Arquitectura de consolidación de la información para seguros de la salud mediante Big Data

Information consolidation architecture for health insurance using Big Data

Arquitetura de consolidação de informações para planos de saúde usando Big Data

José Zerega-Prado [1], Joe Llerena-Izquierdo [2]

Recibido: Febrero 2022 Aceptado: Julio 2022

Resumen. - La identificación de los datos que están en varias fuentes de información y su consolidación para entregarla como útil se logra con Big Data. El objetivo general de este trabajo es desarrollar un diseño de arquitectura de consolidación de la información para seguros de la salud mediante Big Data. Para esta propuesta de investigación se utiliza el método empírico analítico, de tipo cuasi experimental con enfoque cuantitativo, mediante el análisis de referencias relevantes y especificación de los componentes de la arquitectura. Los resultados de esta investigación permiten categorizar diferentes arquitecturas computacionales para seguros de la salud mediante una revisión de literatura relevante, desarrollar un modelo de arquitectura de un sistema computacional para una empresa ecuatoriana de seguros de salud orientado a la consolidación de la información, y evaluar la metodología de estudio utilizada para establecer factores viables del modelo. El aporte de este trabajo permite determinar la aplicabilidad del modelo a empresas de seguros de salud nacionales o extranjeras mediante la contrastación de factores viables en una empresa específica del medio. Se concluye que las distintas fuentes de información o tipos de datos utilizados en el ámbito de los seguros de salud permiten conocer varias aristas del análisis de datos a través de una arquitectura en Big Data, además de obtener indicadores para mejorar la toma de decisiones; el 73% de los factores establecidos son viables en una empresa ecuatoriana de seguros de salud.

Palabras clave: Aplicaciones de Big Data; Salud y seguridad; Arquitectura de la información.

Summary. - *The identification of data that is in various sources of information and its consolidation to deliver it as useful is achieved with Big Data. The overall objective of this work is to develop an information consolidation architecture design for health insurance using Big Data. For this research proposal, the analytical empirical method is used, of a quasi-experimental type with a quantitative approach, through the analysis of relevant references and specification of the architecture components. The results of this research allow categorizing different computational architectures for health insurance through a review of relevant literature, developing an*

[1] Ingeniero en Sistemas, GIEACI Research Group, https://gieaci.blog.ups.edu.ec/, jzerega@est.ups.edu.ec, Universidad Politécnica Salesiana, Guayaquil, Ecuador, ORCID iD: https://orcid.org/0000-0001-7688-5149

[2] Magister en Sistemas de Información Gerencial, GieTICEA Educational Innovation Group, https://gieaci.blog.ups.edu.ec/, jllerena@ups.edu.ec, Universidad Politécnica Salesiana, Guayaquil, Ecuador, ORCID iD: https://orcid.org/0000-0001-9907-7048

architectural model of a computational system for an Ecuadorian health insurance company oriented to the consolidation of information, and evaluating the study methodology used to establish feasible factors of the model. The contribution of this work allows us to determine the applicability of the model to national or foreign health insurance companies by contrasting feasible factors in a specific company of the environment. It is concluded that the different sources of information or types of data used in the field of health insurance allow to know several edges of data analysis through a Big Data architecture, in addition to obtaining indicators to improve decision making; 73% of the established factors are viable in an Ecuadorian health insurance company.

Keywords: *Big Data applications; Health and safety; Information architecture.*

Resumo. **-** *A identificação dos dados que estão em várias fontes de informação e sua consolidação para entregá-los como úteis é conseguida com o Big Data. O objetivo geral deste trabalho é desenvolver um projeto de arquitetura de consolidação da informação para planos de saúde por meio de Big Data. Para esta proposta de pesquisa, é utilizado o método empírico analítico, de tipo quase-experimental com abordagem quantitativa, por meio da análise de referências relevantes e especificação dos componentes da arquitetura. Os resultados desta pesquisa permitem a categorização de diferentes arquiteturas computacionais para seguros de saúde por meio de uma revisão da literatura relevante, o desenvolvimento de um modelo de arquitetura de um sistema de computador para uma seguradora de saúde equatoriana orientado para a consolidação de informações e a avaliação do metodologia de estudo utilizada para estabelecer os fatores viáveis do modelo. A contribuição deste trabalho permite determinar a aplicabilidade do modelo a seguradoras de saúde nacionais ou estrangeiras, comparando fatores viáveis em uma empresa específica no ambiente. Conclui-se que as diferentes fontes de informação ou tipos de dados utilizados na área de seguros de saúde permitem conhecer vários aspectos da análise de dados por meio de uma arquitetura de Big Data, além de obter indicadores para melhorar a tomada de decisões; 73% dos fatores estabelecidos são viáveis em uma seguradora de saúde equatoriana.*

Palavras-chave: *Aplicações de Big Data; Saúde e segurança; Arquitetura de informação.*

1. Introducción.- Los datos en el ámbito de la salud se generan desde diferentes fuentes y formatos; algunos de ellos son el historial médico, ingresos y salidas de hospital, medicamentos, tratamientos y prescripciones, datos de pacientes, imágenes, farmacias, laboratorios, sensores y compañías de seguros; estos datos están en continuo crecimiento, se generan con más velocidad y en varias diversidades [1][2]. A su vez las compañías en seguros de salud basan sus procedimientos y sistemas en datos generados por el sector de la salud [3][4]. Los seguros de salud se aplican en diferentes sectores como, el sector público que atiende personas con financiamiento de impuestos, el sector privado que se financia con el cobro a los clientes, el seguro social que se financia con aportes de los empleados, y finalmente el seguro comunitario que se financia de impuestos y de la participación de la comunidad [5][6].

