Opuntia Brava
ISSN: 2222-081X
RNPS: 2074
Volumen: 12 Número: 1
Recepción: 18/07/19. Aprobado: 08/12/19.

Proceso de diseño de una arquitectura Big Data para el análisis de grandes volúmenes de datos e información

Design process of a Big Data architecture for analysis of large volumes of data and information

Miguel Quiroz Martinez[1] (mquiroz@ups.edu.ec) https://orcid.org/0000-0002-8369-1913

Ricardo Andrés Aguilar Duarte[2] (raguilar@est.ups.edu.ec) https://orcid.org/0000-0002-7934-0440

Diego B. Intriago Cedeño[3] (dintriago@est.ups.edu.ec) https://orcid.org/0000-0002-1389-8740

Resumen

El objetivo de este artículo es presentar un diseño de una Arquitectura Big Data para entidades financieras, que permita el análisis de grandes volúmenes de datos e información y propicie una mejor toma de decisiones y en menor tiempo. Para ello fueron empleados diversos métodos y técnicas científicas que permitieron el análisis, extracción de información y la validación de la arquitectura propuesta. Esta se divide en tres partes: la obtención de datos de manera estructurada y no estructurada de diferentes fuentes, el procesamiento de datos en tiempo real, a partir del empleo del clúster de Hadoop, y el análisis, visualización y toma de decisiones, a partir del procesamiento analítico en líneas y las técnicas de aprendizaje automático. Además, se generó un conjunto de indicaciones para la implementación de la arquitectura Big Data diseñada en entidades financieras. Finalmente, se validó la Arquitectura Big Data diseñada para entidades financieras a partir de criterio de expertos, en lo que se evidenció su pertinencia.

Palabras claves: Análisis de información, arquitectura, Big Data, entidades financieras, toma de decisiones.

Abstract

The objective of this article is to present a design of a Big Data Architecture for financial institutions, which allows the analysis of large volumes of data and information and promotes better decision making in less time. For this purpose, several scientific methods and techniques were used to allow the analysis, information extraction and validation of the proposed architecture. This is divided into three parts: obtaining data in a structured and unstructured manner from different sources, processing data in real time, using the Hadoop cluster, and analysis, visualization and decision making, using online analytical processing and automatic learning techniques. In addition, a set of guidelines was generated for the implementation of the Big Data architecture designed

[1] Profesor Investigador. Universidad Politécnica Salesiana, Sitio Guayaquil. Guayaquil, Ecuador.
[2] Estudiante Investigador, Universidad Politécnica Salesiana, Sitio Guayaquil. Guayaquil, Ecuador.
[3] Estudiante Investigador, Universidad Politécnica Salesiana, Sitio Guayaquil. Guayaquil, Ecuador.

in financial institutions. Finally, the Big Data Architecture designed for financial entities was validated based on expert criteria, in which its relevance was demonstrated.

Key words: Information analysis, architecture, Big Data, financial entities, decision making.

En la actualidad, las organizaciones empresariales incrementan sus esfuerzos por reducir los costos y aumentar el binomio calidad-efectividad en los procesos de atención a la población, el sector financiero no está ajeno a ello (Kim, Trimi y Chung, 2014; Kshetri, 2016; Campbell-Verduyn, Goguen y Porter, 2017). Con el desarrollo del internet y la cantidad de datos que se generan a diario, las empresas han migrado sus necesidades a las tecnologías de la información, para encontrar variantes que propicien darles solución a sus problemas. Uno de estos es el análisis de la gran cantidad de datos e información que se disponen, para mejorar su toma de decisiones, hacerla más efectiva, en el menor tiempo posible y con menor gasto de recursos (Tabares y Hernández, 2014).

Big Data es un paradigma surgido en la última década, término que describe el gran volumen de datos, tanto estructurados como no estructurados, que inundan los negocios cada día (Gandomi y Haider, 2015; Marz y Warren, 2015*).* También se refiere a la velocidad con la que pueden ser procesados esos datos, la variedad o heterogeneidad de los mismos, el valor que se obtiene por la información extraída de los datos y su veracidad. Para ello, se pueden realizar análisis que van desde la aplicación de técnicas de minería de datos, de aprendizaje automático, hasta investigación operacional, entre otras, en dependencia siempre del tipo de análisis a realizar (Fernández, 2017).

