Revista Venezolana de Computación
ISSN: 2244-7040
http://www.svc.net.ve/revecom
Vol. 3, No. 2, pp. 47-56, Diciembre 2016
Selección de los Mejores Artículos de CoNCISa 2016

Arquitectura para la Gestión de Datos Imperfectos en la Era de Big Data

Kity Álvarez[1], Betzaida Romero[1], José Tomás Cadenas[1], David Coronado[1], Rosseline Rodríguez[1]
kjalvarez@usb.ve, betzaidaromero@usb.ve, jtcadenas@usb.ve, dcoronado@usb.ve, crodrig@usb.ve

[1] Departamento de Computación y Tecnología de la Información, Universidad Simón Bolívar, Caracas, Venezuela

Resumen: La gran cantidad de datos que se maneja hoy en día y la manera como estos datos se extraen del contexto obligan a utilizar sistemas de información capaces de tomar decisiones, adecuarlos a la empresa y al usuario final. La digitalización da lugar a distintos tipos de datos en tiempo real de acuerdo al escenario que se plantee. Una gran cantidad de estos datos, en ocasiones se presentan de forma no normalizada: datos en *streaming*, geoespaciales o generados por diferentes tipos de sensores, que no encajan bien en un esquema relacional, tradicional o estructurado. Mucha de esta información puede ser vaga, imprecisa, ambigua o incompleta, muy parecida al lenguaje humano con respecto a términos cualitativos y cuantitativos. El objetivo del presente trabajo es estudiar acerca de la gestión de datos imperfectos sensibles al contexto integrado a la utilización de tecnologías asociadas a *Big Data*. La teoría de conjuntos difusos, en el marco de *Soft Computing*, aporta mecanismos para modelar y representar datos posibilísticos en bases de datos difusas, se propone una arquitectura que gestione este tipo de datos integrada con tecnologías para el manejo de datos masivos. Esta arquitectura incorpora un Módulo de Interacción con una Base de Datos Difusa que permita el almacenamiento y recuperación de datos sensibles al contexto. El propósito es proporcionar una herramienta útil para enfrentar el reto que tienen las organizaciones de obtener un mayor provecho de la gran cantidad de información proporcionada por las tecnologías del mundo actual. Además, se espera obtener los beneficios que agrega la gestión de datos difusos sensibles al contexto en el almacenamiento de datos imperfectos y consultas flexibles que no pueden ser ofrecidas en sistemas de bases de datos tradicionales.

Palabras Clave: Datos Imperfectos; Bases de Datos Difusas; Bases de Datos Sensibles al Contexto; Soft Computing; Big Data.

Abstract: The large amount of data that is handled today and the way these data are extracted from the context require the use of information systems capable of making decisions, adapting them to the company and the end user. Digitization gives rise to different types of data in real time according to the scenario that arises. A large amount of this data, sometimes, is presented in a non-normalized way: data streaming, geospatial or generated data by different types of sensors that do not fit well into a relational, traditional or structured scheme. Much of this information may be vague, imprecise, ambiguous or incomplete, very similar to human language with respect to qualitative and quantitative terms. The objective of the present work is to study of context-aware imperfect data management integrated to the use of technologies associated with Big Data. The fuzzy sets theory, in the framework of Soft Computing, provides mechanisms to model and represent possibilistic data in fuzzy databases. We propose an architecture that manages this data type, integrated with technologies for big data management. This architecture incorporates an Interactive Module with a Fuzzy Database that allows the storage and retrieval of context-aware data. The purpose is to provide a useful tool to face the challenge that has the organizations to make the most of the vast amount of information provided by today's technologies. Also, we expect to obtain the added benefits of context-aware fuzzy data management in storage imperfect data and flexible queries that cannot be offered in traditional database systems.

Keywords: Imperfect Data; Fuzzy Databases; Context-Aware Databases; Soft Computing; Big Data.

I. Introducción

En la actualidad, el tratamiento de los datos ha ido cambiando debido a que el alcance de las aplicaciones va en crecimiento y los retos empresariales al igual que la cantidad de datos disponibles son cada vez mayores. Estos datos provienen de diferentes fuentes y son de naturaleza distinta.

Debido a los avances en las Tecnologías de la Información y Comunicación (TIC), la gran cantidad de información disponible, denominada Big Data [1], constituye una oportunidad sin precedentes, sin embargo, los usuarios de esta información corren el riesgo de verse abrumados, perdiendo la oportunidad de utilizar el valioso conocimiento que puede ser extraído de estos datos. Es así como, las aplicaciones que usan

estas bases de datos, requieren de una gestión inteligente de la información para satisfacer las demandas de almacenamiento y consultas. Esto representa un gran reto para los sistemas informáticos modernos pues los gestores de bases de datos tradicionales no son capaces de satisfacer las necesidades actuales de información de los usuarios.

Con la evolución de los datos masivos (Big Data) [2], las empresas de todo el mundo están descubriendo nuevas formas de competir y ganar. Esto obliga a una comprensión exhaustiva de los mercados, los clientes, los productos, las normativas, los competidores, los proveedores, los empleados, entre otros. Esta comprensión exige un uso eficaz y analítico de la información, que debe ir de la mano de las tecnologías y herramientas adecuadas para el procesamiento del mismo.

Las aplicaciones en el área de Big Data, requieren administrar un volumen de muchos Terabytes de datos cuyo tamaño, complejidad y diversidad, impiden que los Sistemas Gestores de Bases de Datos (SGBD) tradicionales puedan manejarlos eficientemente. Por eso han surgido las Bases de Datos bajo el paradigma NoSQL (No sólo SQL), que permiten resolver problemas de escalabilidad y rendimiento que presentan estos tamaños, así como su gran diversidad, usando nuevos entornos de datos distribuidos y escalables de forma horizontal [1].

Además, con el surgimiento de tecnologías asociadas a Internet de las Cosas (IoT) se puede obtener una gran cantidad datos de diversas fuentes, de forma rápida, confiable y segura. La integración de áreas como sistemas de información, inteligencia artificial, base de datos, sistemas expertos y extracción de conocimiento ha evidenciado la necesidad de gestionar datos imperfectos [3]. Este se considera un tema de investigación importante ya que los datos imperfectos están presentes en muchas aplicaciones del mundo real, la gestión de estos datos contribuyen a comprender y predecir mejor el comportamiento de los usuarios y mejorar sus experiencias [4].

En cuanto a la naturaleza de los datos, se consideran varios tipos de información imperfecta que debe ser incluida en una base de datos, clasificándola como imprecisa, incierta o vaga [5]. La gestión de datos imperfectos es importante en la exploración de información proveniente de las redes sociales y sus tendencias, de fuentes datos sindicados como las tarjetas de fidelidad (RFID), además de información relacionada a la toma de decisiones para la calidad de los servicios prestados a los usuarios.

Hoy en día se considera de vital importancia disponer de técnicas y herramientas de análisis de información que usen grandes repositorios de datos, así como representa un gran reto la gestión de datos imperfectos sensibles al contexto en los Sistemas de Bases de Datos. Por tal razón, el presente artículo se enfoca en integrar la gestión de datos imperfectos, con técnicas que ya han sido empleadas en Bases de Datos Difusas, que tomen en cuenta las nuevas tecnologías que han surgido en el área de Big Data.

La estructura del presente artículo es la siguiente: la sección II trata sobre la Gestión de Datos Imperfectos que constituye el marco teórico de este trabajo. La sección III muestra los retos que presenta el área del Big Data. La sección IV, presenta los antecedentes del trabajo. La sección V, plantea la propuesta de arquitectura para gestión de datos imperfectos sensibles al contexto integrada con tecnologías de datos masivos. Finalmente, en la sección VI, se exponen las conclusiones y trabajos futuros.