En los seguros de salud se evidencia la usabilidad del Big Data (BD) por las siguientes razones: entender la expansión de enfermedades, tendencias en uso de medicamentos, mejorar los servicios médicos, aumentar el conocimiento médico desde diferentes perspectivas, suministrar información médica a los clientes, divisar cambios en prescripciones médicas, obtener sensores sociales, vincular las prescripciones entre medicamentos y enfermedades [7][8][9], búsqueda en aumentar la experiencia del cliente, mejorar los comentarios de los clientes en los procesos y redes sociales [5][10], detectar fraudes generales en los reclamos [11], entre otros. Así Big Data introduce grandes mejoras en el área del seguro de salud como: lealtad y retención de clientes, entender la conducta de los clientes, buscar oportunidades con otros clientes, entender la conducta de los canales de ventas en seguros, aumento en venta de seguros de salud, conducta de los beneficiarios de las pólizas, búsqueda de nuevas redes de personas, reducción de costos y precios en la compañía, detectar fraudes en los reclamos, selección inmediata de reclamos interrelacionadas entre compañías y mejoras en los proceso de reclamaciones [5][12].

El gobierno de India explora los seguros de salud y el financiamiento para analizar el alza de precios en la atención de salud, enfermedades y verificar medicamentos. También utiliza Big Data para reducir los costos de atención médica, proveerla en un menor tiempo, dar un amplio acceso a un número mayor de personas y mejorar la calidad [5][13]. En Japón todos los ciudadanos están obligados a adquirir un seguro de salud, es decir la cobertura es 100%. Naciones que aplican Big Data en seguro de salud son India [5], Japón, Corea, Taiwán [7], Indonesia [14]. Los participantes en los seguros de salud son: pacientes, beneficiarios, compañías farmacéuticas, proveedores de salud públicos y privados, médicos, compañías de seguro médico, compañías de seguro general, intermediarios de seguros, gobiernos locales y los gobiernos nacionales [7][15]. Las existentes leyes locales y normas generales para uso correcto de los datos de salud, deben ser cumplidas por médicos y proveedores de salud, además los cambios en las regulaciones se aplican a los procedimientos y sistemas informáticos; es necesario recordar que la adherencia en normas de salud es esencial en la gestión de información privada de los pacientes, y en este escenario los datos generados por la pandemia COVID-19 también deben respetarse [16][17][18].

Ecuador tiene 21 compañías privadas en seguros de salud legalmente constituidas bajo control de la entidad de la Superintendencia de Compañías, Valores y Seguros. Además, este tipo de compañía de seguros de salud es llamada medicina pre pagada debido a que estas compañías ofrecen los seguros en forma directa o por intermediarios del sistema de seguros existente en el medio, así los clientes contratan este servicio por un pago anual [19][20]; además existe un servicio público del gobierno ecuatoriano que reembolsa dinero a las personas en caso de accidentes, por ejemplo el seguro de tránsito [21]. Este estudio se centra en los datos a nivel general de los reclamos a compañías en seguro de salud. Los sistemas en seguros de salud gestionan la adquisición de datos por medio de: solicitudes de seguro individual, declaración del médico, formulario de reclamos, anexos de explicaciones médicas, anexo de datos de los dependientes, formulario de pago, anexo de cambios o actualización. Los datos que existen en las compañías de seguros de salud tienen características heterogéneas, es decir están en un formato estructurado (bases de datos, hojas

electrónicas) y en formato no estructurado (procesadores de palabras, correos electrónicos, archivos portables «pdf», redes sociales). Desde el área de salud, cada día se generan y capturan datos, y están en revisión desde sistemas que pertenecen a las compañías en seguro de salud, y estos grandes datos necesitan procesos, herramientas y tecnología para una organización sistemática [5][22][23].

En las compañías de seguros de salud, el trabajo sobre análisis de datos es complejo, la confirmación de discrepancias toma mucho tiempo, la detección de documentos fraudulentos es difícil de detectar, existe gran cantidad de documentos de reclamos, se evidencia diversidad en tratamientos y prescripciones médicas, es decir los datos son abundantes [11][24][25]. En los sistemas de seguros de salud, los clientes solicitan el reembolso de los pagos en atención médica y medicina a través de formularios que poseen datos clínicos y detallados.

Se propone un modelo computacional para una empresa ecuatoriana en seguros de salud que gestione los datos heterogéneos y los consolide en información, para el modelo propuesto se debe exceptuar aquellos datos que provienen de pruebas médicas, imágenes o radiografías debido a que éstos, son datos confidenciales del paciente. El objetivo general plantea desarrollar un diseño de arquitectura de consolidación de la información para seguros de la salud mediante Big Data.

2. Metodología. - Para la propuesta de diseño de una arquitectura de consolidación de la información se utiliza una metodología empírico analítica, de tipo cuasi experimental con enfoque cuantitativo para el análisis de las referencias y propuesta de los componentes de la arquitectura. Se revisan las bibliotecas virtuales IEEExplore, Scopus, ACM, WoS entre otras. Se identifican y seleccionan los documentos con base teórica de propuestas sobre Big Data y Health Insurance. Se analizan los componentes de las arquitecturas en Big Data en los artículos científicos relevantes. Se escogen artículos científicos de alto impacto para el análisis de los elementos necesarios que permitan proponer una arquitectura de Big Data en el área de seguros de salud para una empresa en el contexto ecuatoriano. Se elabora un modelo de la arquitectura, su diagrama y debida explicación. Finalmente se evalúa la metodología utilizada para la factibilidad y aplicabilidad en una empresa específica en el medio.

En el desarrollo del modelo de arquitectura se utiliza una metodología iterativa [26] que ayuda a gestionar, examinar y visualizar los datos, la metodología contiene buenas prácticas y técnicas coordinadas que consiste en 5 fases: i) Definir etapas de los datos, ii) Gestionar las fuentes de datos, iii) Valorizar los datos, iv) Selección del almacén de datos, v) Implementación visual del Big Data. En la figura I se muestran las fases, en un proceso iterativo, debido a que se pueden realizar y repetir cada vez que existan más hallazgos de datos, de acuerdo con esto, el diseño de la arquitectura se aplica sobre datos de una empresa ecuatoriana de seguros de salud.