El Instituto de Ingeniería del Conocimiento lo conceptualiza como toda clase de técnicas que permiten el tratamiento de grandes volúmenes de datos, fuera de los análisis y herramientas clásicas, que tiene como objetivo común: extraer información de valor de los datos, de forma que pueda ser de ayuda para las decisiones y procesos de negocio (IIC, 2019). En poco tiempo ha dejado de ser una tecnología del futuro para convertirse en una tendencia que ya aporta importantes beneficios a muchas áreas de la sociedad, como el sector financiero (Kim, Trimi y Chung, 2014; Obermeyer y Emanuel, 2016; Yaqoob y otros, 2016).

El empleo de Big Data permite el desarrollo de sistemas revolucionarios para mejorar la manera en que se toman las decisiones y beneficiar tanto a personal administrativo como a la población (Rodríguez, Torre y Garrote, 2014; Hernández y Hernández, 2015; Aguilar, 2016). Todo ello hace necesario el empleo de nuevas y potentes herramientas y tecnologías computacionales que posibiliten procesar este volumen exponencial de información, para analizarla y tomar decisiones cada vez más rápidas y precisas, que impacten positivamente en la sociedad y la economía (Camargo-Vega, Camargo-Ortega y Joyanes, 2015).

El sector financiero ha sido tradicionalmente uno de los más propensos a la inversión en tecnología, especialmente relacionada con los datos (Fernández y Ferrer, 2016). Es por ello que el mercado de soluciones Big Data para este sector se ha concentrado y solidificado en el análisis de créditos, la detección y prevención de fraudes, la retención de clientes, el mercado de variables, la gestión de patrimonios, el control de riesgos y la banca comercial, entre otros (Srivastava y Gopalkrishnan, 2015; Vasarhelyi, Kogan y Tuttle, 2015).

En el contexto del Ecuador, constituye una necesidad la progresiva adopción de Big Data en la banca, si bien su penetración aún no posibilita el aprovechamiento de todas sus potencialidades. Es por ello que la investigación está alineada a los Objetivos de Desarrollo Sostenible (INEC, 2019) y al Plan Nacional de Desarrollo (2017-2021). En ellos se aborda la necesidad de desarrollar las ciudades inteligentes y las infraestructuras, así como la toma de decisiones basadas en los datos, donde el paradigma del Big Data es un factor primordial.

Teniendo en cuenta la situación anterior, el objetivo de este artículo es diseñar una Arquitectura Big Data para una entidad financiera, que permita el análisis de grandes volúmenes de datos e información, y propicie una mejor toma de decisiones.

La investigación realizada comprende un estudio observacional-descriptivo y explicativo, para comprender las problemáticas existentes en entidades financieras. Ello posibilita llevar a cabo un correcto análisis de la información gestionada, para favorecer la toma de decisiones administrativas. Para darle cumplimiento al objetivo general definido, fueron propuestos los siguientes objetivos específicos:

- Analizar el estado de una entidad financiera, respecto a las problemáticas existentes en el procesamiento de datos e información.
- Validar la arquitectura desarrollada a partir de los métodos y técnicas científicas definidas, como las encuestas a expertos con el empleo del Escalamiento de Likert.

En la investigación se aplicó un enfoque mixto, cualitativo y cuantitativo. Para ello se emplearon diversos métodos científicos para la recolección, análisis y procesamiento estadístico de la información, entre los que se destacan:

- Analítico-sintético: Permite analizar los principales conceptos y características del Big Data, así como las principales investigaciones existentes en el área de aplicación. Todo ello posibilita llevar a cabo análisis más profundo para luego llegar a conclusiones que impacten en una arquitectura con mayor calidad.
- Histórico-lógico: Se emplea para investigar los estudios existentes acerca de la evolución y funcionamiento de las arquitecturas Big Data para darle solución a problemáticas similares existentes, así como el análisis de las características de las entidades financieras.