II. Gestión de Datos Imperfectos

En esta sección se presentan los fundamentos teóricos de la investigación realizada, que incluye la noción de datos imperfectos y las diversas formas como han sido tratados, la teoría de conjuntos difusos como una alternativa posible y el principio de información planteado por Zadeh.

A. Datos Imperfectos

Debido a la gran variedad de imperfecciones que pueden afectar a los datos, se puede enfocar el análisis de los mismos separando aquellos que tienen una apariencia flexible (*soft*) de aquellos que tienen especificaciones precisas y completamente ciertas. Estos últimos corresponden a la información almacenada en Bases de Datos tradicionales. En [5], se definen tres términos relacionados a la imperfección de los datos: imprecisión, incertidumbre y vaguedad.

La *imprecisión* denota la carencia de exactitud en la expresión de la información, ocurre cuando el dato es desconocido o el dominio del atributo es impreciso por su naturaleza, por ejemplo “la imagen es *vieja*”, “Luis es *joven*”, o “el paciente es *obeso*”. La *incertidumbre* revela una situación en donde no se está seguro acerca de la veracidad de la información, tal como “es *posible* que el carro tenga cinco metros de longitud”, “Luis *probablemente* vive en Maracay”, o “*Creo* que la placa del automóvil que produjo el accidente era DBW50S”.

La *vaguedad* sucede cuando la información es afectada por imprecisión, incertidumbre o ambos; pues a veces es difícil establecer los límites entre imprecisión e incertidumbre. Cuando se usa la expresión “el carro es *grande*” puede ser que la persona no conozca el tamaño del carro con certeza, aunque no lo exprese explícitamente. Algunos autores usan este término como sinónimo de imprecisión.

Es de hacer notar que la imperfección puede incluir otras deficiencias, tal como datos erróneos o inconsistentes debido a: falta de precisión de dispositivos sensores, datos que provienen de fuentes heterogéneas o información resultante de técnicas como la consolidación.

La imperfección en los datos está presente en una gran variedad de aplicaciones emergentes, tales como: Servicios basados en Localización, Sistemas de Información Geográficos (GIS), Redes de Sensores Inalámbricos, Flujo de Datos (*data streaming*), Bases de Datos de uso específico (espaciales, biológicas y biométricas), Extracción de Conocimiento (Minería de Datos, Conocimiento Experto y Ontologías), Inteligencia Ambiental, además de Sistemas de Recomendación y Recuperación de Información. En estas aplicaciones es importante gestionar la imperfección de los datos adecuadamente, a fin de facilitar la toma de decisiones de los usuarios y la dotación de servicios de alta calidad.

La gestión de datos imperfectos ha sido de interés debido a la gran cantidad de áreas donde se usan, tales como: minería de datos preservando privacidad, datos probabilísticos inferidos por métodos de predicción estadística, incertidumbre en bases de datos [1]. Debido a esto, se han diseñado e implementado diferentes propuestas de bases de datos que permitan la

representación y gestión de la imperfección de los datos a través de los Sistemas Gestores de Bases de Datos (SGBD).

Las Bases de Datos tradicionales sólo manejan datos y condiciones precisas que en muchas ocasiones no representan las necesidades reales de información de los usuarios. Los datos faltantes son normalmente manejados a través de un pseudovalor denominado *null*. El uso de este pseudovalor puede deberse a varias razones tales como: el dato es desconocido, es impreciso, incierto o puede que no aplique para en una determinada tupla, entre otras [3].

Una estrategia utilizada regularmente por los manejadores de Bases de Datos tradicionales es la llamada imputación, que consiste en cambiar los valores *null* en la Base de Datos por valores sustitutos elegidos con algún tipo de criterio establecido. En [6] se propone el uso de la técnica de Reglas de Asociación Difusa, para obtener un conjunto de reglas que permitan estimar valores nulos en función de los demás valores presentes en los registros y cuya aplicación se extiende a atributos con valores cuantitativos y no solo categóricos.

Según De Tré y Zadrozny [3] para que las organizaciones sean competitivas tienen que utilizar toda la información disponible incluyendo la imperfecta, la cual no debe ser descartada. Por lo tanto, almacenar de manera eficiente y consultar datos imperfectos, sin introducir errores o causar perdida de datos, es considerado actualmente, como uno de los principales retos para gestión de la información.

Una de las áreas que ofrece formalismos y técnicas matemáticas para enfrentar la gestión de datos imperfectos es *Soft Computing* [7]. Su principal objetivo es aprovechar la tolerancia a la imprecisión, incertidumbre y verdad parcial, con el fin de obtener soluciones computacionales que sean tratables, robustas y de bajo costo. Sus técnicas incluyen Lógica Difusa, Neuro-computación, Razonamiento Probabilístico, Computación Evolutiva, además de otras técnicas como Redes de Creencia, Sistemas Caóticos y Aprendizaje Automático (*Machine Learning*).

Dado que la información imperfecta aparece en dominios y situaciones reales, debido a errores instrumentales o la corrupción por ruido durante la recogida de datos, se puede dar lugar a información incompleta para atributos específicos [8]. Además, la extracción de información exacta puede ser excesivamente costosa, de allí la necesidad de aplicar técnicas de *Soft Computing*.

Motro [8] indica que la imprecisión en los valores de los datos toma diferentes formas. Puede que el valor real del dato pertenezca a un conjunto específico de valores, por lo que es llamado dato disyuntivo. Si el conjunto al cual pertenece es el dominio completo entonces es indisponible, desconocido o perdido. Si cada uno de los valores candidatos está acompañado por un número describiendo la probabilidad que el valor sea verdadero (y la suma de todos ellos es uno) entonces se denomina probabilístico. Ocasionalmente, la información disponible en la ausencia de datos precisos es un término descriptivo, estos datos son denominados difusos.

La incertidumbre debido a ambigüedad o inconsistencia en los datos, asociada con situaciones en la cual se debe escoger entre diversas alternativas precisas, puede ser modelada utilizando la teoría de probabilidad [9]. Este es el caso de la medición de datos que pueden ser obtenidos a través de sensores o como resultado de un proceso de consolidación de datos provenientes de diversas fuentes. Es por ello que se han desarrollado una serie de modelos probabilísticos que pueden ser revisados en profundidad en [10].

Por otro lado, la incertidumbre debida a la vaguedad, es modelada con herramientas como la Teoría de Conjuntos Difusos [11], en la cual se hace énfasis en el significado de los términos imprecisos dados por el ser humano en un contexto específico. Zadeh [12] propuso la Teoría de Posibilidad donde se definen distribuciones de posibilidad como conjuntos difusos que permiten hacer limitaciones flexibles sobre los valores que pueden asignarse a una variable. La importancia de esta teoría reside en el hecho de que la intrínseca borrosidad de los lenguajes naturales, como consecuencia lógica de las expresiones utilizadas es *posibilística* y no probabilística, representando cuantiosa información sobre la cual se basan las decisiones del ser humano en un contexto.

Además, Zadeh [13] afirma que una de las más importantes facetas del pensamiento humano es la habilidad de resumir información en etiquetas de conjuntos difusos (denominadas *etiquetas lingüísticas*), que proporcionan una relación aproximada con los datos originales. Los postulados de Zadeh [13] permiten modelar la percepción y los procesos de pensamiento humano.