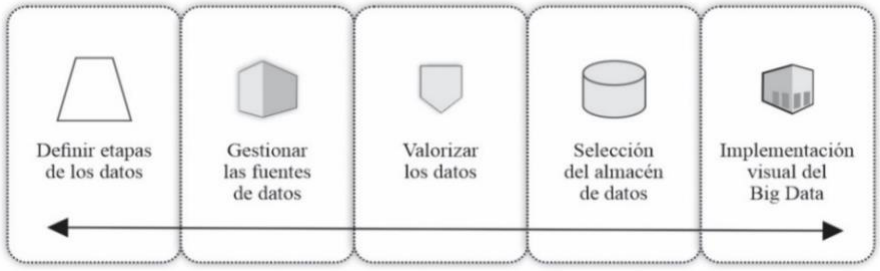

Figura I.- Metodología para desarrollo de Big Data

3. Revisión de la literatura. -

3.1. Big Data y conceptos generales. - Big Data es una colección de conjuntos de datos en considerables volúmenes. Medir o localizar problemas para una organización es complicado por la

variedad de datos que tiene. Con Big Data se abren posibilidades de aplicar herramientas tecnológicas, procesos y métodos para gestionar el conjuntos de datos y luego su posterior almacenamiento [5][27]. El conjunto de datos se almacena en formato estructurado o no estructurado y los datos disponibles sirven como plataforma de predicción [3][28].

De acuerdo a Alim y Abdul [3], Big Data tiene tres características principales: volumen, velocidad y variedad. Debido a que, en volumen, los datos médicos se duplican anualmente. En velocidad, el área de salud de cualquier departamento, redes o equipo genera más datos que otras áreas o sectores. En variedad, los datos generados por los equipos médicos son diferentes por cada equipo. Otros autores consideran más características: en veracidad, la información generada y almacenada es necesario mantener limpia y precisa [5][29]. En valor, va de acuerdo con el resultado. En validez, el dato debe ser correcto. En variabilidad, los datos tienen inconsistencia. En volatilidad, es la persistencia de los datos que determinan el almacenamiento. En visualización, es la facilidad de entender los datos [3].

3.2 Casos del uso de Big Data en Seguros de Salud.- Existen problemas de ética, uso, seguridad y privacidad para implementar Big Data en datos de salud y para alcanzar un buen éxito es necesario garantizar la seguridad y privacidad en los datos médicos, además de adoptar estándares y establecer un buen diseño del procesamiento de datos [1]. En India el gobierno respalda fuertemente la atención médica y los investigadores aplican BD sobre datos de salud, con ello obtienen conocimiento sobre los cuidados de salud que deben tener en los pacientes [5]. En Japón para identificar las causas de reclamos en los seguros, su modelo propone una revisión de medicación en lapsos de tiempo, estas prescripciones se descomponen para visualizar el comportamiento de medicina recomendada por el médico [7]. El seguro de salud nacional en Indonesia tuvo un aumento en las tarifas del año 2019, por esta razón aplicaron Big Data para capturar y estudiar los sentimientos expresados por los usuarios en una red social [14]. La reducción de fraudes en las compañías de seguros de salud fue minimizada a través de un framework que revisa los reclamos de seguros con la identificación de datos en los documentos [11]. Una comisión del gobierno de Australia analizó los gastos anuales de hospitales en 3 mil millones de dólares australianos, para reducir costos y aumentar la calidad de vida [30]. Existen algoritmos que trabajan sobre Big Data de compañías en seguros de salud, estos algoritmos evitan o localizan los posibles fraudes en los reclamos de una provincia de China [31]. En Australia se utilizan dos sistemas computacionales que analizan los reclamos en seguros de salud y utiliza los macro datos para descubrir fraudes, derroches o sobre precios en las transacciones [32], en este caso particular los investigadores aplicaron un modelo cuantitativo de Big Data en seguros de salud con datos de 1500 reclamos [33]. Se presenta un enfoque que se aplica a macro datos generados por el sistema informático en seguros de medicina para encontrar errores de datos o datos falsificados [34].

3.3 Datos generales en las compañías de seguros de salud.- Se explica algunos datos en las compañías en seguros de salud, estos datos contienen a nivel general, una solicitud de seguro individual, declaración del médico, formulario de reclamos, anexos de explicaciones médicas, anexo de datos de los dependientes, formulario de pago, anexo de cambios o actualización, entre los más relevantes. En hojas electrónicas están cotizaciones, calculadoras de constitución física, calculadoras de prorrateos e inclusiones. En procesadores de palabras están las cartas avales que se envían a los proveedores para que den cobertura directa a los clientes, las cartas de asegurabilidad que detallan la información general de la cobertura que tienen los clientes, las cartas de siniestralidad que contiene resumen de los reclamos que han presentado los clientes, el certificado de cobertura, aviso de prima, recibos de pago, tarjetas de membresía y documentos de anexos. En la página web de la empresa los visitantes tienen la disponibilidad de dejar mensajes.

4. Resultados.- En esta fase se obtuvieron tres resultados de acuerdo con los objetivos específicos

del estudio de investigación.

4.1. Categorizar diferentes arquitecturas computacionales para seguros de la salud mediante una revisión de literatura relevante.- Se ha categorizado las arquitecturas de acuerdo a los objetivos encontrados en otros trabajos relevantes, y las categorías son: prevención de enfermedades [5], [33], [35], [36] y [37] es decir, el 25% se concentraron en obtener mejores prácticas de salud o alimentación; causas de reclamos [7] y [38] es decir, el 10% se concentraron en las principales causas o razones que hacen los clientes en sus reclamos; expresión de sentimientos [14] es decir, el 5% se concentró en conocer las opiniones positivas o negativas sobre los reclamos; fraudes en reclamos [11], [31], [32], [34], [39], [40], [41], [42] es decir, el 40% de las arquitecturas se concentraron en detectar fraudes o estafas en los reclamos a los seguros de salud; predicción de enfermedades [30] y [43] es decir, 10% se concentra en conocer qué enfermedades podrían acaecer sobre la población; seguridad de datos [44] y [45] es decir, 10% trabaja sobre seguridad de información sobre salud de los clientes, ver figura II.