- Inductivo-deductivo: Se utiliza para guiar la investigación desde la definición del objetivo hasta la verificación de la solución a partir de la validación, orientando la secuencia lógica de las tareas que se realizan y que quedan recogidas en la investigación.
- Observación: Se emplea para analizar las características del entorno y de las entidades financieras, para ajustar la propuesta de solución en función de los resultados que se desean obtener.
- Encuestas: Se utiliza para diagnosticar la situación problémica, relacionada con las necesidades existentes de análisis de datos e información en entidades financieras, y para validar el cumplimiento del objetivo propuesto.
- Criterio de expertos: A partir de la consulta de expertos en las áreas de informática y financiera, se evalúan los elementos teóricos que fundamentan la arquitectura Big Data diseñada.
- Análisis documental: Se utiliza para el desarrollo de la investigación, a partir de la consulta de libros y artículos científicos digitales indexados en bases de datos de impacto, publicados en su mayoría en los últimos cinco años, comprendido de 2014 a 2019.

Arquitectura Big Data. Características

A continuación, y como resultado fundamental de la investigación, en la Figura 1 se muestra el diseño de la Arquitectura Big Data para el análisis de grandes volúmenes de datos e información en entidades financieras.

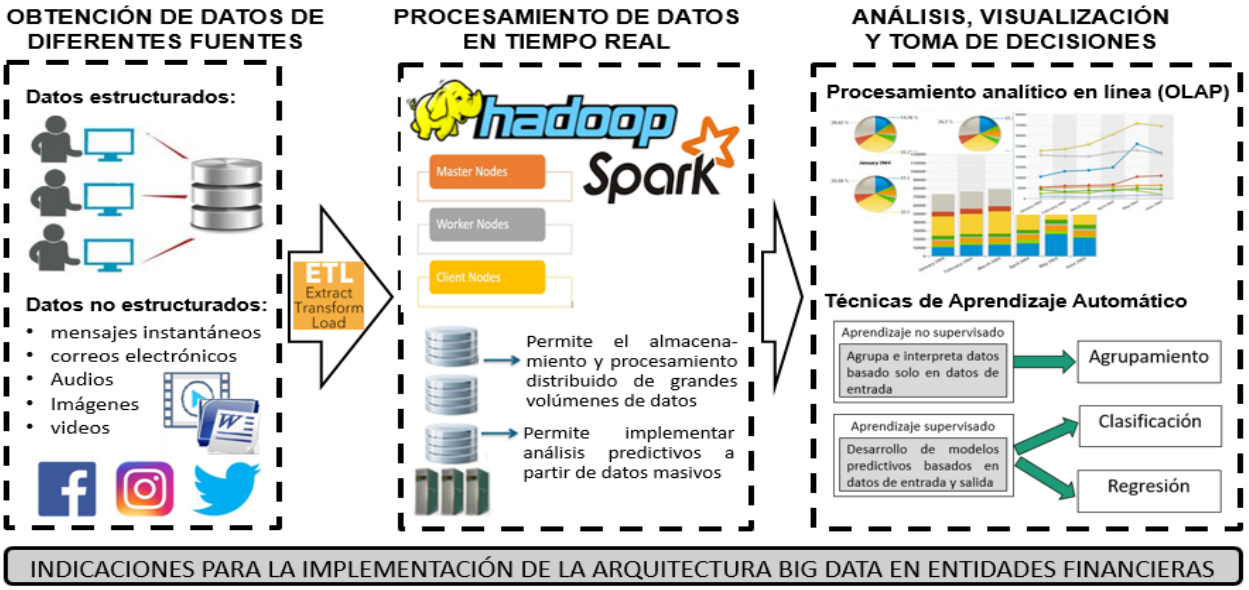

Figura 1. Arquitectura Big Data para el análisis de grandes volúmenes de datos e información en entidades financieras. Fuente: elaboración propia.

La Arquitectura Big Data diseñada para el análisis de grandes volúmenes de datos e información en entidades financieras se divide en tres módulos. Estas se fundamentan en las 3 capas tecnológicas que debe tener todo proyecto de Big Data: almacenamiento, procesamiento y análisis.

- Almacenamiento: Los recursos hardware y software permiten el almacenamiento distribuido y redundante de los datos, lo que facilita su acceso y disponibilidad. Permiten responder la interrogante ¿dónde tener los datos?
- Procesamiento: Las herramientas y técnicas de procesamiento proveen el soporte tecnológico para operar con grandes volúmenes de datos en tiempo real. Permiten responder la interrogante ¿cómo trabajar con los datos?
- Análisis: Los métodos y técnicas computacionales realizan el análisis de los datos, lo que favorece la toma de decisiones oportuna y eficiente en las entidades financieras. Permiten responder a la interrogante ¿qué hacer con los datos?

