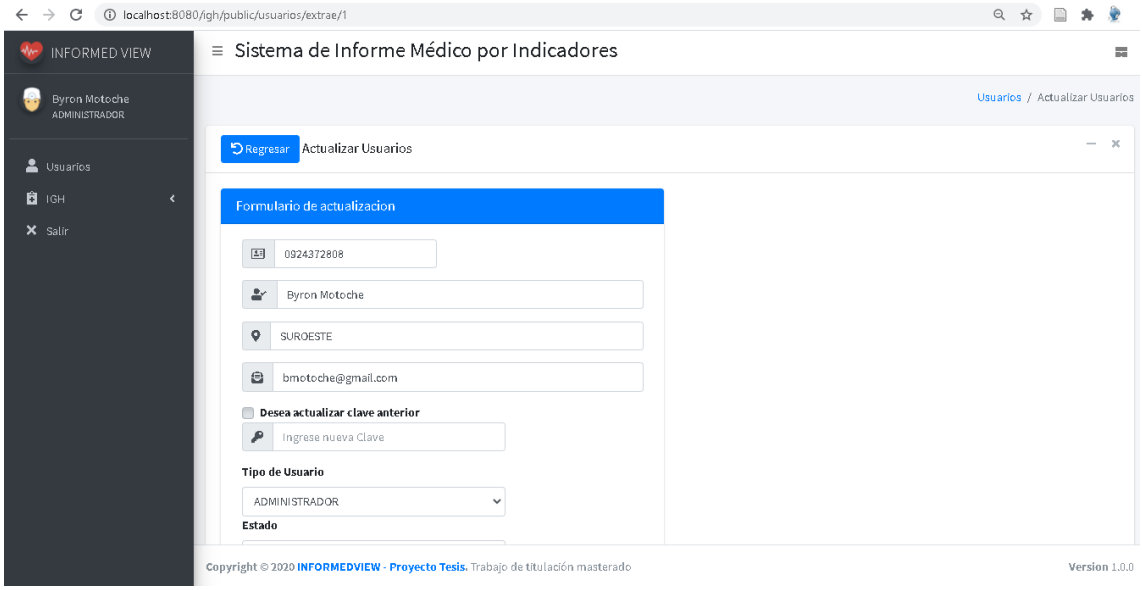


3.4.2.3.3 PANTALLA DE BIENVENIDA

Al iniciar el sistema de consulta, el usuario podrá observar una pantalla de bienvenida la cual consta de una sección de menú y una sección informativa.

3.4.2.3.4 ADMINISTRACIÓN DE USUARIOS

Solo los usuarios que tiene acceso a las opciones de configuración y administración podrán hacer uso de esta pantalla, con el fin de poder dar mantenimiento a los datos de los usuarios.



3.4.2.3.5 CONSULTA DE INDICADORES OCUPACIÓN CAMA

La información procesada por el sistema Big Data, da opción a poder consultar los datos mediante graficas que representan la información de forma más comprensible para su interpretación.

Las figuras pueden variar según las necesidades de los usuarios, en este caso se puede observar la información ya procesada en cuanto a ocupación de camas se refiere, esto muestra, la utilidad del sistema para tomar una decisión en base al parámetro de ocupación de camas dentro del centro hospitalario.

El desarrollo de este capítulo ha sido sustancial para conocer los resultados obtenidos mediante la aplicación de Big Data dentro del Hospital. Así, por ejemplo, las fases de su desarrollo y procesamiento de datos y por ende la forma de cómo se visualizarán. Además, el uso de modelo de datos que permitan evaluar cual es el que mejor se adapta para tomarlo como referencia para otros indicadores que se puedan crear en el tiempo. En el próximo capítulo se analizará, qué modelo es el más adecuado y cual generó menor error comparativo entre lo real y lo ficticio.

CAPÍTULO 4. DISCUSIÓN DE LOS RESULTADOS

4.1 ANÁLISIS Y DISCUSIÓN DE LOS RESULTADOS

Para el análisis de datos se usaron modelos matemáticos que permitan obtener resultados comprensibles y válidos para comparar que modelo es el más adecuado para su uso, para ello se tomó en consideración técnicas predictivas como NNETAR, STML, HOLT WINTER, TBATS. Estos modelos fueron útiles para comprender como se pueden comportar los datos en periodos futuros. Con el fin de adecuar al uso de uno o dos modelos de datos, para ello se destaca el que menor error residual tenga en sus resultados, se tomó en cuenta el RMSE (Error cuadrático medio) y MAE (error absoluto medio), estos índices permiten evaluar de forma exacta cuales son los modelos más adecuados a usar en la estrategia de Big Data Analytics. El RMSE es la métrica que indica que tan lejos está el valor pronosticado con el valor real en un análisis de regresión, mientras que el MAE establece la diferencia entro los dos vectores el real y el ficticio. En la minería de datos se dio énfasis al uso de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) que es la más usada en cuanto al área de Salud se refiere, tomando en consideración que cada fase cumple con las etapas de ejecución en las técnicas ETL.

Como resultado se obtuvo que el modelo NNETAR en relación a RMSE es de 0.0033164 y TBATS en relación a MAE es de 0.00211. Estos son los más adecuados ya que representan un menor índice y mayor acercamiento a la realidad. Para dicho análisis se tomó en consideración un periodo de pronósticos en 12 meses a futuro, teniendo datos desde enero 2014 a octubre 2020, ahora existe la posibilidad de que existan otros modelos que puedan adaptarse mayormente e inclusive establecer modificación de parámetros de entradas que permitan ajustar mayormente a la realidad, por el momento para el desarrollo de este trabajo dicho modelos se encuentran aprobados para su uso.

4.2 INTEPRETACIÓN Y REDACCIÓN DE LOS RESULTADOS

Para comprender de forma cuantitativas los resultados obtenidos para el análisis de validación del modelo adecuado para su uso dentro del proyecto, en la Tabla 12 se describen las métricas de medición de errores en los modelos:

Tabla 12.- Métricas de medición de errores en modelos

IGH	Modelo	Métricas de error						
		ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
PPEA	NNETAR	-1.00E-05	**0.1276**	**0.0943**	-Inf	Inf	0.3583	0.0360
	STLM	-0.0031	0.1869	0.1303	-Inf	Inf	0.4963	-0.021
	H.WINTERS	0.01885	0.25842	0.18186	-Inf	Inf	0.6903	0.1081
	TBATS	-0.0108	0.2065	0.1353	-Inf	Inf	0.5137	0.0494
THMM	NNETAR	5.14E-08	0.0132	0.0034	NaN	Inf	14.056	-0.008
	STLM	-9.81E-20	**0.01089**	0.00309	NaN	Inf	12.522	0.00650
	H.WINTERS	0.00073	0.01323	0.0030	NaN	Inf	12.488	-0.0313
	TBATS	0.00143	0.01235	**0.00211**	-Inf	Inf	0.8557	-0.0341
PHMN	NNETAR	-4.29E-06	0.00483	0.0033	-Inf	Inf	0.6704	0.01141
	STLM	-1.13E-19	**0.00436**	0.00318	-Inf	Inf	0.64548	-0.0989
	H.WINTERS	-0.0006	0.0055	0.00396	NaN	Inf	0.8037	-0.0505
	TBATS	-0.00026	0.00460	**0.00308**	NaN	Inf	0.62579	0.0129
NPLEQ	NNETAR	-0.0182	48.131	22.8	-Inf	Inf	0.3024	0.0299
	STLM	24.921	**46.045**	22.96	-Inf	Inf	0.304	0.020
	H.WINTERS	2.73	53.288	24.658	NaN	Inf	0.3265	-0.0042
	TBATS	2.784	48.92	**19.992**	-Inf	Inf	0.2647	-0.0042
POCP	NNETAR	0.00035	**0.1155**	**0.0569**	-1.245	5.922	0.7747	0.0039
	STLM	-2.74E-17	0.12682	0.0587	-1.335	61.951	0.80041	-0.0756
	H.WINTERS	0.02688	0.1539	0.0785	1.322	8.215	1.068	-0.0611
	TBATS	-0.0100	0.1199	0.0683	-2.331	7.342	0.931	-0.035
TDMH	NNETAR	8.99E-06	**0.00331**	**0.0025**	-3.933	14.301	0.0668	0.2781
	STLM	0.00187	0.03832	0.01491	-22.261	52.854	0.39612	-0.0217
	H.WINTERS	0.00421	0.04424	0.01115	-20.096	20.321	0.29629	0.0131
	TBATS	-0.0008	0.04460	0.01188	-89.112	20.675	0.3155	-0.2667

Fuente: Autor

Nota: -inf, inf, NaN, son referencia de valores que no pudieron ser calculados, por tener en sus datos 0.

En la Tabla 13 se muestran los resultados que cada modelo ha tenido según el indicador evaluado, por ejemplo, se tomará el indicador PPEA:

Tabla 13.- Resumen de métricas de cada modelo

Error Meseaure	PPEA			
	NNETAR	STML	WINTERS	TBATS
ME	-1.00E-05	-0.003132	0.018855	-0.010836
RMSE	0.1276846	0.1869548	0.25842	0.2065308
MAE	0.0943668	0.1307153	0.1818096	0.1353142
MPE	-Inf	-Inf	-Inf	-Inf
MAPE	Inf	Inf	Inf	Inf
MASE	0.3583021	0.4963138	0.6903142	0.5137754
ACF1	0.0360217	-0.021258	0.1080041	0.0494632

Fuente: Autor

Si bien es cierto que el objetivo de análisis es solo las medidas por error, RMSE y MAE, el sumario muestra el procesamiento del modelo en R que detalla otros índices que pueden ser usado a criterio del evaluador, para este caso se centrara la validación en los índices ya establecidos. Para tener en cuenta que modelo es el más adecuado se tomó en consideración el que está más cerca al valor real según los pronósticos generados. En este caso se detecta que el modelo NNETAR es el más adecuado ya que es el que presenta menor valor, este mismo análisis es aplicado a cada indicador, destacando que modelo e índice tiene mayor cercanía a la realidad.

Los resultados condensados de cada modelo y el de menor valor, se puede deducir en un análisis en conjunto, cual o cuales son los más indicados para su uso. Para ello, se recopilo cada indicador con su índice más bajo por modelo de datos y se obtuvo la Tabla 14.

Tabla 14.- Selección del modelo más adecuado

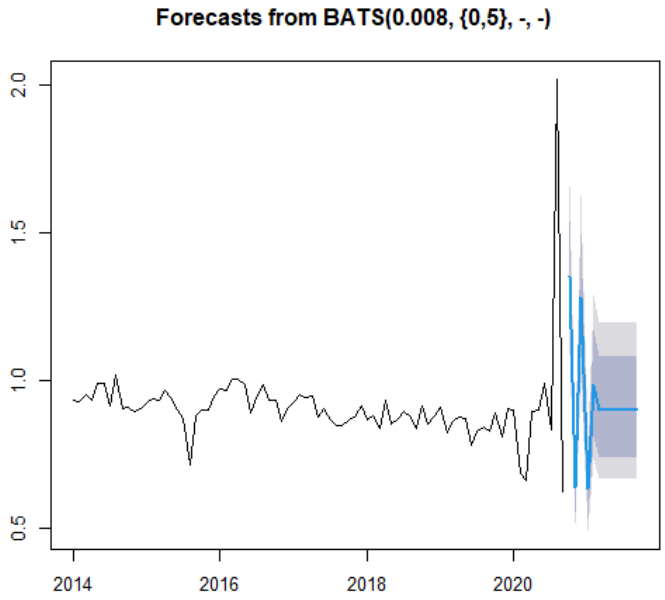
Error	**PPEA**	**THMM**		**PHMN**		**NPLEQ**		**POCP**	**TDMH**
Meseaure	**NNETAR**	**STLM**	**TBATS**	**STLM**	**TBATS**	**STLM**	**TBATS**	**NNETAR**	**NNETAR**
RMSE	0.1268	0.0010	0.0012	0.0043	0.0046	46.0452	48.9228	0.11558	0.0033
MAE	0.0943	0.0030	0.0021	0.0031	0.0030	22.9692	19.9921	0.0569	0.0025

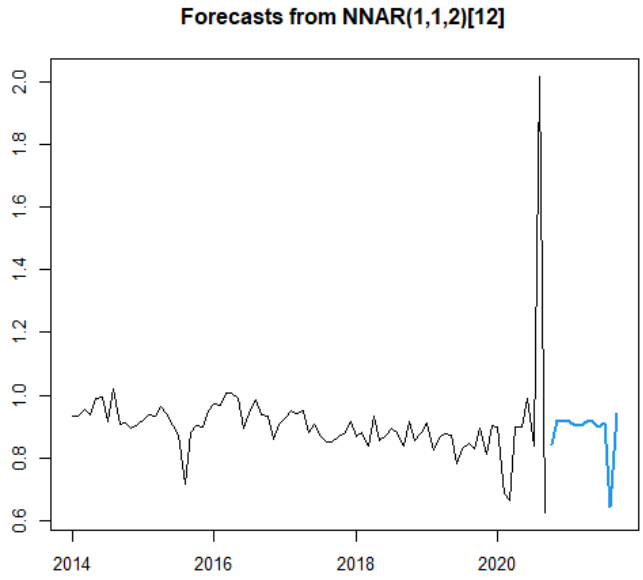
Fuente: Autor

Se puede notar que los índices de menor rango o mucho menores para el modelo TBATS (BATS) en relación a MAE es 0.00211 y NNETAR (NNAR) en relación a RMSE con 0.0033164. Con ello se determina que estos dos modelos de datos pueden ser usados para el análisis de pronósticos. Aun así, se debe tomar en cuenta que para mejorar la forma de cálculo de los modelos se puedan ajustar más para garantizar mayor fiabilidad de acercamiento a la realidad. En la Figura 34, se muestran los dos modelos

aprobados según el análisis previo, para ello se tomará de ejemplo el indicador POCP (Promedio de ocupación camas).

Figura 34.- Modelos aprobados ajustados a la realidad





Cualquier de estos dos modelos se pueden usar, aunque los de menor rango de error fueron TBATS (BATS) y NNETAR (NNAR) en todo caso queda a criterio del evaluador. Se recuerda que todo pronóstico no siempre es real ya que influye otros factores externos como internos.