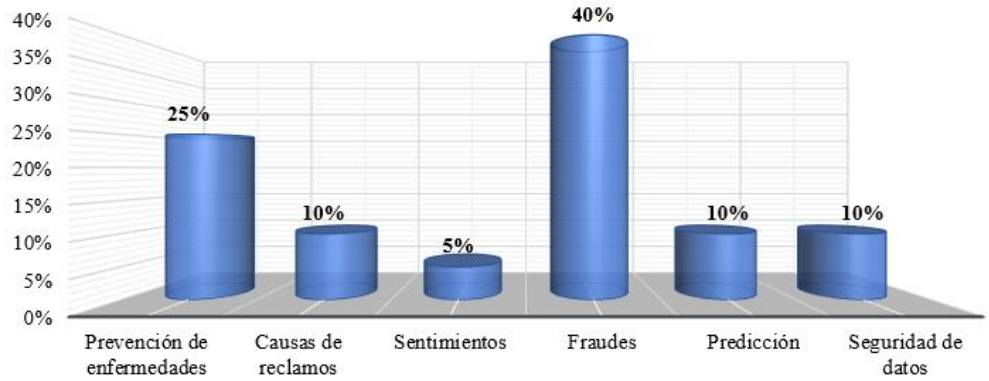

Figura II.- Categorías de los objetivos de las arquitecturas

4.2. Desarrollar un modelo de arquitectura de un sistema computacional para una empresa ecuatoriana de seguros de salud orientado a la consolidación de la información.- De acuerdo a la metodología utilizada [26] se sigue el siguiente orden:

i) Definir etapas de los datos.- Aquí se identifican los requisitos de información funcionales y los requisitos de información no funcionales que sirven para el análisis del proyecto de Big Data; los requisitos funcionales establecen la estructura de datos y los requisitos no funcionales establecen las herramientas apropiadas para las labores de análisis. Entre los requisitos no funcionales se destacan: a) oportunidad-circunstancia de análisis que define la condición de tiempo desde la captura de datos hasta el análisis, b) exigencias en calidad de datos, y c) tiempo de consultas. Se adoptan cuatro tipos de labores de análisis representadas en la tabla I y se aplica los requisitos no funcionales: a) oportunidad - circunstancia en segundos, horas y días, b) exigencias en calidad de los datos en baja, media y alta, c) tiempo de consulta en post segundo y sub segundo.

Ambiente	Circunstancia	Calidad de Datos	Tiempo de consulta
Datos sin procesar	–	baja	post – segundo
Reportes	días	alta	post – segundo
Procesamiento en línea	horas	media	sub – segundo
Tablero de gestión	horas	media	sub – segundo

Tabla I.- Requerimientos no funcionales

En este caso de la compañía de seguros de salud, el procesamiento en línea y el tablero de gestión son bajo tiempo/latencia y tienen la misma calidad de datos y, la misma circunstancia, basados en esto se definen 3 etapas, ver figura III.

Figura III.- Camino de datos para la compañía

ii) Gestionar las fuentes de datos.- Aquí se recolectan datos sin procesar y son cargados al sistema de archivos Big Data para que se almacenen en un clúster HDFS; de acuerdo a las oportunidades de la primera fase y la generación de las fuentes de datos, el responsable del Dataware House puede utilizar dos técnicas; la primera técnica es "Carga por lotes" que utiliza en momentos que los datos están lejano del tiempo real o pueden estar empaquetados y la herramienta a utilizar es MapReduce; la segunda técnica es "Carga de transmisión" que procesa los datos en tiempo real con la herramienta Apache Spark, y los datos en diferido los procesa con la herramienta Apache Flume. En ambas técnicas la conexión debe ser persistente con las fuentes de datos.

Para el caso de una compañía de seguros de salud en el contexto ecuatoriano existen 12 fuentes de datos comunes y se recomienda el uso de la "Carga de transmisión" para la base de datos, la técnica de la "Carga por lotes" para los archivos que están en hojas electrónicas o procesadores de palabras (variedad), debido a que estos archivos se generan a diferentes velocidades (velocidad); la herramienta de almacenamiento frecuente es HDFS para el procesamiento de los archivos distribuidos en diferido semanalmente, Hadoop es el *framework* libre que permite ejecutar aplicaciones en clusters y MapReduce para el soporte de la computación paralela y distribuida de los macrodatos, Kafka es un estándar para recolectar mensajes como clics o comportamientos de usuarios en las páginas web, y finalmente Apache Streaming para procesar los conjuntos de mensajes.

iii) Valorizar los datos.- Los datos sin procesar que están almacenados deben ser explorados para diseñar un modelo de datos multidimensional que genere valor a los datos, y estos datos recolectados en la segunda fase tienen entidades y relaciones no visibles o sin identificarse; el modelo de datos da a conocer estas entidades y relaciones. El Modelado Multidimensional es: simple porque divide el problema en hechos y dimensiones, es eficiente porque implementa un modelo conceptual, y se basa en un descubrimiento iterativo a través de la exploración de los datos sin procesar. El modelo diseñado debe responder incógnitas relevantes para la empresa.

1) Explorar las fuentes de datos sin procesar: Aquí se identifica el requerimiento de información "para analizar los reclamos de clientes" (hecho), mediante la "facturación promedio" (medida) en función de: póliza, reclamos, renovación, pagos, clientes, dependientes, zona, agencia, fecha, tipo transacción (dimensiones); las consultas deben devolver la facturación, reclamos, pagos o renovaciones media agregada por fecha, y esta consulta permite plasmar con el requisito de información "analiza los reclamos de clientes", ver figura IV.

2) Integración: Aquí se propone integrar los varios modelos conceptuales que sean similares en sus hechos. La integración se repite iterativamente al momento de incorporar las fuentes de datos, entonces el resultado debe ser el modelo conceptual en Big Data, por esta razón se determinaron cuatro dimensiones y con ellas se realiza el análisis de las medidas.