Técnicas de computación flexible (*Soft Computing*) hacen que sea posible gestionar la información imperfecta sobre una parte modelada del mundo real y representarla directamente en una base de datos. Si se usa la teoría de conjuntos difusos o la teoría de posibilidad para modelar datos imperfectos en una base de datos, ésta se denomina *difusa*. Otros enfoques incluyen aquellos que se basan en la teoría de conjuntos aproximados (*Rough sets*) y en la teoría de probabilidad, por lo que las bases de datos resultantes son denominadas *aproximadas* y *probabilísticas*, respectivamente.

Las técnicas más importantes de computación flexible para el tratamiento de la información imperfecta, debido a la investigación previa que han realizado diversos autores en el área de bases de datos, están basadas en la Teoría de Conjuntos Difusos y la Teoría de Posibilidad [3]. Es por ello que en la próxima sección se introduce este tema.

B. Teoría de Conjuntos Difusos

La teoría de conjuntos difusos [11] proporciona un marco matemático y computacional formal para representar las nociones de naturaleza vaga o imprecisa. Los conjuntos difusos, extienden el concepto de conjunto clásico, por lo que se caracterizan por una función de pertenencia, sobre un universo, cuyo rango está en el intervalo real [0,1]. Cuanto más se acerca a 1 el grado de pertenencia de un elemento, el mismo está más posiblemente (o certeramente) incluido en el conjunto. Así, 0 es la medida de completa exclusión y 1 la de completa inclusión o pertenencia total. El *borde* se define como el conjunto formado por los elementos parcialmente incluidos, el *núcleo*, como el conjunto de los elementos completamente incluidos y el *soporte*, como el conjunto de los elementos que no están completamente excluidos.

Según Zadeh [7], el término de lógica difusa es utilizado en dos sentidos diferentes. En un sentido limitado es visto como

una extensión de lógica multivaluada con el propósito de servir para el razonamiento aproximado. En su sentido más amplio, es un sinónimo de la teoría de conjuntos difusos. Es importante reconocer que el término lógica difusa es usado predominantemente en su sentido más amplio, así cualquier campo puede ser *fuzzificado*, reemplazando valores de un conjunto preciso por valores de un conjunto difuso.

La lógica difusa es una técnica de inteligencia computacional que permite trabajar con la información con alto grado de imprecisión, en esto se diferencia de la lógica convencional que trabaja con información bien definida y precisa. La lógica difusa permite que haya valores flexibles como: *cerca/lejos*, *grande/pequeño*, *fuerte/débil*, entre otros.

En el conjunto de datos se debe tener en cuenta la imperfección de la información para poder extraer modelos más cercanos a la realidad de acuerdo al contexto. Sin embargo, a pesar de la gran cantidad de técnicas existentes para dicho proceso, es común observar que se siguen utilizando Bases de Datos tradicionales para información clásica, y muy pocas toman en cuentan los datos imperfectos para su posterior análisis.

Por otro lado, la Teoría Computacional de Percepciones (CTP por sus siglas en inglés) propuesta por Zadeh [7], proporciona una base teórica para modelar sistemas complejos que gestionen información imperfecta, convirtiéndose en una poderosa herramienta que permite formalizar procesos de naturaleza humana (tal como conversar, razonar y tomar decisiones). Mediante el uso de etiquetas lingüísticas se pueden representar valores imperfectos para ser almacenados en la base de datos o para hacer posible el uso de términos lingüísticos en las consultas que se asemejan más al lenguaje natural utilizado por los seres humanos, es decir, utilizar el enfoque de representar la información en forma lingüística (computación por palabras) basado en la percepción más que en forma de medidas basadas en números.

Además, es necesario resaltar la habilidad extraordinaria del cerebro humano para manipular percepciones con respecto a diversos aspectos tales como: distancia, tamaño, peso, color, velocidad, tiempo, dirección, fuerza, verdad, probabilidad y otras características de objetos físicos y mentales; jugando un papel crucial en el reconocimiento de patrones, ejecución de actividades y la toma de decisiones.

Un ejemplo del uso de la lógica difusa, en la toma de decisiones, es el escenario en el que existe un entrenador de básquet que debe seleccionar a los candidatos para su equipo [11]. En la solución clásica de este ejemplo, la altura debe ser mayor a 185 cm y debe haber encestado al menos 13 de 16 tiros a cesta.

En la Tabla I se muestran los resultados de la solución clásica, donde sólo se seleccionarían a los candidatos F e I. Sin embargo, se observa que el candidato E tuvo 16 aciertos. La solución que utiliza conjuntos difusos es diferente. Se definen términos lingüísticos para cada variable (estatura y aciertos) y se da una respuesta que usa lógica difusa. Por ejemplo, el candidato exitoso es el que reúne los criterios: estatura *alta* y encestador *bueno*. El resultado se muestra en la Tabla II.

Utilizando lógica difusa se puede lograr una selección con discriminación entre los candidatos (*ranking* valorado) como se observa en la Tabla III. Esto evita dejar fuera del equipo a un buen encestador que mide 183 cm, tal como haría un entrenador. En lugar de tomar una decisión limitada a sólo dos alternativas "se rechaza" o "se acepta" (como el 0 ó 1 de la lógica clásica), se toma en cuenta la gradualidad.

Tabla I: Candidatos para el Equipo de Básquet (Solución Clásica)

CANDIDATO	ESTATURA (CMS)	ACIERTOS (16 TIROS)	SOLUCIÓN (CLÁSICA)
A	167	12	0
B	169	6	0
C	175	15	0
D	179	12	0
E	183	16	0
F	186	13	1
G	187	12	0
H	190	10	0
I	200	13	1

Tabla II: Candidatos para el Equipo de Básquet (Solución Difusa)

CANDIDATO	ESTATURA (CMS)	ACIERTOS (16 TIROS)	LÓGICA DIFUSA
A	167	12	0
B	169	6	0
C	175	15	0,33
D	179	12	0,50
E	183	16	0,87
F	186	13	0,75
G	187	12	0,50
H	190	10	0
I	200	13	0,75

Tabla III: Candidatos para el Equipo de Básquet (*Ranking* Valorado)

CANDIDATO	ESTATURA (CMS)	ACIERTOS (16 TIROS)	LÓGICA CLÁSICA	LÓGICA DIFUSA
E	183	16	0	0,87
F	186	13	1	0,75
J	200	13	1	0,75
D	179	12	0	0,50
G	187	12	0	0,50
C	175	15	0	0,33
A	167	12	0	0
B	169	6	0	0
H	190	10	0	0

La Teoría de Conjuntos Difusos es más adecuada que la Lógica Clásica para representar el conocimiento humano, ya que permite que los fenómenos y observaciones tengan más de dos estados lógicos. La Lógica Difusa produce resultados exactos a partir de datos imprecisos, por lo cual es particularmente útil en aplicaciones electrónicas o computacionales.

Los conjuntos difusos permiten representar y manipular información imperfecta a nivel de atributos [14]. Para ello, se definen dominios difusos de diferentes naturalezas que permiten especificar los diversos atributos de una entidad que presentan imprecisión. También, se definen diferentes operadores difusos para cada uno de estos dominios, tal como la igualdad difusa (*Fuzzy Equal* o FEQ) para comparar datos difusos.

La imperfección inherente al valor de un atributo está dada por la naturaleza del dominio del atributo. Los valores imperfectos en el modelo pueden estar representados por una etiqueta lingüística que dispondrá de diferentes interpretaciones dependiendo de la semántica en el contexto del dominio donde se defina. En base a la cardinalidad se pueden distinguir dos tipos de dominios difusos: *atómicos* o *conjuntivos*.

Los dominios *atómicos* son aquellos donde el atributo puede tomar un único valor que puede o no tener representación semántica asociada. Si no tienen representación semántica se conocen como categóricos. Por ejemplo, cuando se habla de la calidad de un producto el dominio se puede componer de etiquetas lingüísticas tales como *buena*, *regular* y *mala*.