Los tres módulos en los que se materializan las capas tecnológicas anteriores, se abordan a continuación:

- Obtención de datos de diferentes fuentes: En este módulo se asegura la captación de datos e información de todas las fuentes conocidas y divididas en datos estructurados, datos semiestructurados y datos no estructurados, asociados con la información necesaria a analizar por las instituciones financieras. Esta clasificación depende esencialmente de la forma en que se encuentre el dato:
- Datos estructurados: son datos con estructura conocida. Se encuentran en campos fijos dentro de un archivo o registro.
- Datos no estructurados: información que no tiene una estructura definida. Puede tener un formato textual como: mensajes instantáneos y correos electrónicos; o formato no textual como: audios, imágenes y videos.
- Procesamiento de datos en tiempo real: Para este módulo se adoptaron las potencialidades y ventajas del clúster de Hadoop para el almacenamiento de grandes volúmenes de datos, el cual combinado con el framework Spark posibilitan además el procesamiento en tiempo real. Es por ello que en la arquitectura diseñada se emplearán ambas tecnologías, las cuales pueden coexistir entre sí, lo que constituye un escenario ideal para el trabajo con Big Data. Algunas de las ventajas de esta arquitectura son:

Tabla 1. Características de Apache Spark y Hadoop como escenario ideal para el almacenamiento y procesamiento Big Data. Fuente: elaboración propia.

Hadoop	**Apache Spark**
Ambos son proyectos de Apache y productos de software libre y de código abierto, así como compatibles entre sí.	
Sistema de archivos: Consta de un sistema de archivos distribuido.	Tiempo real: Proporciona procesamiento en tiempo real en la memoria.
Seguridad: Proporciona a sus usuarios todos los beneficios de los avances obtenidos en los proyectos de seguridad de Hadoop	Facilidad de uso: Viene con API fáciles de usar para Scala, Java, Python y Spark SQL.
Procesamiento de datos: Permite el almacenamiento y procesamiento distribuido de grandes volúmenes de datos, así como la ejecución de análisis predictivos.	Velocidad: Es mucho más rápido porque el procesamiento lo hace en memoria.

- Análisis, visualización y toma de decisiones: En este módulo se concibió la utilización de 2 de las principales tecnologías y/o áreas del conocimiento para el análisis y apoyo a la toma de decisiones: el procesamiento analítico en línea (OLAP) y el aprendizaje automático. Sus principales bondades se abordan a continuación:
- Procesamiento analítico en línea (OLAP): Las bases de datos de OLAP facilitan las consultas de inteligencia empresarial. Es una tecnología sofisticada que usa estructuras multidimensionales para proporcionar acceso rápido a datos para su análisis. De esta manera se posibilita la generación de informes detallados, tablas y gráficos dinámicos asociados con los análisis de créditos y datos de la banca comercial, así como el comportamiento del mercado de variables y la gestión de patrimonios.
- Técnicas de Aprendizaje Automático: Este tipo de técnicas de la Inteligencia Artificial se dividen en aprendizaje supervisado y aprendizaje no supervisado. Dentro de estos grupos existen técnicas para la clasificación, regresión y agrupamiento, las cuales permiten un conjunto de análisis, así como la predicción de hechos, que pueden prevenir y detectar la ocurrencia de fraudes, a partir de la aparición de determinados patrones en los datos. De igual manera, posibilita el control de riesgos financieros y la retención de clientes (Sosa, 2011).

Adicionalmente, las cualidades de la Arquitectura Big Data diseñada son (Hernández y Hernández, 2015):

- Consta de un procesamiento distribuido de los datos, a partir de los nodos, que permite la optimización en los tiempos de ejecución de los análisis a realizar.

Miguel Quiroz Martinez
Ricardo Andrés Aguilar Duarte
Diego B. Intriago Cedeño

Opuntia Brava
ISSN: 2222-081X
RNPS: 2074
Volumen: 12 Número: 1
Recepción: 18/07/19. Aprobado: 08/12/19.