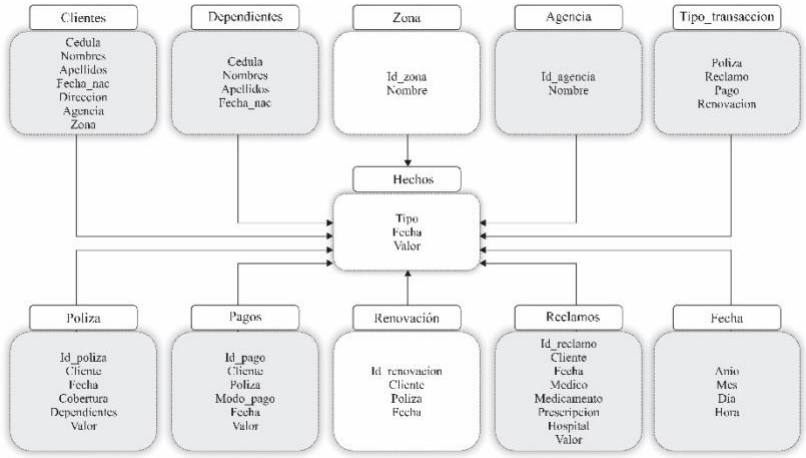

Figura IV.- Modelo conceptual del Big Data Warehouse

3) Enriquecimiento de Big Data Warehouse: Con el modelo conceptual se puede entregar consultas a través de Apache Pig o Hive para consultar datos sin procesar, además este almacén es flexible, es decir, el modelo puede ser actualizado y puede ser enriquecido con dos alternativas como adicionar otros datos fuentes y aplicar minería de datos.

Para el caso de una compañía de seguro de salud en general, para adicionar otros datos se consideran *brokers* (personas intermediarias en operaciones transaccionales en el ámbito financiero), vendedores independientes, sucursales, puntos de venta, puntos de reclamos y redes sociales. Para minería de datos se evidencia que hay diferentes tipos de clientes como personas o empresas, y se obtiene una nueva jerarquía, ejemplo, zona → tipo de cliente. En esta propuesta de investigación se aporta con una forma sistemática de reducir la complejidad en la integración de datos que se presenta como un desafío o problema de Big Data, además este estudio se enfoca en requisitos analíticos para cumplir y responder con criterios validados de datos relevantes.

iv) Selección del almacén de datos. - Se selecciona un Big Data Warehouse para su posible y sencilla implementación, este almacén debe soportar herramientas de Inteligencia Empresarial y tener buena respuesta en la latencia de consultas. Para el caso de una compañía de seguros de salud, se selecciona Apache Hive con herramientas de Inteligencia Empresarial en memoria, que gestionan aplicaciones OLAP y Dashboard, debido a que en esta fase se debe implementar el modelo conceptual con el uso de la herramienta Hive luego, transformar los datos del esquema conceptual en herramienta de Apache Pig finalmente, es posible consultar y analizar el Big Data o pasar a una base de datos en memoria.

v) Implementación visual del Big Data.- Con todos los datos cargados en el almacén de datos, estos se utilizan a través de las herramientas de Inteligencia Empresarial para diseñar y visualizar tablas o cuadros de mando, debido a que estas herramientas generan más valor e interpretación a los datos en forma práctica y sencilla. En este caso de la compañía de seguros de salud se propone el uso de Power BI y con ello, la verificación de los siguientes indicadores:

- Indicadores área de ventas: mide la cantidad de prima (valor de pólizas) ingresado por mes, por agencia y por zona para determinar la región y agencia que está vendiendo más mensualmente, también se suele realizar revisiones anuales para promociones.
- Indicadores de área de reclamos: mide la cantidad de prima de la póliza versus la cantidad de reclamos, esta comparación es anual de acuerdo con la vigencia que tenga la póliza, este indicador evalúa el comportamiento de ciertas carteras de clientes para brindar

incentivos adicionales.

- Indicadores del área de grupos: mide la cantidad de prima que ingresa por todo el grupo versus la cantidad de reclamos de todo el grupo, este indicador evalúa el comportamiento del grupo con el fin de renovarlo o mejorar sus beneficios.
- Indicadores del área de retenciones: mide la cantidad de clientes que son retenidos o clientes con intención de cancelar su póliza.
- Indicadores del área de renovaciones: mide la cantidad de clientes que solicitan actualización de cobertura durante su periodo de renovación anual.
- Indicadores del área de pagos: mide los clientes, agencias o carteras que tienden a no cumplir con sus pagos a tiempo.
- Indicadores de cancelación: mide las zonas, agencias o carteras que tienden a cancelar su cobertura antes de su finalización.

Se especifica que los datos médicos son confidenciales, y como tal no se publican indicadores referentes a esta clase de datos, si algún investigador decide implementar el proyecto en la compañía de seguros solo puede ser utilizado por esa compañía.

La figura V representa el modelo de arquitectura basada en la metodología utilizada. En la *capa fuentes de datos* están los diferentes tipos de datos de las distintas áreas físicas que contiene la compañía, los detalles de estos datos fuentes fueron descritos en la sección "Datos generales en las compañías de seguros de salud", estos datos brutos sirven para a la fase de transformación. En la *capa transformación* se realiza el proceso ETL (*Extraction, Transformation and Load*) que requiere mucho tiempo de procesamiento; la extracción de datos de sistemas de origen que pueden ser bases de datos, archivos y otros tipos o diferentes sistemas, utiliza Hadoop, MapReduce y Kafka; la transformación de datos contiene una serie de funciones y reglas a través de las cuales deben pasar los datos, como limpiar, conciliar, dividir, seleccionar, combinar, estandarizar, eliminar datos duplicados entre otros, para ello se utiliza Apache Streaming; la carga de datos es el paso al almacén de datos desde la capa anterior, para esto se utiliza herramienta HDFS. Además, se realiza el proceso de envío al almacén de datos en un formato unificado, se utiliza Apache Pig y Hive. En la capa *herramientas* se nombra el software que se utiliza en todo el entorno desde datos fuentes hasta presentar los indicadores. En la capa *visualización* se presentan los datos sumarizados en reportes, tablas o gráficos mediante herramienta Power BI, el personal responsable aplica analítica de datos.