CONCLUSIONES

- El estudio de las diferentes fuentes bibliográficas para la comprensión de la Big Data, fue un eje fundamental para poder establecer los parámetros adecuados en el desarrollo y arquitectura aplicable para este proyecto. Si embargo, en nuestro país falta mucho por investigar este campo.
- Las técnicas de predicción dentro de Big Data Analytics han permitido conocer los patrones y comportamiento de los datos analizados. Con el fin de adecuar el mejor resultado y modelo a utilizar, se han evaluado cuatro modelos, lo que ha permitido determinar el que mejor resultado ha ofrecido mediante el análisis de los índices de error en pronósticos.
- Para el análisis de datos, se usó la metodología CRISP-DM. Fue una labor inmensa recopilar la información. Para poder identificar correctamente los datos y poder tenerlos listos para su procesamiento y visualización, fue de vital importancia el aporte de todos los recursos.
- El uso de las técnicas de modelamiento de datos dependerá mucho del mínimo error que se obtenga en sus resultados. En este estudio, se determinó que TBATS (BATS) y NNETAR (NNAR) son los de menor error en la predicción. Sin embargo, es criterio del evaluador o analista de la información, el poder seleccionar el modelo a usar.
- A nivel de instituciones públicas médicas, se observó que se da menos interés al desarrollo de tecnologías para mejorar los procesos. Por lo tanto, este trabajo de tesis, es la pauta para el desarrollo de otras implementaciones que aporten con necesidades cotidianas dentro de estas instituciones, más aún cuando cada día aparecen nuevos datos por analizar.

RECOMENDACIONES

- Es de vital importancia determinar correctamente las fuentes de información mediante el uso de técnicas de recopilación de datos en software, para la comprensión y entendimiento del tema. Aun cuando se pretende seguir una investigación más profunda en el área de Big Data.
- Cuando se trata de establecer mejorías o cambios en la forma de cómo trabajan las personas, se presentan resistencias al cambio. Más aún cuando se piensa que será un problema tener algo automatizado o la falta de interés por mejorar los procesos, por ello el aporte de cada uno de los integrantes es fundamental para que el proyecto, salga a flote. Es de vital importancia que el personal se instruya en los nuevos modelos tecnológicos que benefician a los procesos internos en el campo de la Salud.
- El desarrollo de este proyecto establece los beneficios de la Big Data, para mantener los datos de forma más centralizada y dar mayor acceso para su procesamiento. La existencia de modelos matemáticos para la interpretación de datos a futuro siempre debe ir acompañado de otros análisis.
- Es válido destacar que no existe un modelo establecido para determinar una arquitectura de Big Data, por lo que es adecuado tomar en consideración el que más se acopla al proyecto de implementación, sumando beneficios y ventajas en el procesamiento de los datos.
- El resultado de un modelo predictivo es la pauta para la toma de una decisión ante un análisis previo, más no es el resultado de lo que se debe de hacer o no hacer. No es recomendable tomarlo plenamente como algo seguro, siempre debe llevarse junto a otros criterios que el analista deba de considerar.

TRABAJOS FUTUROS

El desarrollo de esta tesis abre el camino a la implementación o estudio de analítica utilizando Big Data, de forma general en las demás instituciones públicas de Salud, qué, de una manera u otra, requieren de este tipo de soluciones para mantenerse a la vanguardia como en otros países. Incluso puede ser útil como hilo de investigación para un trabajo de doctorado que especifique directamente las necesidades a cubrir en cuanto a un área hospitalaria, creando algo innovador que permita ser útil para todos.

BIBLIOGRAFÍA

[1] M. Tascón, «Introducción: Big Data. Pasado, presente y futuro,» *Fundación Dialnet,* vol. 1, nº 95, pp. 47-50, 2013.

[2] J. L. Samprieto, «Transformación Digital de la Industria 4.0,» 28 Agosto 2020. [En línea]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=7554338. [Último acceso: 14 Octubre 2020].

[3] S. Naya, «Nuevo paradigma de big data en la era de la industria 4.0,» 31 Mayo 2018. [En línea]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=6489657. [Último acceso: 18 Octubre 2020].

[4] T. Hwang, «World Economic Forum,» Cómo los datos grandes y abiertos pueden transformar América Latina, 14 Marzo 2018. [En línea]. Available: https://www.weforum.org/agenda/2018/03/latin-america-smart-cities-big-data/. [Último acceso: 25 Octubre 2020].

[5] J. Gubbioli, «El valor de la información y el Big Data,» 15 Agosto 2016. [En línea]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=5640378. [Último acceso: 18 Septiembre 2020].

[6] S. K. S. ,. S. y. K. Sabyasachi Dash, «Big data en salud: gestión, análisis y perspectivas de futuro,» Revista de Big Data, 19 Junio 2019. [En línea]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0217-0#citeas. [Último acceso: 01 Octubre 2020].

[7] Secretaria Técnica Panifica Ecuador, «Directrices para la Planificación del GPR,» GPR Gobierno por Resultado, 15 Enero 2018. [En línea]. Available: https://soportegpr.administracionpublica.gob.ec/otrs/public.pl?Action=PublicFAQZoom;ItemID=142;ZoomBackLink=QWN0aW9uPVB1YmxpY0ZBUVNlYXJjaDtTdWJhY3Rpb249U2VhcmNoO0Z1bGx0ZXh0PXBhYzIwMjA7%0AU29ydEJ5PUZBUUlEO09yZGVyPURvd247U3RhcnRIaXQ9MQ%3D%3D%0A;. [Último acceso: 15 Octubre 2020].

[8] ESPOL, «El Boom del Big Data,» Revista cientifica, 08 Septiembre 2017. [En línea]. Available: http://www.espol.edu.ec/sites/default/files/docs_escribe/El%20Boom%20del%20Big%20Data.pdf. [Último acceso: 16 Octubre 2020].

[9] M. B. Vega, «Tecnología de la información y comunicación sanitaria,» *Revista PUCE,* vol. 1, nº 102, pp. 271-290, 3 Mayo 2016.

[10] L. S. Carlos Guaipatin, «Análisis del sistema nacional de innovación,» 15 Octubre 2014. [En línea]. Available: https://www.epn.edu.ec/wp-content/uploads/2017/03/CTI-MON-Ecuador-An%C3%A1lisis-del-Sistema-Nacional-de-Innovaci%C3%B3n.pdf. [Último acceso: 28 Octubre 2020].

[11] B. Kitchenham y S. Charters, «Guidelines for performing Systematic Literature Reviews in Software Engineering,» 15 Octubre 2007. [En línea]. Available: https://userpages.uni-koblenz.de/~laemmel/esecourse/slides/slr.pdf. [Último acceso: 26 Agosto 2020].

[12] V. Mayer y K. Cukier, Big Data. La revolución de los datos, Madrid: Turner Publicaciones, 2013.

[13] S. Sicular, «Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s,» Gartner, 02 Abril 2013. [En línea]. Available: https://blogs.gartner.com/svetlana-sicular/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/. [Último acceso: 22 Octubre 2020].

[14] Red Hat, «El concepto del big data,» 15 Enero 2017. [En línea]. Available: https://www.redhat.com/es/topics/big-data. [Último acceso: 23 Octubre 2020].