- Es escalable, tanto en hardware como en software. Ello le permite aumentar la capacidad de almacenamiento, procesamiento y análisis de ser necesario, sin afectar su funcionamiento en un ambiente real.
- La disponibilidad de la información es una premisa, para lo cual el procesamiento distribuido en nodos constituye otra fortaleza. Ante cualquier afectación de hardware, no existen fallas en el sistema y no se producen pérdidas en los datos.
- La distribución del gran volumen de datos se realiza hacia cada uno de los nodos disponibles, lo que evita la centralización de la información y favorece la rapidez de ejecución y análisis de la información en tiempo real.

Indicaciones para la implementación de la Arquitectura Big Data en entidades financieras

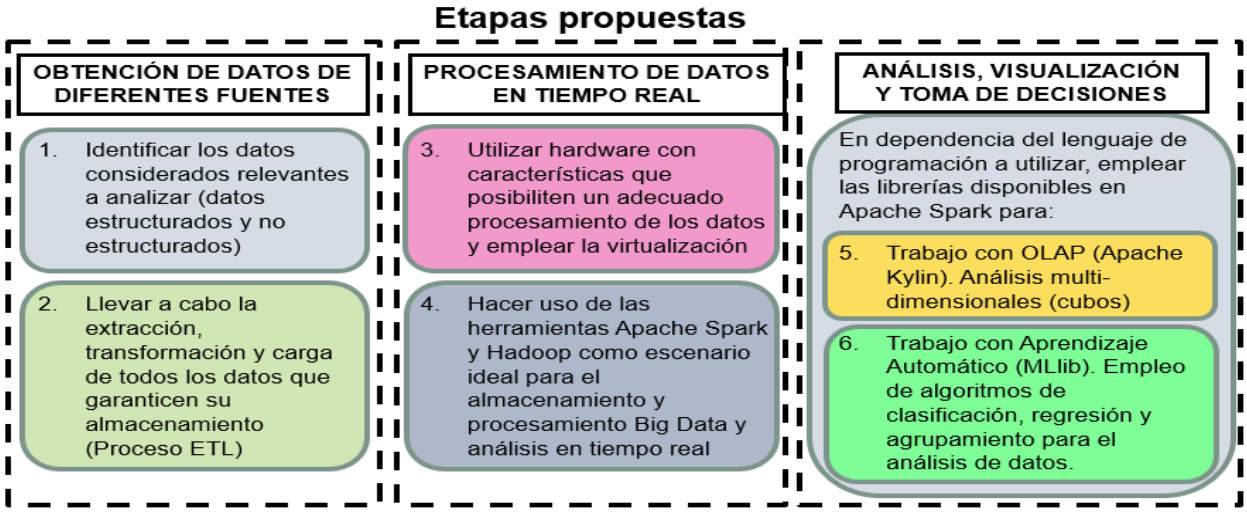

Figura 2. Indicaciones para la implementación de la arquitectura Big Data en entidades financieras. Fuente: Elaboración propia.

Validación de la Arquitectura Big Data diseñada para entidades financieras

La arquitectura diseñada fue validada en varios escenarios, a partir del criterio de expertos, mediante el Escalamiento de Likert y la satisfacción del usuario a partir de la técnica IADOV, donde se evidenció una reacción favorable de los usuarios encuestados. A partir de la relevancia que tiene para la investigación la evaluación de los referentes teóricos y principios evaluados e implementados en la Arquitectura Big Data, se abordará el criterio de expertos realizado.

El método criterio de expertos permite obtener valoraciones de expertos sobre temas relacionados con la propuesta de solución. Como método para el procesamiento estadístico de estos criterios o evaluaciones fue aplicada la escala psicométrica creada por Rensis Likert en 1932 (Ramírez, Estrada, Morejón y Arza, 2017).

Se confeccionó un cuestionario donde se definieron 7 preguntas, las cuales están relacionadas con aspectos relevantes presentes en la arquitectura diseñada, tales como:

- Pertinencia de los módulos propuestos como parte de la arquitectura diseñada.
- Obtención de datos de diferentes fuentes y proceso ETL para la extracción, transformación y carga.
- Empleo de las herramientas Apache Spark y Hadoop para el procesamiento de datos en tiempo real.
- Empleo de técnicas de aprendizaje automático para el análisis de datos, predicción y toma de decisiones.
- Utilización de Procesamiento Analítico en Línea (OLAP) para análisis multidimensionales.
- Impacto y aplicabilidad de la Arquitectura Big Data diseñada para el análisis de datos en instituciones financieras.
- Utilidad de la información visualizada para la toma de decisiones organizacionales.

