Prototipo de software para la evaluación de la calidad de datos abiertos.

Nicolás Estefan López Beltrán
John Ferney Mahecha Moyano

Universidad Católica de Colombia

Facultad de Ingeniería

Programa de Ingeniería de Sistemas

Modalidad

Bogotá, Colombia

2017

Prototipo de software para la evaluación de la calidad de los datos abiertos.

Nicolás Estefan López Beltrán
John Ferney Mahecha Moyano

Tesis presentada como requisito parcial para optar al título de:

Ingeniero de Sistemas

Director (a):

M.Sc. John Velandia

Universidad Católica de Colombia
Facultad de Ingeniería
Programa de Ingeniería de Sistemas
Modalidad
Bogotá, Colombia
2017

creative commons

Atribución-NoComercial-SinDerivadas 2.5 Colombia (CC BY-NC-ND 2.5)

La presente obra está bajo una licencia:
Atribución-NoComercial-SinDerivadas 2.5 Colombia (CC BY-NC-ND 2.5)
Para leer el texto completo de la licencia, visita:
http://creativecommons.org/licenses/by-nc-nd/2.5/co/

Usted es libre de:

Compartir - copiar, distribuir, ejecutar y comunicar públicamente la obra

Bajo las condiciones siguientes:

Atribución — Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciante (pero no de una manera que sugiera que tiene su apoyo o que apoyan el uso que hace de su obra).

No Comercial — No puede utilizar esta obra para fines comerciales.

Sin Obras Derivadas — No se puede alterar, transformar o generar una obra derivada a partir de esta obra.

Nota de aceptación

Aprobado por el comité de grado en cumplimiento de los requisitos exigidos por la Facultad de Ingeniería y la Universidad Católica de Colombia para optar al título de Ingenieros de Sistemas.

_____
_____
_____

_____

M.Sc. John Alexander Velandia Vega
Director de trabajo de grado

_____

M.Sc. Alexandra López
Coordinador de trabajo de grado

Bogotá, Mayo de 2017

AGRADECIMIENTOS

A aquellas personas que contribuyeron a la realización de este proyecto, a nuestros padres y familia por su apoyo incondicional, a nuestro director de trabajo de grado MSc. John Velandia por guiarnos en la elaboración del proyecto y a Antonio Vetrò por guiarnos en la resolución de dudas con respecto a las métricas trabajadas.

CONTENIDO

Pág.

LISTA DE FIGURAS

Pág.

LISTA DE ANEXOS

Pag.

RESUMEN

No hay un conocimiento de que la ley de datos abiertos se esté cumpliendo éticamente y constitucionalmente, esto genera la necesidad de desarrollar un proyecto que permita realizar la validación. Dentro de los puntos necesarios para llegar a la conclusión del proyecto se necesita una herramienta que permita verificar la calidad de los datos. El proyecto específico se centra en el desarrollo de un prototipo para la medición de la calidad de los datos utilizando tres métricas: completitud, trazabilidad y exactitud. Los datos son extraídos desde el repositorio www.datos.gov.co utilizando Rapidminer y JAVA.

Con el proceso de medición aplicado a los datos se encuentran problemas que tiene la plataforma www.datos.gov.co, por ejemplo, el identificador de cada conjunto de datos no posee una manera de diferenciarlos.

Para el desarrollo del proyecto se utiliza una metodología de investigación que inicia con la definición del problema de investigación. Una vez identificado el problema se realiza la revisión de la literatura en diferentes fuentes. Luego de obtener información se plantea la hipótesis a desarrollar a lo largo del proyecto, preparando el diseño de la investigación y coleccionando la información realmente útil para el desarrollo. Finalmente se realiza un análisis de los datos, categorizando las métricas y definiciones generales del proyecto.

Para la codificación del prototipo se utiliza la metodología de desarrollo de software espiral. Esta ofrece una etapa de planeación, lo que beneficia al proyecto en el análisis y codificación de las métricas. Se crearon cálculos en cada fase y dentro de los datos se encontraron inconsistencias en la plataforma.

INTRODUCCIÓN

El concepto de datos abiertos está tomando fuerza en Colombia, la información que manejan las organizaciones cada vez aumenta en su tamaño de manera incontrolable, y es normal, cada pequeño detalle que se tome para realizar estudios o tomar decisiones influye en el futuro de la compañía[1].

Las compañías públicas y privadas están empezando a acceder a los datos suministrados en la página web www.datos.gov.co, tanto para consumir los datos que existen ahí como para subir la información que generan con ellos, lo que se desconoce es si la información que ingresa al sistema cumple con lo que dice la ley de datos abiertos o lo hacen para "cumplir" con las normas. El objetivo de la ley es regular el derecho de acceso a la información pública[2].

El determinar si la organización está tomando un buen camino con los datos que se están utilizando es importante, porque la buena calidad de los mismos ofrecen competitividad estratégica[3], si se utiliza una información de mala calidad las decisiones serán en un contexto diferente al real[4]. La investigación verificará si los datos que utilizan las compañías son de calidad, y con ello, saber si a las compañías es útil el repositorio para sus tareas.

El origen de datos abiertos en Colombia surge de la implementación de la ley 1712 de 2014, las compañías públicas obtienen un repositorio para aquellos interesados que quieran utilizar esta información para diferentes entornos. Además de subir información a un repositorio, el gobierno nacional con el ministerio de las TIC abrió la página www.datos.gov.co, un sitio web donde las organizaciones pueden subir datos y hacer cumplimiento de la ley.