En caso de tener una representación semántica, pueden ser de dos tipos dependiendo del universo sobre el cual estén definidos: *discreto* o *continuo*. Por ejemplo, se puede definir un dominio atómico continuo para el atributo estatura, formado por las etiquetas lingüísticas *gigante*, *alto*, *mediano*, *bajo* y *enano*. Cada una de estas etiquetas es asociada a un conjunto difuso cuyo universo o conjunto soporte es continuo (los números reales).

Otro ejemplo, es el atributo dureza de un mineral, la cual puede ser medida en una escala discreta en el rango de 1 a 10, pero normalmente se usan etiquetas lingüísticas, tales como, *muy dura*, *dura*, *media*, *blanda* y *muy blanda*. En este tipo de dominio, las etiquetas lingüísticas se representan con conjuntos difusos por extensión que indican la asociación de éstas a la escala del 1 al 10, el cual es un universo discreto.

Los dominios difusos donde el atributo puede tener diversos valores representados por un conjunto difuso, se denominan *conjuntivos*. Por ejemplo, para indicar que un investigador domina tres idiomas, se puede representar con un conjunto difuso tal como: {1.0/español, 0.8/inglés, 0.5/italiano}. Aquí el valor que acompaña al idioma determina nivel de experticia que tiene el investigador (mientras más cercano al 1 es más alto, más cercano al cero es más bajo). El universo es el conjunto de idiomas posibles {inglés, francés, ruso, italiano, español} y el valor del atributo es la conjunción de los idiomas que conforman el conjunto difuso.

Debido a la importancia que tiene actualmente la gestión de datos imperfectos en la siguiente sección se estudia el Principio de la Información postulado por Zadeh [13].

C. El Principio de la Información

El Principio de la Información se basa en identificar información con restricción, donde una restricción es una limitación a los valores que una variable puede tomar. Este principio consta de tres partes, a saber:

1) La información es una restricción.

2) Hay tres tipos principales de información, que provienen de los tres tipos principales de restricciones:

- *Información posibilística*: La variable toma el valor de un conjunto difuso.
- *Información probabilística*: La variable toma un valor con cierta probabilidad.
- *Información bimodal*: combinación de información posibilística y probabilística.

3) La información posibilística y la información probabilística no son derivables (son ortogonales), linealmente independientes, en el sentido de que ninguna se puede deducir de la otra.

Zadeh [13] afirma que, a partir de observaciones empíricas, se puede concluir que la mayoría de la información que se maneja a diario es posibilística o bimodal. Además, reconoce que la mayoría de los sistemas existentes no están capacitados para trabajar con la información bimodal. La diferencia entre la información posibilística y la probabilística, es que la posibilidad es la respuesta a la pregunta "¿puede ocurrir?", mientras que la probabilidad responde a "¿con qué frecuencia?". Para que una información sea probable, primero debe ser posible, por lo tanto, se tiene información bimodal.

Además de ser posible, probable y bimodal, la información puede ser precisa o difusa de acuerdo con el tipo de restricción. Por ejemplo, la probabilidad de que la estatura del candidato A está entre 1,65 y 1,69 cm es 0,8, es información bimodal precisa (CBI por sus siglas en inglés *Crisp Bimodal Information*). Por otro lado, decir que es muy probable que el candidato A no sea alto, es información bimodal difusa (FBI, *Fuzzy Bimodal Information*).

A continuación se presenta un ejemplo que permite observar cómo representar la información bimodal. Se quiere formar un equipo pequeño de baloncesto (tres contra tres), con los jugadores de la Tabla I, de la sección anterior. El universo estaría dado por el conjunto U = {A, B, C, D, E, F, G, H, I}. Durante el torneo, después de cada partido, los jugadores pueden ser cambiados, tal que a cada jugador X se le asocian dos números μ_X que representa la posibilidad de que X sea seleccionado y p_X la probabilidad de que X sea escogido. Ambos valores están en el intervalo real [0,1]. Si T es el conjunto de los jugadores seleccionados, estos valores se representan con la notación μ_X = Posibilidad ($X \in T$) y p_X = Probabilidad ($X \in T$). Así, se puede representar la información bimodal para el conjunto T, con las siguientes distribuciones de posibilidad y probabilidad:

$$\text{Poss}(T) = \mu_A/\text{A} + \mu_B/\text{B} + \ldots + \mu_I/\text{I}$$

$$\text{Prob}(T) = p_A/\text{A} + p_B/\text{B} + \ldots + p_I/\text{I}$$

Donde el símbolo + funciona como un separador. En particular, Poss(T) coincide con la representación de un dominio difuso conjuntivo. La información bimodal del equipo estaría dada entonces por Poss(T) y Prob(T).

III. Retos que Presenta la Era de *Big Data*

El constante uso de las tecnologías ha traído consigo un crecimiento explosivo en la cantidad de datos, los cuales son generados a grandes velocidades y en distintos formatos. A partir de este aumento de información, se da la necesidad de

extraer de ella, patrones y/o conocimientos, de forma rápida y eficiente, para lo cual, los métodos tradicionales han tenido que evolucionar en búsqueda de rendimiento y escalabilidad. El gran contenido de valor que genera este tipo de información está permitiendo a las organizaciones una mejora en la toma de sus decisiones, lo que conlleva a la obtención de ventajas competitivas en los diferentes campos de acción.

En un principio, las tecnologías informáticas apoyaron las funciones operativas de compañías y organismos mediante sistemas transaccionales internos, siempre basados en datos perfectamente normalizados, dotados de un formato sistemático y común. Posteriormente, los datos procedentes de los procesos operativos, generalmente almacenados en bases de datos relacionales, se usaron para sustentar los procesos de toma de decisión, siendo la fuente más importante de los sistemas estratégicos que constituían la estructura de la empresa.

Más allá de estos datos críticos de formato fijo, hay un tesoro escondido en otros tipos de datos menos tradicionales o aparentemente menos susceptibles de ser tratados de un modo automatizado, pero con los que también cuentan las empresas y entidades públicas. Los datos no tradicionales proceden de fuentes tales como portales, canales de acceso y de relación con clientes, redes sociales, correo electrónico, fotografías, dispositivos, sensores, objetos, medidores electrónicos, posicionamiento geográfico, entre otros.

Big Data [15] es un término que se ha acuñado para referirse a la manipulación de gran cantidad de datos. El volumen masivo, variedad y velocidad que ahora toma la información hace imprescindible capturar, almacenar y analizar todo este complejo repositorio.

Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Entre los beneficios que se obtienen cuando se asume un proyecto de gestión de grandes volúmenes de datos, está la toma de decisiones más asertivas. Así, entre otras cosas, se obtiene un mercadeo personalizado acorde con las características de la empresa, como es el caso de *Twitter*, *Facebook*, *Instagram* y *Amazon*, con grandes ventajas competitivas. En el ámbito de seguridad, el uso de *Big Data* permite descifrar vulnerabilidades computacionales existentes y protegerse a tiempo de cualquier eventualidad.

Big Data también está emparentado con el área de Minería de Datos, o en forma más general, Extracción de Conocimiento, donde se intenta descubrir patrones de comportamiento en grandes volúmenes de datos.

Dentro de los tipos de datos masivos se pueden mencionar los siguientes [15]:

1) Empresariales tradicionales que incluyen información de clientes proveniente de sistemas como Relaciones con el Cliente (CRM), Sistemas Empresariales (ERP), transacciones realizadas en la Web, inventarios de ventas, entre otros.

2) Detallados que son datos o hallazgos derivados de actividades como campañas de atención telefónica, históricos (*logs*) de equipos, medidores inteligentes, información de planta o producción, sistemas de ventas y comercio.