[15] M. Labbe, «Big Data: Nuevos desafíos en materia de libre competencia,» Scielo, 14 Junio 2020. [En línea]. Available: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0719-25842020000100033&lang=es. [Último acceso: 26 Octubre 2020].

[16] Bundeskartellamnt, «Competition Law and Data,» 10 Mayo 2016. [En línea]. Available: https://www.bundeskartellamt.de/SharedDocs/Publikation/DE/Berichte/Big%20Data%20Papier.pdf;jsessionid=733957F598A9E665658AE12CAB1615F2.2_cid387?__blob=publicationFile&v=2. [Último acceso: 31 Octubre 2020].

[17] V. Dhar, «Data Science and Prediction,» 12 Mayo 2012. [En línea]. Available: https://archive.nyu.edu/bitstream/2451/31553/2/Dhar-DataScience.pdf. [Último acceso: 30 Octubre 2020].

[18] UNIR, «Las tres V del Big Data: todo un reto por su volumen, variedad y velocidad,» 26 Junio 2020. [En línea]. Available: https://www.unir.net/ingenieria/revista/3-v-big-data/#:~:text=Las%20tres%20V%20del%20Big%20Data%20se%20refiere%20a

%20los,adem%C3%A1s%20de%20sus%20principales%20retos.. [Último acceso: 26 Octubre 2020].

[19] ESIC Business & Marketing School, «Las ventajas del Big Data,» 15 Enero 2018. [En línea]. Available: https://www.esic.edu/rethink/tecnologia/las-ventajas-del-big-data. [Último acceso: 26 Octubre 2020].

[20] R. Coello y J. Parrales, «Análisis de las ventajas y desventajas del Big Data y el Cloud Computing en el proceso de la toma de decisiones de lasempresas que practican comercio electrónico,» 01 Noviembre 2019. [En línea]. Available: http://cienciaytecnologia.uteg.edu.ec/revista/index.php/cienciaytecnologia/article/view/279/422. [Último acceso: 28 Octubre 2020].

[21] E. Hernandez, N. Duque y J. Moreno, «Big Data: una exploración de investigaciones, tecnologias y casos de aplicación,» 15 Marzo 2017. [En línea]. Available: http://www.scielo.org.co/pdf/teclo/v20n39/v20n39a02.pdf. [Último acceso: 23 Octubre 2020].

[22] SAS, «Hadoop,» 15 Mayo 2018. [En línea]. Available: https://www.sas.com/es_mx/insights/big-data/hadoop.html. [Último acceso: 20 Octubre 2020].

[23] E. M. Julio Santana, «El arte de programar en R,» 27 Noviembre 2014. [En línea]. Available: https://ftp.uni-sofia.bg/CRAN/doc/contrib/Santana_El_arte_de_programar_en_R.pdf. [Último acceso: 26 Octubre 2020].

[24] I. Perez, «El lenguaje de programación Python/The programming language Python,» 05 Junio 2014. [En línea]. Available: https://www.redalyc.org/pdf/1815/181531232001.pdf. [Último acceso: 01 Octubre 2020].

[25] IBM, «Understanding the architectural layers of a big data solution,» 14 Octubre 2013. [En línea]. Available: https://developer.ibm.com/articles/bd-archpatterns3/. [Último acceso: 08 Diciembre 2020].

[26] H.-K. Lin, «A Hyperconnected Manufacturing Collaboration System Using the Semantic Web and Hadoop Ecosystem System,» 18 Diciembre 2016. [En línea]. Available: https://www.researchgate.net/figure/Apache-Hadoop-Ecosystem_fig3_307619823. [Último acceso: 02 Enero 2021].

[27] L. Diaz, J. Garcia, B. Lopez y L. Gonzalez, «Técnicas para capturar cambios en los datos y mantener actualizado un almacén de datos,» 15 Diciembre 2015. [En línea]. Available:

http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992015000400007. [Último acceso: 25 Septiembre 2020].

[28] C. Barriga, «Tecnicas de Analisis de Datos en BIG DATA,» 15 Septiembre 2017. [En línea]. Available: https://www.researchgate.net/figure/Apache-Hadoop-Ecosystem_fig3_307619823. [Último acceso: 18 Noviembre 2020].

[29] Secretaria de la Administración Pública, «GPR,» 12 Noviembre 2011. [En línea]. Available: https://www.gestionderiesgos.gob.ec/wp-content/uploads/downloads/2012/06/Acuerdo1.pdf. [Último acceso: 15 Septiembre 2020].

[30] M. Sanchez, «Indicadores de gestión hospitalaria,» 25 Junio 2005. [En línea]. Available: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-75852005000200009. [Último acceso: 15 Octubre 2020].

[31] F. Redrovan y M. Zea, «Big Data y la Salud Pública. Detección de brotes epidémicos de enfermedades infecciosas,» 25 Junio 2018. [En línea]. Available: https://www.academia.edu/8945367/Articulo_Big_Data_y_Salud_Publica. [Último acceso: 26 Octubre 2020].

[32] F. Villalta, «CRISP-DM: una metodología para minería de datos en salud,» 12 Diciembre 2020. [En línea]. Available: https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/. [Último acceso: 4 Enero 2021].

[33] J. Camacho, «Analítica de grandes datos en Salud: Retos por superar,» 06 Marzo 2019. [En línea]. Available: ANALÍTICA DE GRANDES DATOS EN SALUD: RETOS POR SUPERAR. [Último acceso: 01 Noviembre 2020].

[34] C. P. Guevara, «Big Data, a tool to avoid failures and support in decisions of the hotel sector in Quito-Ecuador,» 12 Julio 2019. [En línea]. Available: https://revistas.uide.edu.ec/index.php/innova/article/view/1062/1615. [Último acceso: 25 Octubre 2020].

[35] J. Valls, «El Big Data y los consumidores,» 07 Diciembre 2017. [En línea]. Available: https://es.weforum.org/agenda/2017/12/el-big-data-y-los-consumidores/. [Último acceso: 30 Octubre 2020].

[36] D. Riskin, «The Next Revolution in Healthcare,» 01 Octubre 2012. [En línea]. Available: https://www.forbes.com/sites/singularity/2012/10/01/the-next-revolution-in-healthcare/?sh=6887304455cc. [Último acceso: 01 Noviembre 2020].

[37] E. Benites, «El uso de big data en medicina,» 15 Octubre 2018. [En línea]. Available: https://www.eluniverso.com/opinion/2018/10/15/nota/7000838/uso-big-data-medicina. [Último acceso: 01 Noviembre 2020].

[38] ESIC Business & Marketing, «Apache Spark: Introducción, qué es y cómo funciona,» 15 Octubre 2018. [En línea]. Available: https://www.esic.edu/rethink/tecnologia/apache-spark-introduccion-que-es-y-como-funciona. [Último acceso: 22 Octubre 2020].

[39] Intelipaat, «Introduction to Apache Storm,» 15 Diciembre 2016. [En línea]. Available: https://intellipaat.com/blog/what-is-apache-storm/. [Último acceso: 23 Octubre 2020].