Se eligieron 26 personas como posibles expertos, se les aplicó la encuesta para determinar el coeficiente de competencia de los expertos, quedando finalmente 22 personas con nivel de competencias alto o medio. Posteriormente, se aplicó el cuestionario y se computaron los resultados. El experto expresa su valoración de cada indicador mediante la siguiente escala: 5- muy de acuerdo (MA), 4- de acuerdo (DA), 3- ni de acuerdo ni en desacuerdo (Sí-No), 2- en desacuerdo (ED) y 1- completamente en desacuerdo (CD).

A continuación, se procesan los resultados mediante la escala Likert. Con esta técnica son calculados los porcientos de concordancia de los expertos con cada una de las posibles respuestas para los planteamientos formulados. Luego se calcula un índice porcentual (IP) que integra en un solo valor la aceptación de cada planteamiento por los evaluadores mediante la siguiente fórmula:

$$IP = \frac{5(\%\text{ de MA})+4(\%\text{ de DA})+3(\%\text{ de Si-No})+2(\%\text{ de ED})+1(\%\text{ de CD})}{5}$$

La Figura 3 muestra que el índice porcentual relacionado con la valoración de los expertos, sobre los aspectos planteados, es superior a 86 en todos los casos, lo cual evidencia la alta valoración de los expertos con el modelo desarrollado.

Miguel Quiroz Martinez
Ricardo Andrés Aguilar Duarte
Diego B. Intriago Cedeño

Opuntia Brava
ISSN: 2222-081X
RNPS: 2074
Volumen: 12 Número: 1
Recepción: 18/07/19. Aprobado: 08/12/19.

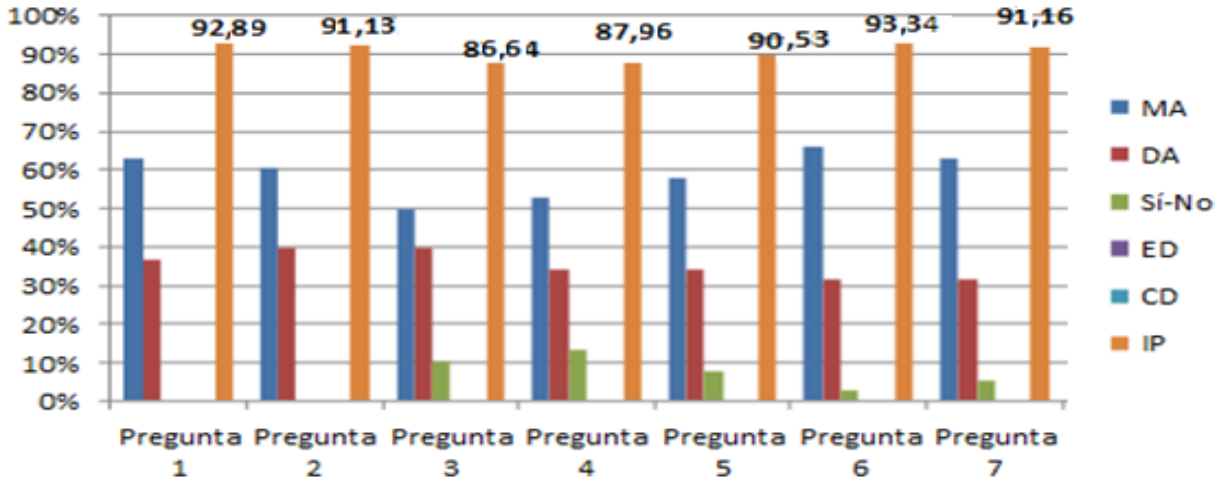

Figura 3. Índice porcentual de concordancia de los expertos. Fuente: elaboración propia.

Con el desarrollo de la investigación se mostró la Arquitectura Big Data diseñada para entidades financieras, la cual posibilita arquitectónicamente el análisis de grandes volúmenes de datos e información. Además, propicia una mejor toma de decisiones, de manera más oportuna y en menor tiempo, todo lo cual impacta en una mayor calidad y efectividad en los procesos de negocios financieros.

Las indicaciones abordadas posibilitan la implementación de la Arquitectura Big Data presentada para el análisis de grandes volúmenes de datos e información en entidades financieras. Para ello, se deben tener en cuenta las premisas abordadas y que definen el correcto funcionamiento de la Arquitectura Big Data.