3) Sociales, que son datos obtenidos a través del comportamiento e interacción de los usuarios con las redes sociales, los cuales incluyen retroalimentaciones (*feedbacks*), opiniones y tendencias.

Esta contribución a la acumulación masiva de datos se puede encontrar en diversas industrias. Las organizaciones mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores y operaciones. En muchos países [16], se administran enormes bases de datos que contienen datos de censo de población, registros médicos, impuestos, ubicación geográfica mediante coordenadas *GPS*, además de transacciones financieras realizadas en línea o por dispositivos móviles.

Por otro lado, están los análisis de redes sociales. Sólo en Twitter son cerca de 12 terabytes de *tweets* creados diariamente. *Facebook* almacena alrededor de 100 *petabytes* de fotos y videos [16]. Si se cuentan todas aquellas actividades, que la mayoría de las personas realizan varias veces al día con los teléfonos inteligentes, se está hablando de alrededor de 2.5 quintillones de *bytes* generados diariamente en el mundo. Por tal razón, en la actualidad, los sistemas computacionales deben trabajar en el orden de los *terabytes* [17].

En la Figura 1 se observa cómo ha sido la evolución de los datos en cuanto a su tamaño, de acuerdo a las operaciones realizadas y los sistemas utilizados. Además, se evidencia el paso con respecto a la tecnología que se empleará, desde un simple dato transaccional, hasta un significativo aumento de los datos generados por los sitios web, los mensajes de texto, el contenido multimedia y las aplicaciones que se encuentran en las populares redes sociales.

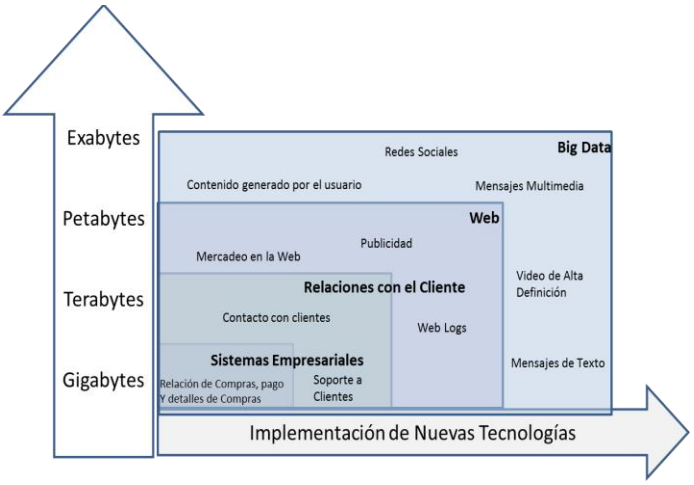

Figura 1: Evolución del Big Data [15]

Big Data incluye una nueva generación de tecnologías y arquitecturas, diseñadas para extraer valor económico de grandes volúmenes de datos heterogéneos, habilitando la captura, identificación y/o análisis a alta velocidad. *Big Data* se caracteriza por tener cinco dimensiones [18]: volumen, variedad, velocidad, veracidad y valor. Estas, son realmente las que definen la frontera entre si se debe o no utilizar *Big Data*. También, dan parámetros a considerar cuándo se piensa contar con el uso tradicional de las bases de datos relacionales. Además, ayudan a decidir si se debe utilizar herramientas existentes para el manejo de datos que no pueden ser tratados

bajo el esquema de tablas normalizadas, facilitando la elección de la herramienta o la aplicación que corresponda [17].

El Volumen es la dimensión que genera la necesidad de procesamiento intensivo y complejo de datos masivos que contienen información de valor para una organización. Se habla de al menos más de un millón de registros [19].

La Variedad indica que la información de valor es el resultado de la combinación de datos de múltiple origen y tipología, presente en forma estructurada, semiestructurada o no estructurada [17].

La Velocidad está asociada a los requerimientos de los procesos y los usuarios [18]. Hoy en día, los datos se generan de forma continua a una velocidad que los sistemas tradicionales no pueden captar, almacenar y/o analizar. Se espera que los sistemas tarden 5 segundos en dar respuesta al procesamiento de los datos.

El Valor hace referencia a los beneficios que se obtienen por el uso de *Big Data*, como, reducción de costes, eficiencia operativa, mejoras del negocio [20].

La Veracidad se refiere al nivel de fiabilidad asociado a ciertos tipos de datos. Conseguir datos de alta calidad es un requisito importante y un reto fundamental de *Big Data*. Sin embargo, aún los mejores métodos de limpieza de datos no pueden eliminar lo imprevisible de algunos datos, como el tiempo, la economía o las futuras decisiones [20].

Fundamentalmente, los sistemas gestores de grandes volúmenes de datos NoSQL están diseñados para aprovechar las nuevas arquitecturas de computación en la nube que han surgido en la última década para permitir cálculos masivos que se ejecuten con bajo costo y de manera eficiente. Esto hace que la gestión de grandes cargas de trabajo sea más fácil, más económica y más rápida de implementar.

La Figura 2 muestra las etapas de la gestión de *Big Data*. La captura es la recolección de datos, procedentes principalmente de la Web, las redes sociales, la interconexión de objetos Máquina a Máquina (M2M), sensores presentes en el Internet de las Cosas (IoT), biometría o directamente de las personas.

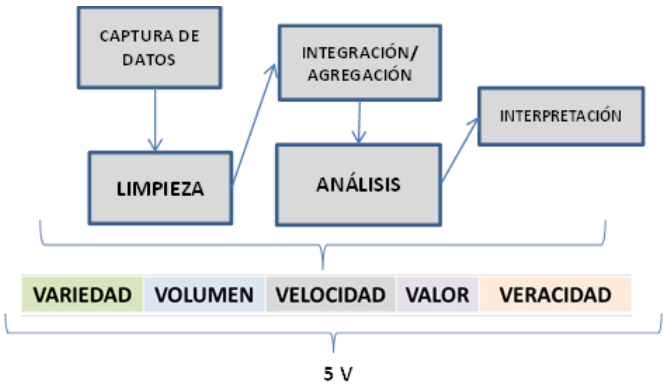

Figura 2: Etapas del Big Data [22]

Boyd y Crawford [21] consideran que la novedad de *Big Data* se encuentra en las capacidades de búsqueda y agregación de grandes cantidades de conjuntos de datos relacionados, más que en el volumen.

La etapa de filtrado y limpieza de los datos puede demandar hasta el 80% del trabajo de análisis [23]. Actualmente, esta etapa se considera una fase previa y separada de los procesos ETL (extraer, transformar y cargar datos). Sin embargo, esto no significa que sea de menor importancia, pues ella asegura la calidad de los datos que se van a procesar, evita información errónea y ayuda en la toma de decisiones correctas.

IV. Antecedentes

En [24], se muestran los avances de la epidemiología gracias a la tecnología digital y al procesamiento de datos masivos (*Big Data*), obtenidos a tiempo real de las poblaciones. Se buscan correlaciones que de forma rápida permitan establecer estadísticas precisas para inferir realidades o causas de fenómenos epidemiológicos, con al menos cierta probabilidad de resultados confiables. Este tipo de datos son muy útiles en investigaciones médicas, para esclarecer la evolución de una posible pandemia, así como, la toma de decisiones relacionadas a los resultados de tales investigaciones. En esas investigaciones no se han considerado uso de datos difusos para modelar las preferencias de los usuarios en la definición de términos médicos, ni tampoco con semántica definida de acuerdo al contexto del usuario.