[40] Mongo DB, «La base de datos líder para aplicaciones modernas,» 15 Enero 2020. [En línea]. Available: https://www.mongodb.com/es. [Último acceso: 30 Octubre 2020].

[41] Elastic, «El corazón del Elastic Stack, gratuito y abierto,» 15 Enero 2020. [En línea]. Available: https://www.elastic.co/es/elasticsearch/. [Último acceso: 01 Noviembre 2020].

[42] S. Haloi, «Aplicaciones distribuidas en Zookeper,» de *Apache ZooKeeper Essentials*, Birminghan, UK., Packt Publishing, 2015, pp. 25-26.

[43] A. Bustamante y E. Galvis, «Técnicas del modelado de procesos ETL,» 23 Enero 2012. [En línea]. Available: https://dialnet.unirioja.es/descarga/articulo/4271531.pdf. [Último acceso: 15 Septiembre 2020].

[44] J. Thomas y D. Stefan, Near Real-Time Data Warehousing Using State-of-the-Art ETL Tools, Berlin: Springer-Verlag Heidelberg, 2010.

[45] J. Perez, «Big Data for Health,» 15 Julio 2015. [En línea]. Available: https://ieeexplore.ieee.org/abstract/document/7154395. [Último acceso: 26 Septiembre 2020].

[46] D. Bates, S. Saria y L. Ohno, «Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients,» 12 Julio 2014. [En línea]. Available: https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2014.0041. [Último acceso: 15 Agosto 2020].

[47] L. Raymond, H. Demirkan y A. Kazam, «Smart health: Big data enabled health paradigm within smart cities,» 30 Noviembre 2017. [En línea]. Available: https://www.sciencedirect.com/science/article/abs/pii/S095741741730444X. [Último acceso: 15 Septiembre 2020].

[48] S. Sukumar, R. Ratajaran y R. Ferrel, «Quality of Big Data in health care,» 13 Julio 2015. [En línea]. Available: https://www.emerald.com/insight/content/doi/10.1108/IJHCQA-07-2014-0080/full/html. [Último acceso: 26 Agosto 2020].

[49] T. Furche, L. Libkin, G. Orsi y N. Paton, «Big-ETL: Extracting-Transforming-Loading Approach for Big Data,» 25 Marzo 2015. [En línea]. Available: https://d1wqtxts1xzle7.cloudfront.net/54520864/PDP3312.pdf?1506257335=&response-content-disposition=inline%3B+filename%3DBig_ETL_extracting_transforming_loading.pdf&Expires=1604586964&Signature=XUHpozC8aDbOdrJDV74u7TwFGoy1~hBwH~D1J6GD0rx5oydTkWtEuv-5BAf7s. [Último acceso: 30 Agosto 2020].

[50] A. Katal y M. Wazid, «Big data: Issues, challenges, tools and Good practices,» 13 Agosto 2013. [En línea]. Available: https://ieeexplore.ieee.org/abstract/document/6612229. [Último acceso: 15 Octubre 2020].

[51] M. Zorrila y D. García, «Arquitecturas y tecnologías para big data,» 28 Enero 2017. [En línea]. Available: https://ocw.unican.es/pluginfile.php/2396/course/section/2473/tema%203.2%20Arquitecturas%20y%20tecnologi%CC%81as%20para%20el%20big%20data.pdf. [Último acceso: 17 Diciembre 2020].

[52] Cloudera. Apache Kafka, 4 de Enero del 2019. [En línea]. Available: https://docs.cloudera.com/documentation/kafka/1-2-x/topics/kafka.html. [Último acceso: 4 Enero 2021]

ANEXO 1

INSTALACIÓN DEL SISTEMA BIG DATA

En esta sección se establecerá los pasos necesarios para montar y configurar un sistema de Big Data según el esquema propuesto de tipo distribuido, con su respectivo entorno, para que sea funcional y completo con la opción de tener la capacidad de procesar los datos y obtener los resultados requeridos.

PREPARACIÓN DEL ENTORNO

En la sección de diseño de recursos se determinó los equipos que se usarán para el desarrollo del proyecto, para ello se tomará en cuenta el uso del sistema operativo UBUNTU SERVER 18.04 la más estable, por ser gratuito y además tener bien soporte para la configuración e instalación de HADOOP. A nivel de software solo necesitamos:

- Ubuntu Server 18.04, debido a la naturaleza del proyecto, no se tomará en cuenta el proceso a seguir para la instalación de Ubuntu, por lo que se recomienda consultar los manuales y documentos que se encuentran en internet.
- Apache Hadoop 3.1.3
- Openjdk versión 11.0.4

INSTALACIÓN DE JAVA Y HADOOP EN LINUX

Para la instalación de HADOOP primero se instalará el entorno de java y luego Hadoop (cada máquina aplicará los mismos pasos), el cual se tomará los siguientes pasos:

1. Configuración del directorio

   sudo mkdir -p /opt/{hadoop/{logs},hdfs/{datanode,namenode},yarn/{logs}

2. Instalación del JDK

   ```
   apt-get update
   apt-get install default-jdk
   update-alternatives --config java
   ```

3. Se creará un entorno en java llamado JAVA_HOME, seguimos los pasos en cada línea del código.

   ```
   vi /etc/profile.d/java.sh
   #/bin/bash
   ```

export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
reboot
env | grep JAVA_HOME
JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
java -version
openjdk version "11.0.4" 2019-07-16
OpenJDK Runtime Environment (build 11.0.4+11-post-Ubuntu-1ubuntu219.04)
OpenJDK 64-Bit Server VM (build 11.0.4+11-post-Ubuntu-1ubuntu219.04, mixed mode, sharing)

4. Se creará un usuario que se llama hadoop

adduser hadoop
Adding user `hadoop' ...
Adding new group `hadoop' (1002) ...
Adding new user `hadoop' (1002) with group `hadoop' ...
Creating home directory `/home/hadoop' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hadoop
Enter the new value, or press ENTER for the default
Full Name []:
Room Number []:
Work Phone []:
Home Phone []:
Other []:
Is the information correct? [Y/n] Y

5. Generamos las claves SSH para el usuario HADOOP

su hadoop
ssh-keygen
su hadoop
ssh-copy-id 127.0.0.1
exit

6. Descargar HADOOP

mkdir /downloads
cd /downloads
wget http://mirror.nbtelecom.com.br/apache/hadoop/common/hadoop-3.1.3/hadoop-3.1.3.tar.gz
cd /downloads
tar -zxvf hadoop-3.1.3.tar.gz
mv hadoop-3.1.3 /usr/local/hadoop
chown hadoop.hadoop /usr/local/hadoop -R

7. Configuración de las variables de entorno

sudo gedit / etc / profile *Al ejecutar este comando debe de verse las siguientes líneas, modificar lo necesario.
if [“$PS1”];
then if [“$BASH”] && [“$BASH” != “/bin/sh”]; then# The file bash.bashrc already sets the default PS1.# PS1=’\h:\w\$ ‘if [-f /etc/bash.bashrc]; then. /etc/bash.bashrcfielseif [“`id -u`” -eq 0];
thenPS1=’# ‘elsePS1=’$ ‘fififiif [-d /etc/profile.d]; thenfor i in