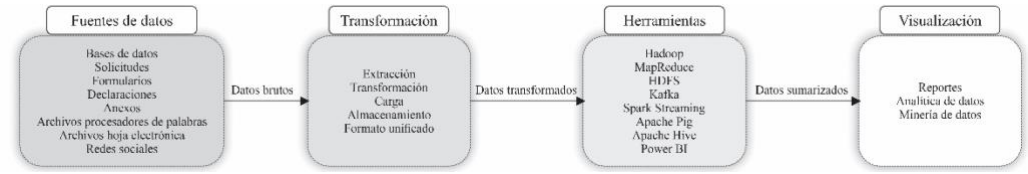

Figura V.- Modelo de arquitectura en empresa ecuatoriana de seguros de salud

4.3 Evaluar la metodología de estudio para establecer factores viables del modelo aplicable a una empresa ecuatoriana de seguros de salud mediante la contrastación de trabajos previos. - Para evaluar los factores viables del modelo sobre una empresa ecuatoriana de seguros de salud, se considera los siguientes niveles: Bajo, Medio y Alto, además si es o no aplicable en una empresa, ver tabla II. El 65% son de Alto factor aplicable (34 factores), entre los factores destacan las investigaciones que ofrecen seguridad, presentan arquitecturas/modelos, buscan disminuir costos, minimizar fraudes/errores, y utilizan herramientas de libre acceso. El 8% son de Medio factor aplicable (4 factores), entre los factores destacan los accesos a datos confidenciales del paciente,

aunque la compañía de seguros es dueña de esta clase de datos, y son aplicables en la compañía. El 27% son de Bajo factor aplicable (14 factores), entre los factores se destaca la aplicación de algoritmos de Inteligencia Artificial porque no son objeto de estudio en esta investigación. El 73% de los factores son viables en la compañía de seguros de salud (38 factores) es decir, factores de Alta y Media aplicabilidad se consideran de utilidad en la empresa. El 27% de los factores son no viables (14 factores) es decir, factores de Baja aplicabilidad porque no se pueden utilizar en la empresa. Entre los 23 trabajos previos contrastados: 12 investigaciones son completamente aplicables, 8 investigaciones no son completamente aplicables, y 3 investigaciones no se pueden utilizar o aplicar.

5. Discusión. - Nuestra investigación se enfoca en Big Data para empresas de seguros de salud, por esta razón se categorizaron diferentes arquitecturas computacionales. Se diseñó un modelo de arquitectura de un sistema computacional, y establecieron factores viables del modelo aplicable a una empresa ecuatoriana, todo esto basado en investigaciones científicas y relevantes. Los tres resultados de esta investigación están enlazados por la literatura analizada y utilizada. Entre las investigaciones, se conoció que cuatro de ellas aplican algoritmos de Inteligencia Artificial para encontrar patrones o tendencias, los datos provienen de una Big Data en área de seguros de salud. El modelo de arquitectura es teórico. Una limitante es que la empresa no tenga computación distribuida y otros recursos para cargar el Big Data Warehouse. Otras limitantes son desconfianza en el sistema, habilidades en la operación del sistema, fragmentación de datos y concientización sobre el uso de datos en salud. Las herramientas nombradas en el modelo de arquitectura son de libre acceso o licencia libre. No es objeto de esta investigación conocer los tiempos, ni talento humano, ni especificaciones de equipos requeridos para la implementación de la arquitectura.

Para continuidad de esta investigación se propone un prototipo de Big Data, que tome como datos todas las fuentes de la empresa de seguros e implementada con las herramientas de software nombradas en el modelo de arquitectura.

Art	Factores	Aplicable	Empresa
[1]	Garantiza la seguridad y privacidad	Alto	Sí
	Adopta estándares	Alto	Sí
[3]	Aplica algoritmo de predicción	Bajo	No
	Presenta modelo de predicción	Bajo	No
[5]	Disminuye costos	Alto	Sí
	Conoce comportamiento de pacientes y proveedor de salud	Alto	Sí
	Presenta herramientas de software que utiliza	Alto	Sí
	Dirigido al sector público o privado	Alto	Sí
[7]	Revisa la medicación de los pacientes	Medio	Sí
	La compañía de seguros entrega el acceso a los datos	Medio	Sí
	Aplica un modelo de estados	Medio	Sí
[11]	Revisa los datos en los reclamos para reducir fraudes	Alto	Sí
	Utiliza un framework	Alto	Sí
[14]	Analiza los sentimientos expresados en una red social	Alto	Sí
	Usa herramientas libres	Alto	Sí
	Aplica algoritmo de Machine Learning	Bajo	No
[26]	Usa metodología para aplicar análisis en Big data	Alto	Sí
[30]	Analiza los gastos anuales de hospitales	Alto	Sí
	Utiliza un algoritmo de árbol de regresión	Bajo	No
[31]	Localiza los posibles fraudes en los reclamos	Alto	Sí
	Utiliza un algoritmo árbol de decisión	Bajo	No
[32]	Descubre fraudes, derroches o sobre precios	Alto	Sí
	Genera modelos predictivos e inteligencia de negocios	Alto	Sí