En el Hospital Birgmingham, de Reino Unido [25], los médicos usaron un sistema conocido como Telemetría para monitorear la recuperación de un niño de un paro cardíaco. Este sistema, normalmente es empleado para evaluar el desempeño del monoplaza en las pistas de carrera de la Fórmula 1. Es la primera vez que se adapta este tipo de sistemas en los seres humanos. Los Laboratorios de la *MCLaren Electronics*, tomaron esta tecnología que recaba una enorme cantidad de información a tiempo real, ya que en cada auto se evalúan cerca de 130 parámetros aproximadamente, que además es capaz de procesarla en corto tiempo.

Estos antecedentes muestran la utilidad del uso de la tecnología de *Big Data* en un área como la médica donde normalmente se presentan datos imperfectos. En [2], se presenta un trabajo de investigación que muestra la gestión de datos imperfectos haciendo uso de tecnologías en el entorno de *Big Data*. El trabajo muestra un caso de estudio en EEUU sobre la atención médica de lesiones de cadera y de antebrazo. Además, exhiben una comparación con el método tradicional de base de datos relacionales.

Desde hace varios años, muchos investigadores han utilizado la Lógica Difusa para modelar e implementar la imperfección de los datos en los sistemas gestores de bases de datos como puede verse en la literatura [26]. La mayoría de estas propuestas están desarrolladas para ser implementadas en Sistemas Gestores de Bases de Datos Relacionales (SGBDR) [27]. Por otra parte, se han abordado diversas aproximaciones teóricas de modelos de datos conceptuales difusos [28].

Si bien la utilización de la Lógica Difusa ha tenido mucho éxito en Ingeniería de Control de la industria manufacturera, existen también un gran número de desarrollos en el área de la Informática, ya que es uno de los pilares para el desarrollo de la Computación Flexible (*Soft Computing*) y Sistemas Inteligentes.

La influencia de la Lógica Difusa sobre sistemas informáticos y en bases de datos, se apoya en que representa un área muy

activa de investigación ya que permite una adaptación al mundo real en que vivimos [28][29]. Así, se han desarrollado trabajos en áreas tales como: Bases de datos Difusas Temporales [30], gestión de datos semánticos [31], recuperación flexible de imágenes médicas [32], citas médicas [33], recursos humanos [34], evaluación de profesores [35].

Un antecedente importante a este artículo lo presentan Cadenas, Marín y Vila [36] quienes proponen una arquitectura para el desarrollo de Sistemas de Bases de Datos Difusas Sensibles al Contexto que sirve de marco para el desarrollo de sistemas inteligentes, flexibles y personalizados al usuario. Una adaptación de esta arquitectura será parte de la propuesta presentada en la próxima sección.

V. Propuesta de Arquitectura para la Gestión de Datos Imperfectos usando Big Data

En este artículo se propone integrar las etapas de desarrollo que considera datos masivos y el trabajo realizado por Cadenas, Marín y Vila [36], que gestiona datos imperfectos sensibles al contexto utilizando una base de datos difusa. En la Figura 3, se observa la arquitectura propuesta que se compone de la gestión de datos imperfectos sensibles al contexto y la tecnología utilizada en *Big Data*. Este modelo mixto plantea el desarrollo de un Módulo de Interacción (MI). En el modelo se muestran las tres fases de administración de *Big Data* que se integran a la arquitectura adaptada de [36], la cual aparece en el lado izquierdo de la Figura 3. Del lado derecho se encuentran esas fases correspondientes al uso de herramientas y tecnologías para datos masivos [37].

La primera parte de la arquitectura corresponde a un Sistema de Bases de Datos Difusas Sensible al Contexto. Como ejemplo, de este tipo de sistema, en [14] se presenta una aplicación para el Examen Físico Articular (EFA) realizado por el Laboratorio de Marcha del Hospital Ortopédico Infantil. Esta aplicación contiene definiciones de etiquetas lingüísticas para diferentes atributos del EFA que eran factibles de ser tratados como difusos. Entre ellos están el peso, la talla, los tonos musculares, el tipo de marcha y los dispositivos utilizados por un paciente. Tales definiciones se adecuan a las preferencias de cada usuario (médicos o fisioterapeutas).

Los usuarios a través de los módulos de sensibilidad al contexto interactúan con el sistema para aportar instrucciones e información de contexto de manera explícita, como la que recoge su perfil. La interacción puede ser efectuada a través de instrucciones del lenguaje estándar de un Sistema Gestor de Base de Datos Objeto Relacional, o a través de aplicaciones sensibles al contexto diseñadas para usuarios inexpertos, tal como el ejemplo mencionado para el EFA.

Los módulos TCP de Entrada/Salida, se encargan del almacenamiento (ASC) y recuperación (RSC) de los datos sensibles al contexto en la base de datos difusa. Estos módulos se comunican con el Gestor de Apoyo de Sensibilización al Contexto para que sean adaptadas las entradas (o salidas) de datos de acuerdo al contexto, antes de ser almacenadas (o recuperadas) en (o de) la base de datos.

El Gestor de Apoyo de Sensibilización al Contexto (GASC) es el encargado de gestionar los datos contextuales en el Catalogo Contextual y las reglas que permitan inferir comportamiento de acuerdo al contexto. Este sistema se convierte en la meta-base de conocimiento que permite a los Módulos TCP hacer su trabajo de transformación.

En la base de datos se almacenan todos los datos que dan soporte a la información de los dominios contextuales, así como, sus metadatos, además de los datos difusos.

La segunda parte, iniciada en el Módulo de Interacción (MI), corresponde a una capa de integración entre las otras dos partes de la arquitectura. Este módulo se encarga de interpretar los datos sensibles al contexto, que se obtienen a través de la "interacción" con los módulos TCP de entrada y salida, para que puedan ser entendidos por los módulos de Análisis y