/etc/profile.d/*.sh; doif [-r $i]; then. $ifidoneunset ifi export
HADOOP_HOME=/opt/hadoopexport
PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:$HADOO
P_HOME/bin:$HADOOP_HOME/sbinexport
HADOOP_CONF_DIR=/opt/hadoop/etc/hadoopexport
HDFS_NAMENODE_USER=rootexport
HDFS_DATANODE_USER=rootexport
HDFS_SECONDARYNAMENODE_USER=rootexport
JAVA_HOME=/usr/lib/jvm/java-8-oracleexport
HADOOP_MAPRED_HOME=/opt/hadoopexport
HADOOP_COMMON_HOME=/opt/hadoopexport
HADOOP_HDFS_HOME=/opt/hadoopexport
YARN_HOME=/opt/Hadoop

8. Aplicación de un enlace simbólico
   sudo ln -sf / etc / profile /root/.bashrc source / etc / profile

9. Se actualiza la siguiente información, / opt / hadoop / etc / hadoop / hadoop-env.sh y establezca la variable JAVA_HOME y las variables HADOOP_HOME, HADOOP_CONF_DIR Y HADOOP_LOG_DIR

   export JAVA_HOME=/usr/lib/jvm/java-8-oracleexport
   HADOOP_HOME=/opt/hadoopexport
   HADOOP_CONF_DIR=/opt/hadoop /etc/hadoopexport
   HADOOP_LOG_DIR=/opt/hadoop/logs

10. Comprobar la versión de hadoop
    hadoop versión

CONFIGURACIÓN DE HADOOP MULTINODO

Para la configuración multinodo se considera la siguiente tabla:

Tabla 1.- Configuración de los equipos en multinodo

IP	HOSTNAME	IS NAME NODE	IS DATE NODE
192.168.10.1	Hadoop-namenode	SI	SI
192.168.10.2	Haddop-datanode-2	NO	SI
192.168.10.3	Haddop-datanode-3	NO	SI
192.168.10.4	Haddop-datanode-4	NO	SI
192.168.10.5	Haddop-datanode-5	NO	SI
192.168.10.6	Haddop-datanode-6	NO	SI

Fuente: Autor

La primera máquina con IP: 192.168.10.1, actuara como nodo principal o maestro, con un nodo de dato esclavo, los demás equipos serán nodos de datos esclavos. Para proceder con la configuración se tomará en cuenta los siguientes pasos:

1. Modificación de los archivos host de cada máquina:

   sudo nano / etc / hosts

   127.0.0.1 localhost
   192.168.10.1 hadoop-namenode
   192.168.10.2 hadoop-datanode-2
   192.168.10.3 hadoop-datadnode-3
   192.168.10.4 hadoop-datadnode-4
   192.168.10.5 hadoop-datadnode-5
   192.168.10.6 hadoop-datadnode-6

CONFIGURACIÓN DEL NODO MAESTRO - NAMENODE

1. Se procede actualizar el archivo xml hdfs-site:

   sudo gedit /opt/hadoop/etc/hadoop/hdfs-site.xml

   <configuration>
   <property><name>dfs.namenode.name.dir</name><value>file:///opt/hdfs/namenode</value><description>
   NameNode directory for namespace and transaction logs
   storage.</description></property><property><name>dfs.datanode.data.dir</name><value>file:///opt/hdfs/datanode</value><description>DataNode
   directory</description></property><property><name>dfs.replication</name><value>3</value></property><property><name>dfs.permissions</name><value>false</value></property><property><name>dfs.datanode.use.datanode.hostname</name><value>false</value></property><property><name>dfs.namenode.datanode.registration.ip-hostname-check</name><value>false</value></property>
   </configuration>

2. Se procede actualizar el siguiente archivo xml core-site:

   sudo gedit /opt/hadoop/etc/hadoop/core-site.xml

   <configuration><property><name>fs.defaultFS</name><value>hdfs://hadoop-namenode:9820/</value><description>NameNode
   URI</description></property><property><name>io.file.buffer.size</name><value>131072</value><description>Buffer
   size</description></property>
   </configuration>

3. Se procede actualizar el siguiente archivo xml yarn-site:

   sudo gedit /opt/hadoop/etc/hadoop/yarn-site.xml

   <configuration><property><name>yarn.nodemanager.aux-services</name><value>mapreduce_shuffle</value><description>Yarn Node Manager Aux
   Service</description></property><property><name>yarn.nodemanager.aux-

services.mapreduce.shuffle.class</name><value>org.apache.hadoop.mapred.ShuffleHandler</value></property><property><name>yarn.nodemanager.local-dirs</name><value>file:///opt/yarn/local</value></property><property><name>yarn.nodemanager.log-dirs</name><value>file:///opt/yarn/logs</value></property>
</configuration>

4. Se procede actualizar el siguiente archivo xml mapre-site:

   sudo gedit /opt/hadoop/etc/hadoop/mapre-site.xml

   <configuration><property><name>mapreduce.framework.name</name><value>yarn</value><description>MapReduce framework name</description></property>
   <property><name>mapreduce.jobhistory.address</name><value>hadoop-namenode:10020</value><description>Default port is 10020.</description></property><property><name>mapreduce.jobhistory.webapp.address</name><value> hadoop-namenode:19888</value><description>Default port is 19888.</description></property><property><name>mapreduce.jobhistory.intermediate-done-dir</name><value>/mr-history/tmp</value><description>Directory where history files are written by MapReduce jobs.</description></property><property><name>mapreduce.jobhistory.done-dir</name><value>/mr-history/done</value><description>Directory where history files are managed by the MR JobHistory Server.</description></property>
   </configuration>

5. Se procede a formatear el nombre del nodo

   hdfs namenode –format

6. Se proceden agregar los nodos datos esclavos en la siguiente ruta / opt / hadoop / etc / hadoop, dentro del archivo workers

   192.168.10.1 192.168.10.2 192.168.10.3 192.168.10.4 192.168.10.5 192.168.10.6

7. Asegurarse que los nodename tiene una clave de acceso ssh:

   ssh-copy-id -i /home/hadoopuser/.ssh/id_rsa.pub hadoopuser@192.168.10.2
   ssh-copy-id -i /home/hadoopuser/.ssh/id_rsa.pub hadoopuser@192.168.10.3
   ssh-copy-id -i /home/hadoopuser/.ssh/id_rsa.pub hadoopuser@192.168.10.4
   ssh-copy-id -i /home/hadoopuser/.ssh/id_rsa.pub hadoopuser@192.168.10.5
   ssh-copy-id -i /home/hadoopuser/.ssh/id_rsa.pub hadoopuser@192.168.10.6

CONFIGURACIÓN DEL NODO ESCLAVOS - DATANODE

En cada máquina que servirá de nodo esclavo se procederá a realizar los siguientes pasos:

- Para no proceder a realizar la instalación de Hadoop nuevamente en cada máquina esclava, se ejecutará el siguiente comando, para poder copiar la estructura del nodo principal (este proceso solo se ejecuta una sola vez en el nodo principal).

```
ssh-copy-id -i /home/hadoopuser/.ssh/id_rsa.pub
hadoopuser@192.168.10.2
ssh-copy-id -i /home/hadoopuser/.ssh/id_rsa.pub
hadoopuser@192.168.10.3
ssh-copy-id -i /home/hadoopuser/.ssh/id_rsa.pub
hadoopuser@192.168.10.4
ssh-copy-id -i /home/hadoopuser/.ssh/id_rsa.pub
hadoopuser@192.168.10.5
ssh-copy-id -i /home/hadoopuser/.ssh/id_rsa.pub
hadoopuser@192.168.10.6
```