	Alerta al personal administrativo	Alto	Sí
[33]	Clasifica enfermedades	Alto	Sí
	Aplica algoritmo supervisado de inteligencia artificial	Bajo	No
[34]	Encuentra errores de datos o datos falsificados;	Alto	Sí
	Combina un algoritmo de árbol con un clasificador neuronal	Bajo	No
[35]	Evidencia patrones de enfermedades	Alto	Sí
	Aplica algoritmo no supervisado de inteligencia artificial	Bajo	No
[36]	Determina causas que inciden en la salud	Alto	Sí
	Aplica algoritmo de Machine Learning	Bajo	No
[37]	Análisis de salud	Alto	Sí
	Predice riesgos	Bajo	Sí
[38]	Conoce el gasto en salud	Alto	Sí
	Presenta un framework	Alto	Sí
[39]	Identifica fraudes	Alto	Sí
	Conoce interesados en el sistema de salud	Alto	Sí
[40]	Controla gastos financieros	Alto	Sí
	Presenta modelo basado en Big data	Alto	Sí
	Presenta herramientas de software para su uso	Alto	Sí
[41]	Previene errores	Alto	Sí
	Presenta estructura del modelo	Medio	No
	Aplica algoritmo de Machine Learning	Bajo	No
[42]	Reduce la tasa de contribución	Bajo	No
	Modelo matemático	Bajo	No
[43]	Seguimiento de salud	Alto	Sí
	Aplica analítica predictiva	Bajo	No
[44]	Preserva la confidencialidad	Alto	Sí
	Presenta arquitectura	Alto	Sí
[45]	Preserva la seguridad	Alto	Sí
	Presenta índices de gestión	Alto	Sí

Tabla II.- Factores viables

6. Conclusiones. - En este trabajo se obtienen las siguientes categorías, prevención de enfermedades, causas de reclamos, expresión de sentimientos, fraudes en reclamos, predicción de enfermedades y seguridad de datos, desde las arquitecturas computacionales para seguros de la salud mediante una revisión de literatura relevante. Se desarrolla un modelo de arquitectura de un sistema computacional que consiste en 4 capas: Capa Fuentes de Datos, Capa Transformación, Capa Herramientas y Capa Visualización, el diseño es para una empresa ecuatoriana de seguros de salud y está orientado a la consolidación de la información.

Se evalúa la metodología de estudio mediante la contrastación de trabajos previos, y se conoce que el 73% de los factores (Alto y Medio) son viables en una empresa ecuatoriana de seguros de salud, es decir, son de aplicabilidad para la empresa. El 27% de los factores (Bajo) son no viables, es decir no se pueden aplicar en la empresa. El aporte de este trabajo permite determinar aquellos factores viables que determinan la aplicabilidad del modelo a empresas de seguros de salud nacionales o extranjeras inclusive.

7. Referencias

[1] I. Olaronke and O. Oluwaseun, "Big data in healthcare," in *FTC*, Dec. 2016, no. December, pp. 1152–1157, doi: 10.1109/FTC.2016.7821747.

[2] G. Melendrez-Caicedo and J. Llerena-Izquierdo, "Secure Data Model for the Healthcare Industry in Ecuador Using Blockchain Technology," *Smart Innov. Syst. Technol.*, vol. 252, pp. 479–489, 2022, doi: 10.1007/978-981-16-4126-8_43.

[3] A. Alim and D. Shukla, "A parameter estimation model of big data setup," *IDEA*, 2020, doi: 10.1109/IDEA49133.2020.9170664.

[4] A. de la Nube Toral Sarmiento *et al.*, "4to. Congreso Internacional de Ciencia, Tecnología e Innovación para la Sociedad. Memoria académica," 2018. .

[5] S. Gupta and P. Tripathi, "An emerging trend of big data analytics with health insurance in India," in *ICICCS*, Feb. 2016, no. Iciccs, pp. 64–69, doi: 10.1109/ICICCS.2016.7542360.

[6] R. Ayala Carabajo and J. Llerena Izquierdo, "Tercer Congreso Internacional de Ciencia, Tecnología e Innovación para la Sociedad." 2017, [Online]. Available: https://dspace.ups.edu.ec/handle/123456789/14450.

[7] K. Umemoto and K. Goda, "A Prescription Trend Analysis using Medical Insurance Claim Big Data," in *ICDE*, Apr. 2019, vol. 2019-April, pp. 1928–1939, doi: 10.1109/ICDE.2019.00209.

[8] R. Ayala Carabajo and J. Llerena Izquierdo, "Segundo Congreso Salesiano de Ciencia, Tecnología e Innovación para la Sociedad." 2016, [Online]. Available: https://dspace.ups.edu.ec/handle/123456789/12776.

[9] J. Llerena-Izquierdo and M. Merino-Lazo, "Aplicación móvil de control nutricional para prevención de la anemia ferropénica en la mujer gestante," *Rev. InGenio*, vol. 4, no. 1, pp. 17–26, 2021, doi: 10.18779/ingenio.v4i1.364.

[10] W. M. Soto Eras, "Desarrollo del portal web de la fundación nuestra Señora del Cisne para la gestión de servicios en el Cantón Durán," 2021. .

[11] S. Kareem and R. Binti Ahmad, "Framework for the identification of fraudulent health insurance claims," in *ICBDA*, Nov. 2017, vol. 2018-Janua, pp. 99–104, doi: 10.1109/ICBDAA.2017.8284114.

[12] N. M. Morán Maldonado, "Estado de la Ciberseguridad en las Empresas del Sector Público del Ecuador: Una Revisión Sistemática," 2021.

[13] C. O. Sánchez Guzmán, "Modelo de red segura en un entorno distribuido para la transferencia de datos con mecanismos básicos de seguridad," 2021. .

[14] M. A. Laagu and A. Setyo Arifin, "Analysis the Issue of Increasing National Health Insurance Rates," *ICoSTA*, 2020, doi: 10.1109/ICoSTA48221.2020.1570615599.

[15] Y. J. Terán Terranova, "Seguridad en la Gestión de la información para las organizaciones públicas desde el enfoque ISO/IEC 2700: un Mapeo Sistemático," 2021. .

[16] D. Kim and K. P. Joshi, "A Semantically Rich Knowledge Graph," in *IEEE*, May 2021, pp. 7–12, doi: 10.1109/BigDataSecurityHPSCIDS52275.2021.00013.

[17] C. J. Guaigua Bucheli, "Algoritmos de seguridad para mitigar riesgos de datos en la nube: un mapeo sistemático," 2021. .

[18] J. G. Ponce Larreategui, "Indicadores de compromiso (IOC) para detección de amenazas en la seguridad informática con enfoque en el código malicioso," 2021.