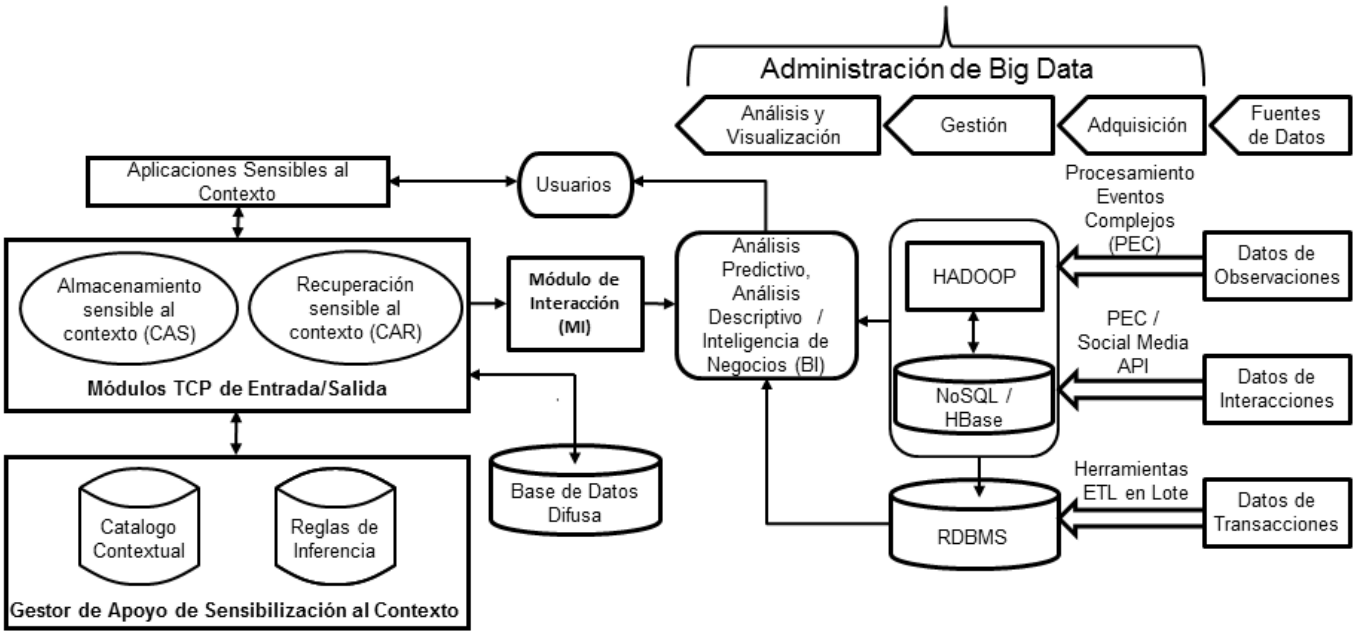

Figura 3: Arquitectura para la Gestión de Datos Imperfectos con *Big Data*

Visualización de la administración *Big Data*. Es importante destacar que los gestores de bases de datos relacionales (SGBDR) no tienen la capacidad de representar y manipular datos imperfectos. Por tal motivo, es necesaria la presencia de una capa que sirva de intérprete, función que en esta arquitectura la haría el Módulo de Interacción.

La tercera parte de la arquitectura corresponde a la administración *Big Data*, que inicia con la fase de adquisición de datos provenientes de diversas fuentes, como son: los datos de observaciones realizadas en procesamientos de eventos complejos (PEC), datos de interacciones obtenidos principalmente de las redes sociales y datos transaccionales. Luego, aparece la fase de gestión, donde no sólo se considera el tratamiento de datos a través de operaciones SQL realizadas en un SGBDR, sino también de datos no estructurados (No sólo SQL ó NoSQL).

Las Bases de Datos NoSQL, como tendencia [38], han venido ganando espacio especialmente por la escalabilidad y velocidad en sus tiempos de respuestas, superiores a los de los sistemas relacionales. En el caso de PostgreSQL, siendo un SGBD Objeto Relacional, ha adquirido varias características de tipo NoSQL, como el almacenamiento momentáneo o fugaz y el manejo de datos en notación de objetos JavaScript (JSON). En [39], dichas características fueron evaluadas con respecto al gestor NoSQL MongoDB, dando como resultado, mejores tiempos de respuestas en todas las operaciones realizadas. Entre las plataformas existentes de tratamiento masivo de datos se pueden mencionar, HADOOP, CouchDB, Neo4J, MapReduce, MongoDB, entre otras.

Finalmente, la administración *Big Data* culmina con la fase de Análisis y Visualización. Aquí se realiza el análisis con técnicas de Inteligencia de Negocio (*Business Intelligence*), predictivas o descriptivas, a fin de producir los resultados o reportes (visualización) que servirán como estrategia para la toma de decisiones de la organización, los cuales están disponibles al usuario de la arquitectura. Es en esta fase donde el módulo de interacción le proporciona los datos difusos y contextuales transformados para que sean considerados durante el análisis. De tal forma que el usuario dispondrá de resultados en base a sus preferencias y contexto en que se ubica.

En resumen, se observan tres partes en esta arquitectura. La primera parte, que trata de la gestión de datos imperfectos sensibles al contexto. La segunda parte, corresponde al módulo de interacción, que sirve de mediador entre esta gestión y la tecnología *Big Data* a utilizar. Finalmente, la tercera parte, se refiere a las etapas del desarrollo *Big Data*.

VI. Conclusiones y Trabajos Futuros

El manejo de grandes cantidades de datos ha ayudado a los investigadores a hacer descubrimientos que les podrían haber tomado años en descifrar por si mismos sin el uso de herramientas para procesar y analizar dichos datos. Debido a la abundancia de sensores, micrófonos, cámaras, escáneres médicos, imágenes, entre otros; presentes en la vida cotidiana, los datos generados a partir de estos elementos serán dentro de poco el segmento más grande de toda la información disponible.

Desde hace varios años, los investigadores han utilizado la teoría de los conjuntos difusos y la lógica difusa, permitiendo modelar la incertidumbre en las Bases de Datos, así como representar las diferentes preferencias de los usuarios. A través de los conjuntos difusos, se modelan los datos posibilísticos, conceptos comúnmente utilizados en el lenguaje natural. Por otro lado, se ha usado esta teoría para representar la información contextual obteniendo aplicaciones inteligentes que gestionan datos imperfectos sensibles al contexto.

Los resultados obtenidos en estas investigaciones previas se han considerado en la propuesta presentada en este trabajo a fin de obtener una arquitectura de un sistema de Base de Datos que gestiona la imperfección, así como, la información del contexto del usuario, con tecnología para la gestión de datos masivos. Se plantea el uso de un Módulo de Interacción (MI), el cual se encarga de la integración entre el módulo de Bases de Datos Difusas Sensibles al Contexto y la fase de análisis y visualización de la administración *Big Data*.

De esta forma, en esta propuesta se incluyen avances ya logrados en cuanto a la gestión de datos imperfectos mediante su almacenamiento y recuperación en Sistemas Gestores de Bases de Datos flexibles. La gestión de datos imperfectos, que incluye ingreso, actualización, eliminación, consulta y extracción de conocimiento, proporciona una mejor aproximación de la información manejada en el mundo virtual con respecto al universo real de los usuarios. Por otro lado, se abarca también las funcionalidades referentes a la formulación de consultas flexibles, por parte de los usuarios, que utilicen términos imperfectos a pesar de que los datos sean precisos.

En este trabajo, se pudieron identificar las características más importantes de la gestión de datos masivos, que incluyen los diferentes formatos, existentes hoy en día, manejados por los usuarios; así como las tecnologías necesarias para convertir datos no estructurados en información y conocimiento que beneficie tanto a personas como a empresas en la toma de decisiones.

Parte de la investigación arrojó que existe una gran variedad de herramientas tecnológicas que permiten el análisis de datos. En su mayoría, dichas herramientas están basadas en Apache HADOOP, disponibles online para ser trabajadas en ambiente web, otras para ser instaladas en computadores personales y algunas para ser usadas en la nube.

Como trabajos futuros, se plantea el desarrollo del Módulo de Interacción propuesto, según los requerimientos descritos. Además, estudiar la utilización de PostgreSQL como Sistema Gestor de Bases de Datos, pues posee las bondades de ser código abierto, incluye características Objeto Relacional e incorpora propiedades del paradigma NoSQL.

También, se plantea la aplicación de la arquitectura propuesta en un caso de estudio. Finalmente, sería deseable lograr incorporar otros tipos de datos imperfectos a esta arquitectura, además de los posibilísticos, tales como los probabilísticos y bimodales tal como lo plantea Zadeh en el Principio de la Información [13].

Agradecimientos

Los autores de esta investigación expresamos agradecimiento por el apoyo incondicional al Prof. Leonid Tineo, Prof. Orlando Luna, Prof. Josué Ramírez y a nuestros familiares por su consideración y aliento. “Te damos gracias, oh Dios, te

damos gracias, pues cercano está tu nombre; los hombres declaran tus maravillas" (Salmos 75:1).

REFERENCIAS

[1] J. Torres, *Big Data: Por qué las Bases de Datos no Sirven para Todos*, TecnoNews, http://www.tecnonews.info/opiniones, Mayo 2013.

[2] M. Hilbert, *Big Data for Development: from Information to Knowledge Societies*, Social Science Research Network, http://www.martinhilbert.net/wp-content/uploads/2015/01/BigData4Dev _Hilbert2014.pdf, 2013.

[3] G. De Tré and S. Zadrozny, *The Application of Fuzzy Logic and Soft Computing in Information Management*, Fuzzy Sets and Systems, vol. 160, no. 15, pp. 2117-2119. 2009.