1. Se procede actualizar el archivo hdfs-site.xml

```
sudo gedit /opt/hadoop/etc/hadoop/hdfs-site.xml

<configuration><property><name>dfs.datanode.data.dir</name><value
>file:///opt/hdfs/datanode</value><description>DataNode
directory</description></property><property><name>dfs.replication</na
me><value>3</value></property><property><name>dfs.permissions</n
ame><value>false</value></property><property><name>dfs.datanode.
use.datanode.hostname</name><value>false</value></property>
</configuration>
```

2. Se procede actualizar el archivo core-site.xml

```
sudo gedit /opt/hadoop/etc/hadoop/core-site.xml

<configuration><property><name>fs.defaultFS</name><value>hdfs://ha
doop-namenode:9820/</value><description>NameNode
URI</description></property>
</configuration>
```

3. Se procede actualizar el archivo yarn-site.xml

```
sudo gedit /opt/hadoop/etc/hadoop/yarn-site.xml
```

<configuration><property><name>yarn.nodemanager.aux-services</name><value>mapreduce_shuffle</value><description>Yarn Node Manager Aux Service</description></property>
</configuration>

4. Se procede actualizar el archivo mapre-site.xml

sudo gedit /opt/hadoop/etc/hadoop/mapre-site.xml

<configuration><property><name>mapreduce.framework.name</name><value>yarn</value><description>MapReduce framework name</description></property>
</configuration>

Una configurada cada máquina esclava, se procede a inicializar Hadoop al igual que su administrador de recursos

start-dfs.sh
start-yarn.sh

Para validar que todo está correctamente instalado, se ejecutara el comand jps, si no muestra algún error, toda la configuración esta correcta, si existe algún error, por favor revisar nuevamente los pasos, o en su efecto se debe de buscar en la documentación que hay en internet sobre el proceso de instalación en HADOOP APACHE.

ANEXO 2

PROCEDIMIENTO PARA LA GENERACIÓN DE SERIES TEMPORALES EN R

```
install.packages("forecast")
install.packages("tseries")
install.packages("readr")
install.packages("ggplot2")
install.packages("ggfortify")
library(ggfortify)
library(ggplot2)
library(forecast)
library(tseries)
library(readr)

datos <- read_csv(file.choose())
datos.ts<-ts(datos, start = c(2014,1), frequency = 12)
print(datos.ts)
datos.ts.desc = decompose(datos.ts)
plot(datos.ts.desc, xlab='Año')
```

MODELOS DE PRONÓSTICO EN SERIES TEMPORALES

Modelo NNETAR

```
f7 <- nnetar(serie1)
fc7<- forecast(f7, h=36)
plot(fc7)
fc7
accuracy(fc7)
fc7
```

Modelo STML

```
f1 <- stlm(serie1, modelfunction=ar)
fc1<- forecast(f1, h=36)
plot(fc1)
accuracy(fc1)
fc1
```

Modelo Holt Winters

```
f3 <- HoltWinters(serie1)
fc3 <- forecast(f3, 36)
plot(fc3)
accuracy(fc3)
fc3
```

Modelo TBATS

```
f8 <- tbats(serie1, biasadj=TRUE)
fc8<- forecast(f8, h=36)
plot(fc8)
fc8
accuracy(fc8)
fc8
```

ANEXO 3

MODELO DE PRONÓSTICO PARA INDICADOR PPEA

Forecast method: NNAR(2,1,2)[12]

Model Information:
Average of 20 networks, each of which is
a 3-2-1 network with 11 weights
options were - linear output units

	Error measures						
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-1.00E-05	0.1276846	0.09436684	-Inf	Inf	0.3583021	0.03602173

Forecast method: STL + AR(2)

Model Information:
Call: modelfunction (x = x.sa)
Coefficients:
1 2
0.5551 0.2361

Order selected 2 sigma^2 estimated as 0.03636

	Error measures						
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.0031321	0.1869548	0.1307153	-Inf	Inf	0.4963138	-0.0212582

Forecast method: HoltWinters

Model Information:
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = seriePPEA)

Smoothing parameters:
alpha: 0.4529059
beta : 0
gamma: 1

Coefficients:
[,1]
a 0.43818162
b -0.01530859
s1 -0.04833654

s2 0.04190920
s3 0.16041281
s4 0.11519136
s5 0.14881788
s6 0.17501127
s7 -0.21217982
s8 -0.04357750
s9 0.12960779
s10 0.13941178
s11 0.08537402
s12 0.02012300

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.01885496	0.25842	0.1818096	-Inf	Inf	0.6903142	0.1080041

Forecast method: BATS(1, {0,0}, -, -)

Model Information:
BATS(1, {0,0}, -, -)
Call: tbats(y = seriePPEA, biasadj = TRUE)
Parameters
  Alpha: 0.4629803
Seed States:
        [,1]
[1,] 0.8123107
Sigma: 0.2065308
AIC: 104.4269

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.01083568	0.2065308	0.1353142	-Inf	Inf	0.5137754	0.0494632

MODELO DE PRONÓSTICO PARA INDICADOR THMM

Forecast method: NNAR(1,1,2)[12]

Model Information:

Average of 20 networks, each of which is
a 2-2-1 network with 9 weights
options were - linear output units

	Error measures						
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	5.14E-08	0.01321677	0.00347113	NaN	Inf	1.405816	-0.00864864

Forecast method: STL + AR(0)

Model Information:
Call:
modelfunction(x = x.sa)
Order selected 0 sigma^2 estimated as 0.0001201

	Error measures						
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-9.81E-20	0.01089321	0.00309186	NaN	Inf	1.252212	0.00650533

Forecast method: HoltWinters

Model Information:
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = serieTHMM)

Smoothing parameters:
alpha: 0
beta : 0
gamma: 0

Coefficients:
[,1]
a 2.539279e-03
b 3.391197e-05
s1 -5.852000e-04
s2 -5.852000e-04

s3 -5.852000e-04
s4 -5.852000e-04
s5   2.579357e-03
s6   2.280130e-03
s7 -5.852000e-04
s8 -5.852000e-04
s9 -5.852000e-04
s10 -8.270944e-05
s11 -2.145660e-04
s12 -4.658112e-04

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.00073505	0.01323168	0.00308359	NaN	Inf	1.24886	-0.03131861

Forecast method: BATS(1, {0,0}, -, -)

Model Information:
BATS(1, {0,0}, -, -)
Call: tbats(y = serieTHMM, biasadj = TRUE)
Parameters
  Alpha: 0.01323461
Seed States:
          [,1]
[1,] 0.0002675098

Sigma: 0.01235617
AIC: -351.8128

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.00143347	0.01235617	0.00211307	-Inf	Inf	0.8557986	-0.03411855

MODELO DE PRONÓSTICO PARA INDICADOR PHMN

Forecast method: NNAR(1,1,2)[12]