La profundización en el análisis del estado del arte permitió establecer las bases de la arquitectura presentada, las cuales se fundamentan en la utilización conjunta de Big Data, Procesamiento Analítico en Línea (OLAP) y Machine Learning. Además, se definieron las principales cuestiones de interés y seguimiento en el sector financiero. Estas definiciones que se materializan en la arquitectura Big Data presentada pueden marcar en un futuro la fidelización con el cliente y su seguridad, ya que posibilitan la detección y prevención de fraudes, el análisis de créditos y el control de riesgos, entre otros elementos de gran importancia para clientes y para el propio sector financiero.

La validación de la arquitectura diseñada, a partir de la aplicación de los métodos científicos de criterio de expertos y satisfacción de potenciales usuarios, permitió comprobar la pertinencia y aplicabilidad de la propuesta de solución presentada, así como los principales análisis realizados y resultados arrojados. La misma constituye un aporte práctico importante a las arquitecturas big data existentes en la actualidad.

Referencias

Aguilar, L. J. (2016). *Big Data, Análisis de grandes volúmenes de datos en organizaciones*. México: Alfaomega Grupo Editor.

Camargo-Vega, J. J., Camargo-Ortega, J. F. y Joyanes-Aguilar, L. (2015). Knowing the big data. *Facultad de Ingeniería, 24*(38), 63-77.

Campbell-Verduyn, M., Goguen, M. y Porter, T. (2017). Big Data and algorithmic governance: the case of financial practices. *New Political Economy, 22*(2), 219-236.

Fernández, Y. A. y Ferrer, D. C. (2016). Big Data: una herramienta para la administración pública. *Ciencias de la Información, 47*(3), 3-8.

Fernández, C. O. (2017). Cómo las empresas pueden impulsar su negocio a través de las plataformas e-commerce con el Big Data, el aprendizaje automático y el management científico. *Economía industrial*, (405), 75-86.

Gandomi, A. y Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management, 35*(2), 137-144.

Hernández, A. y Hernández, A. (2015). Acerca de la aplicación de MapReduce+ Hadoop en el tratamiento de Big Data. *Revista Cubana de Ciencias Informáticas, 9*(3), 49-62.

Instituto de Ingeniería del Conocimiento (2019). *¿A qué llamamos Big Data?* Recuperado de http://www.iic.uam.es/big-data

Kim, G. H., Trimi, S. y Chung, J. H. (2014). Big-data applications in the government sector. *Communications of the ACM, 57*(3), 78-85.

Kshetri, N. (2016). Big data's role in expanding access to financial services in China. *International journal of information management, 36*(3), 297-308.

Marz, N. y Warren, J. (2015). *Big Data: Principles and best practices of scalable real-time data systems*. New York: Manning Publications Co.

Obermeyer, Z. y Emanuel, E. J. (2016). Predicting the future - big data, machine learning, and clinical medicine. *The New England journal of medicine, 375*(13), 1216.

Ramírez, J. F., Estrada, V., Morejón, M. y Arza, L. (2017). Modelo para la gestión y análisis de conocimiento para la selección de equipos de trabajo quirúrgico en sistemas de información en salud mediante técnicas de inteligencia organizacional. *Revista cubana de información en ciencias de la salud, 28*(1), 43-60.

Rodríguez, S., Torre, A. I. y Garrote, E. (2014). Tecnologías Big Data para análisis y recuperación de imágenes web. *El profesional de la información, 23*(6).

Sosa, D. C. (2011). Inteligencia artificial en la gestión financiera empresarial. *Revista científica Pensamiento y gestión* (23).

Srivastava, U. y Gopalkrishnan, S. (2015). Impact of big data analytics on banking sector: Learning for Indian banks. *Procedia Computer Science*, 50, 643-652.

Tabares, L. F. y Hernández, J. F. (2014). *Big Data Analytics: Oportunidades, Retos y Tendencias*. Universidad de San Buenaventura.

Vasarhelyi, M. A., Kogan, A. y Tuttle, B. M. (2015). Big Data in accounting: An overview. *Accounting Horizons, 29*(2), 381-396.

Yaqoob, I., Hashem, A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B. y Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management, 36*(6), 1231-1247.