[4] L. Yan and Z. M. Ma, *Modeling Fuzzy Information in Fuzzy Extended Entity-Relationship Model and Fuzzy Relational Databases*, Journal of Intelligent & Fuzzy Systems, vol. 27, no. 4, pp. 1881–1896, 2014.

[5] M. A. Vila, J. C. Cubero, J. M. Medina, and O. Pons, *A Conceptual Approach for Dealing with Imprecision and Uncertainty in Object based Data Models*, Int. J. Intell. Syst, vol. 11, no. 10, pp. 791-806, 1996.

[6] J. Ramirez and L. Tineo, *Un Mecanismo de Respuesta a Consultas en Presencia de Nulos*, Revista Venezolana de Computación (ReVeCom), vol. 2, no. 1, pp. 48-59, 2015.

[7] L. Zadeh. *Soft Computing and Fuzzy Logic*. IEEE Software, vol. 11, no. 6, pp. 48-56, 1994.

[8] A. Motro, *Accommodating Imprecision in Database Systems: Issues and Solutions*, ACM Sigmod Record, vol. 19, no. 4, pp. 69-74, 1990.

[9] S. Sarkar and D. Dey, *Relational Models and Algebra for Uncertain Data*, In Managing and mining Uncertain Data, C. Aggarwal (ed.), vol. 35, series Advances in Database Systems, Springer, USA, 2009.

[10] C. Aggarwal, *Managing and Mining Uncertain Data*, vol. 35 of the series Advances in Database Systems, Springer, USA, 2009.

[11] L. Zadeh, *Fuzzy Sets*, Information and Control, vol. 8, no. 3, pp. 338-353, 1965.

[12] L. Zadeh, *Fuzzy Sets as a Basis for a Theory of Possibility*, Fuzzy Sets and Systems, vol. 1, pp. 3-28, 1978.

[13] L. Zadeh, *The Information Principle*, Information Sciences, vol. 294, pp. 540–549, 2015.

[14] J. T. Cadenas, *Sistemas de Bases de Datos Difusas Sensibles al Contexto*, Disertación Doctoral, Universidad de Granada, España, 2015.

[15] M. L. Tascon, *Introducción al Big Data. Pasado, Presente y Futuro*, Revista Telos, no. 95, https://telos.fundaciontelefonica.com/seccion=1268&idioma=es_ES&id=2013062110090002&activo=6.do, 2013.

[16] A. Rattinger, *Mejora tu Marketing con Big Data*, Revista Mercado 2.0, vol. 6, no. 3, http://www.merca20.com/big-data, 2014.

[17] J. Curto, *Big Data: Un Mercado Emergente*, Madrid, España, 2012.

[18] IBM Big Data and Analytics Platform, http://www-01.ibm.com /software/data/bigdata, 2012.

[19] L. Joyanes, J. F. Camargo, and J. J. Camargo, *Conociendo Big Data*, Revista Facultad de Ingeniería, vol. 24, no. 38, pp. 63-77, 2014.

[20] M. Chroeck, R. Shockley, J. Smart, D. Romero-Moreno, and P. Tufano, *Analytics: El Uso de Big Data en el Mundo Real*, IBM Global Business Services, IBM Institute for Business Value, 2012.

[21] D. Boyd and K. Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, Information, communication & society, vol. 15, no. 5, pp. 662-679, 2012.

[22] F. Malvicino and G. Joguel, *Big Data: Avances Recientes a Nivel Internacional y Perspectivas para el Desarrollo Local*, CIETI, Universidad Nacional General Sarmiento, Buenos Aires, Argentina, Agosto 2015.

[23] E. Dumbell, *Plannig for Big Data*, A CIO's Handbook to the Changing Data Landscape, 1st Edition, O'Really Radar Team, USA, March 2012.

[24] D. Casacuberta, *Innovación, Big Data y Epidemiología*, Revista Iberoamericana de Argumentación, no.7, pp. 1-12, 2013.

[25] H. Campos, *La Tecnología de MCLaren Salva Vidas en un Hospital Infantil de Inglaterra*, http://www.caranddriverthef1.com, Madrid, España, 2016.

[26] F. E. Petry, *Fuzzy Set Theory Utility for Database and Information Systems*, On Fuzziness, R. Seising, E. Trillas, C. Moraga and S. Termini (eds.), vol. 299, pp. 547-552, series Studies in Fuzziness and Soft Computing, Springer Berlin Heidelberg, 2013.

[27] J. Galindo, *Handbook of Research on Fuzzy Information Processing in Databases*, 1st edition, IGI Global, May 2008.

[28] Z. M. Ma and Y. Li, *A Literature Overview of Fuzzy Conceptual Data Modeling*, J. Inf. Sci. Eng, vol. 26, no. 2, pp. 427-441, 2010.

[29] L. Borjas, J. Ramírez, R. Rodríguez, and L. Tineo, *Automated System for Tests Preparation and Configuration Using Fuzzy Queries*, Computational Intelligence, K. Madani, A. D. Correia, A. Rosa and J. . Filipe (eds.), pp. 199-212, 1st edition, Springer International Publishing, 2015.

[30] J. M. Medina, J. E. Pons, C. D. Barranco, and O. Pons, *A Fuzzy Temporal Object‐Relational Database: Model and Implementation*, International Journal of Intelligent Systems, vol. 29, no. 9, pp. 836-863, 2014.

[31] J. R. Campaña, J. M. Medina, and M. A. Vila, *Semantic Data Management Using Fuzzy Relational Databases*, Flexible Approaches in Data, Information and Knowledge Management, O. Pivert and S. Zadrozny (eds.), vol. 497, pp. 115-140, Serie Studies in Computational Intelligence, Springer International Publishing, September 2013.

[32] J. M. Medina, S. Jaime-Castillo, C. D. Barranco, and J. R. Campana, *On the Use of a Fuzzy Object-Relational Database for Flexible Retrieval of Medical Images*, IEEE Transactions on Fuzzy Systems, vol. 20, no. 4, pp. 786-803, 2012.

[33] T. Toshiro, K. Asai, and M. Sugeno, *Applied Fuzzy Systems*, 1st edition, Academic Press, San Diego, USA, 1994.

[34] L. Lien-Fu, W. Chao-Chin, H. Yi-Ta, and H. Liang-Tsung, *A Fuzzy Query Approach to Human Resource Web Services*, in proceeding of the 10th International Conference on e-Business Engineering (ICEBE), pp. 461-466, Coventry, United Kindom, September 2013.

[35] A. Aguilera, L. Borjas, R. Rodríguez, and L. Tineo, *Experiences on Fuzzy DBMS: Implementation and Use*, in proceeding of the XXXIX Latin American Computing Conference (CLEI 2013), pp. 1-8, Naiguatá, Venezuela, October 2013.

[36] J. T. Cadenas, N. Marín, and M. A. Vila, *Context-Aware Fuzzy Databases*, Applied Soft Computing, vol. 25, pp. 215-233, 2014.

[37] H. Cheng, R. Chiang, and V. Storey, *Business Intelligence and Analytics: from Big Data to Big Impact*, MIS Quarterly, vol. 36, no. 4, pp. 1165-1188, December 2012.

[38] N. Leavitt, *NoSQL Databases Live Up to Their Promise?*, Journal Computer, vol. 43, no. 2, pp.12-14, USA, 2010.

[39] A. Sotolongo, L. Vasquez, and Y. Vazquez, *Evaluación de Características NoSQL en PostgreSQL*, http://es.slideshare.net/asotolongo/caractersticas -nosql-de-postgresql, La Habana, Cuba, 2013.