Model Information:

Average of 20 networks, each of which is
a 2-2-1 network with 9 weights
options were - linear output units

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-4.29E-06	0.00483871	0.00330839	-Inf	Inf	0.6704148	0.01141185

Forecast method: STL + AR(0)

Model Information:

Call:
modelfunction(x = x.sa)

Order selected 0 sigma^2 estimated as 1.927e-05

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-1.13E-19	0.00436297	0.00318539	-Inf	Inf	0.6454891	-0.09893209

Forecast method: HoltWinters

Model Information:
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = seriePHMN)

Smoothing parameters:
alpha: 0.05168247
beta : 0.08167691
gamma: 0.07442131

Coefficients:
[,1]
a 3.953750e-03
b -1.786963e-05
s1 -2.121911e-03

s2 -1.953788e-03
s3   2.404642e-03
s4   5.971105e-03
s5   7.042766e-03
s6 -2.549412e-03
s7 -2.983848e-03
s8 -3.395256e-03
s9 -3.402101e-03
s10 -1.746753e-03
s11 -2.193994e-03
s12 1.671040e-03

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.00066894	0.00554702	0.0039664	NaN	Inf	0.8037541	-0.05051132

Forecast method: BATS(1, {0,0}, 0.836, -)

Model Information:
BATS(1, {0,0}, 0.836, -)
Call: tbats(y = seriePHMN, biasadj = TRUE)
Parameters
  Alpha: -0.2195016
  Beta: 0.04302668
  Damping Parameter: 0.836489
Seed States:
            [,1]
[1,] 0.0001105095
[2,] 0.0006510642
Sigma: 0.004603153
AIC: -505.7738

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.00020582	0.00460315	0.00308819	NaN	Inf	0.6257927	0.01292611

MODELO DE PRONÓSTICO PARA INDICADOR NPELQ

Forecast method: NNAR(1,1,2)[12]

Model Information:

Average of 20 networks, each of which is
a 2-2-1 network with 9 weights
options were - linear output units

	Error measures						
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.01828562	48.13157	22.84	-Inf	Inf	0.3024875	0.0299678

Forecast method: STL + AR(1)

Model Information:

Call:
modelfunction(x = x.sa)

Coefficients:
1
0.9315
Order selected 1 sigma^2 estimated as 2404

	Error measures						
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	2.492106	46.04524	22.96927	-Inf	Inf	0.3041995	0.02037114

Forecast method: HoltWinters

Model Information:
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = serieNPLEQ)

Smoothing parameters:
alpha: 0.975642
beta : 0
gamma: 0

Coefficients:
[,1]
a 232.2334583
b 0.5426865
s1 2.9756944

s2 -5.3159722
s3   4.8090278
s4   0.3923611
s5 -4.8159722
s6 -4.8159722
s7 -6.5243056
s8 -4.6076389
s9 12.7256944
s10 -4.2743056
s11 2.0173611
s12 7.4340278

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	2.734578	53.28861	24.65928	NaN	Inf	0.3265816	-0.00429028

Forecast method: BATS(1, {0,0}, -, -)

Model Information:
BATS(1, {0,0}, -, -)
Call: tbats(y = serieNPLEQ, biasadj = TRUE)
Parameters
  Alpha: 0.9958775
Seed States:
        [,1]
[1,] 13.61174
Sigma: 48.9228
AIC: 990.1699

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	2.784865	48.9228	19.99219	-Inf	Inf	0.2647718	-0.00429534

MODELO DE PRONÓSTICO PARA INDICADOR POCP

Forecast method: NNAR(1,1,2)[12]

Model Information:

Average of 20 networks, each of which is
a 2-2-1 network with 9 weights
options were - linear output units

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.0003547 2	0.115581 6	0.0569095 5	-1.24538 8	5.9227 3	0.774770 4	0.0039774 2

Forecast method: STL + AR(0)

Model Information:
Call:
modelfunction(x = x.sa)

Order selected 0 sigma^2 estimated as 0.01628

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-2.74E-17	0.126822 3	0.0587952 9	-1.3351 7	6.19513 2	0.800443 1	-0.0756526 9

Forecast method: HoltWinters

Model Information:
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = seriePOCP)

Smoothing parameters:
alpha: 0.009803122
beta : 0.414738
gamma: 0

Coefficients:
[,1]
a 0.882256747
b 0.002679321
s1 -0.035556951

s2 -0.050200900
s3 -0.032891922
s4 -0.012452867
s5 0.019266454
s6 0.026458578
s7 0.059695022

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.0268844 7	0.153982 4	0.0785177 5	1.3222 3	8.21549 7	1.06894 6	-0.0611979 9

Forecast method: BATS(0.008, {0,5}, -, -)

Model Information:
BATS(0.008, {0,5}, -, -)
Call: tbats(y = seriePOCP, biasadj = TRUE)
Parameters
Lambda: 0.008314
Alpha: 0.03728777
MA coefficients: -0.182197 0.585217 -0.257256 0.562674 -0.444253
Seed States:
[,1]
[1,] -0.06789185
[2,] 0.00000000
[3,] 0.00000000
attr(,"lambda")
[1] 0.008313656
Sigma: 0.1066126
AIC: 2.407522

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.0100693 4	0.119906 4	0.0683853 1	-2.33187 5	7.34226 6	0.931002 2	-0.0354713 1

MODELO DE PRONÓSTICO PARA INDICADOR TDMH

Forecast method: NNAR(8,1,5)[12]

Model Information:

Average of 20 networks, each of which is
a 9-5-1 network with 56 weights
options were - linear output units

	Error measures						
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	8.99E-06	0.00331636	0.00251762	-3.933186	14.30175	0.06684957	0.278447 1

Forecast method: STL + AR(8)

Model Information:

Call:
modelfunction(x = x.sa)

Coefficients:

1	2	3	4	5	6	7	8
0.9612	-0.0024	-0.0828	0.0916	-0.0702	-0.0138	-0.3556	0.3242

Order selected 8 sigma^2 estimated as 0.00184

	Error measures						
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.0018764 1	0.0383292 9	0.0149185 8	-22.2693 1	52.8541 7	0.396128 2	-0.02175178

Forecast method: HoltWinters

Model Information:
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = serieTDMH)

Smoothing parameters:
alpha: 0.8885082
beta : 0
gamma: 1

Coefficients:

	[,1]
a	2.867110e-01
b	1.406941e-04

s1 -5.566722e-07
s2 5.253744e-04
s3 -1.804851e-03
s4 -1.396698e-04
s5 9.484811e-04
s6 3.875841e-02

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.004219 7	0.04424608	0.01115885	-2.009614	20.3218 2	0.296297 1	0.01313 9

Forecast method: BATS(0, {0,0}, -, -)

Model Information:
BATS(0, {0,0}, -, -)

Call: tbats(y = serieTDMH, biasadj = TRUE)

Parameters
Lambda: 0
Alpha: 1.066058

Seed States:
[,1]
[1,] -3.963128
attr(,"lambda")
[1] 1.864251e-08

Sigma: 0.3776843
AIC: -408.4706

Error measures							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.00084691	0.0446013 8	0.011884 2	-8.91123	20.6754 4	0.315557 2	-0.2667701