[19] Superintendencia-de-Compañias-Valores-y-Seguros, "Companias de Medicina Prepagada Ecuador," *Listado*, 2021. https://appscvsmovil.supercias.gob.ec/portaldeinformacion/cias_medicina_prepagada.zul (accessed Sep. 27, 2021).

[20] M. J. Aguirre Sánchez, "Tecnologías de Seguridad en Bases de Datos: Revisión Sistemática," 2021. .

[21] G. Ecuatoriano, "Servicio Publico para pagos de accidentes de transito," 2021.

https://www.gob.ec/sppat/tramites/solicitud-pago-proteccion-gastos-medicos-victimas-accidentes-transito (accessed Sep. 27, 2021).
[22] M. J. Chévez Morán, "Estudio de los patrones de seguridad para la atenuación de las irregularidades, las debilidades y amenazas en empresas de servicios de telecomunicaciones," 2021. .
[23] O. A. Escalante Quimis, "Prototipo de sistema de seguridad de base de datos en organizaciones públicas para mitigar ataques cibernéticos en Latinoamérica," 2021. .
[24] I. N. Coello Ochoa, "Análisis de ciberataques en organizaciones públicas del Ecuador y sus impactos administrativos," 2021. .
[25] C. A. Orozco Bonilla, "Estrategias algorítmicas orientadas a la ciberseguridad: Un mapeo sistemático," 2021.
[26] R. Tardio and A. Mate, "An iterative methodology for big data management," in *IEEE*, Oct. 2015, pp. 545–550, doi: 10.1109/BigData.2015.7363798.
[27] P. S. Muñoz Campuzano, "Modelos de seguridad para prevenir riesgos de ataques Informáticos: Una revisión sistemática," 2021.
[28] N. A. Vera Navas, "Modelo de seguridad informática para riesgos de robo de información por el uso de las redes sociales," 2021.
[29] J. N. Miranda Jiménez, "Mapeo sistemático de metodologías de Seguridad de la Información para el control de la gestión de riesgos informáticos," 2021.
[30] Y. Xie *et al.*, "Predicting Days in Hospital Using Health Insurance Claims," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 4, pp. 1224–1233, Jul. 2015, doi: 10.1109/JBHI.2015.2402692.
[31] W. Yang and W. Hu, "Research on Algorithm for Health Insurance Data Fraud Detection," in *IEEE*, May 2021, pp. 57–62, doi: 10.1109/BigDataSecurityHPSCIDS52275.2021.00021.
[32] U. Srinivasan and B. Arunasalam, "Leveraging Big Data Analytics to Reduce Healthcare Costs," *IT Prof.*, vol. 15, no. 6, pp. 21–28, Nov. 2013, doi: 10.1109/MITP.2013.55.
[33] L. Jiang, "Quantitative Model of Insurance Risk Management System Based on Big Data," in *ICSCSE*, Nov. 2016, pp. 590–593, doi: 10.1109/ICSCSE.2016.0159.
[34] A. Sysoev and R. Scheglevatych, "Combined Approach to Detect Anomalies in Health Care Datasets," in *SUMMA*, Nov. 2019, pp. 359–363, doi: 10.1109/SUMMA48161.2019.8947605.
[35] S. Zahi and B. Achchab, "Clustering of the population benefiting from health insurance," in *ICSCA*, Oct. 2019, pp. 1–6, doi: 10.1145/3368756.3369103.
[36] Y. Katsis and N. Balac, "Big Data Techniques for Public Health," in *IEEE/ACM*, Jul. 2017, pp. 222–231, doi: 10.1109/CHASE.2017.81.
[37] S. Karim and E. Gide, "The Impact of Big Data on Health Care Services in Australia," in *ICMSTTL*, 2019, pp. 34–38, doi: 10.1145/3348400.3348414.
[38] Y.-K. Chen and Y.-H. Tao, "Analyzing the Healthcare Expenditure of National Health Insurance," in *ICAMCM*, Dec. 2014, pp. 353–355, doi: 10.1145/2684103.2684178.
[39] V. Chandola and S. R. Sukumar, "Knowledge discovery from massive healthcare claims data," in *ACM*, Aug. 2013, vol. Part F1288, pp. 1312–1320, doi: 10.1145/2487575.2488205.
[40] Y. Liu and J. Peng, "Big Data Platform Architecture," *BDET*, pp. 31–35, 2018, doi: 10.1145/3297730.3297743.
[41] M. Kumar and R. Ghani, "Data mining to predict and prevent errors in health insurance claims processing," in *ACM*, 2010, p. 65, doi: 10.1145/1835804.1835816.
[42] Y. Chen, K. She, and S. Zhao, "Social Insurance Contribution Rate Reduction and Firm Activity Evidence," in *ICSEIM*, Jan. 2021, no. 24, pp. 113–117, doi: 10.1145/3451471.3451490.
[43] X. Deng and D. Wu, "Senior health management through internet of things and real-time big data analytics," in *ACM*, Sep. 2015, pp. 674–674, doi: 10.1145/2808719.2816981.
[44] A. Sara and T. Yassine, "Secure confidential big data sharing in cloud computing," in *iCBDCA*, Mar. 2017, vol. Part F1294, pp. 1–4, doi: 10.1145/3090354.3090388.

[45] L. Kantner and S. D. Goold, “Web tool for health insurance design by small groups,” in *CHI*, Apr. 2006, pp. 141–146, doi: 10.1145/1125451.1125484.

Nota contribución de los autores:

1. Concepción y diseño del estudio
2. Adquisición de datos
3. Análisis de datos
4. Discusión de los resultados
5. Redacción del manuscrito
6. Aprobación de la versión final del manuscrito

JZP ha contribuido en: 1, 2, 3, 4, 5 y 6.

JLI ha contribuido en: 1, 2, 3, 4, 5 y 6.

Nota de aceptación: Este artículo fue aprobado por los editores de la revista Dr. Rafael Sotelo y Mag. Ing. Fernando A. Hernández Gobertti.