El objetivo de la investigación consiste en evaluar los datos alojados en esta página, ¿Cómo se pueden evaluar? Existen métricas para medir la calidad, como lo son[5]: Trazabilidad[6], completitud[7] y conformidad[8]. Son un punto de partida para determinar

---

[1] LOSHIN, David. The practitioner's guide to data quality improvement. 1 ed. Burlington: Morgan Kaufmann, 2010. 432p.
[2] COLOMBIA. CONGRESO DE LA REPÚBLICA. Ley 1712. (06, marzo, 2014). Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones. MINTIC. Bogotá, D.C., 2014. p. 1-14.
[3] AZUMAH, Kenneth y QUARSHIE, Henry. Towards Higher Quality Data: Impact of Perception of Data Quality on IT Investment - Ghana. En: International Journal of Emerging Trends in Computing and Information Sciences. Enero, 2013. vol. 3, no. 12, p. 1614-1621.
[4] LOSHIN. Op. cit.
[5] HERZOG, Thomas; SCHEUREN, Fritz y WINKLER, William. Data Quality and Record Linkage Techniques. 1 ed. New York: Springer-Verlag New York, 2007. 234 p.
[6] VETRÒ, Antonio, et al. Open Data Quality Measurement Framework: Definition and Application to Open Government Data. En: ScienceDirect. Abril, 2016. vol. 33, no. 2, p. 325-337.
[7] PIPINO, Leo; LEE, Yang y WANG, Richard. Data Quality Assessment. En: Communications of the ACM. Abril, 2002. vol. 45, no. 4, p. 211-218.
[8] VANÍČEK, Jiří. Software and data quality. En: Czech Academy of Agricultural Sciences. Febrero, 2006. vol. 52, no. 3, p. 138-146.

si una persona al momento de acceder a cualquier tipo de archivo alojado lo puede obtener de manera oportuna y precisa.

Esto es un punto de partida para futuras investigaciones, la investigación se centra únicamente en el análisis de la página www.datos.gov.co. Entrando a un enfoque más técnico, la codificación de las métricas planteadas se realizará mediante la metodología espiral. La mayor ventaja de utilizarla es que permite desarrollar de manera iterativa el proyecto[9], lo que facilita la comprensión, la codificación y las entregas de cada módulo correspondiente a las mediciones, además permite una reevaluación de las fases.
La facultad de derecho conjunto con la facultad de ingeniería de sistemas plantea un proyecto que busca identificar si éticamente y constitucionalmente se está cumpliendo la ley de datos abiertos. Para la conclusión del proyecto se necesitan 2 herramientas, una enfocada a la madurez de la implementación de la política de los datos abiertos con sus principios, y la otra enfocada a la calidad de los datos generados por las entidades públicas del gobierno. Otra razón del desarrollo del prototipo es que la calidad de datos es importante para las empresas y el gobierno porque da reputación, satisfacción del cliente y/o asegurar datos financieros si aplica[10].

[9] BOEHM, Barry. Spiral Development: Experience, Principles, and Refinements Spiral Development Workshop February 9, 2000. Los Angeles, Software Engineering Institute, 2000. 37 p. CMU/SEI-2000-SR-008.
[10] HERZOG, Thomas; SCHEUREN, Fritz y WINKLER, William. Data Quality and Record Linkage Techniques. 1 ed. New York: Springer-Verlag New York, 2007. 234 p.

1 PLANTEAMIENTO DEL PROBLEMA

1.1 DESCRIPCIÓN DEL PROBLEMA

Se desconoce si en Colombia se está cumpliendo la ley de datos abiertos de manera ética y constitucional. Para comprobar esto es necesaria una herramienta a la medida que permita extraer, calcular la calidad, y evaluar la madurez de los datos del repositorio www.datos.gov.co. La facultad de ingeniería se va a encargar de dar solución al desarrollo de las herramientas necesarias para la continuidad del proyecto. El enfoque de este proyecto es la extracción y la evaluación de calidad de los datos.

1.2 FORMULACIÓN DEL PROBLEMA

¿Los datos alojados en www.datos.gov.co que consultan y manejan las diferentes organizaciones que trabajan en Colombia para el desarrollo de proyectos y servicios cumplen las métricas de trazabilidad, completitud y conformidad?

2 OBJETIVOS

2.1 OBJETIVO GENERAL

Implementar un prototipo que permita calcular las métricas de trazabilidad, completitud y conformidad de los datos obtenidos del repositorio www.datos.gov.co.

2.2 OBJETIVOS ESPECÍFICOS

- Identificar las métricas que permitan evaluar la calidad de los datos.
- Diseñar la arquitectura de software y la estructura del prototipo.
- Implementar el prototipo de software de acuerdo a la arquitectura propuesta.
- Realizar las pruebas al prototipo para verificar su funcionalidad.

3 ALCANCE Y LIMITACIONES

El proyecto toma como referencia el repositorio www.datos.gov.co, el cual contiene una fuente de datos respecto a información pública de diferentes entidades públicas como lo exige la ley 1712 de 2014. Teniendo en cuenta el tiempo disponible para el desarrollo del proyecto, y la necesidad que tiene la facultad de derecho en la utilización del prototipo con el objetivo de verificar si los datos abiertos están cumpliendo la ley éticamente y constitucionalmente, se toman la información alojada de los conjuntos en formato JSON para la extracción de datos y el cálculo de las métricas de trazabilidad, completitud y conformidad. Al final del proyecto el prototipo será ubicado en un servidor web para el libre acceso de la herramienta.

La API del repositorio www.datos.gov.co no permite descargar más de un conjunto de datos simultáneamente, por lo que la actividad de extraer múltiples se realiza de manera manual mediante programación. El prototipo acepta cualquier modelo de datos que esté guardado en el sistema mediante el módulo de ingresar estructura.

Para la lectura y análisis de los datos se utiliza la herramienta RapidMiner, ya que posee algoritmos avanzados que facilita el tratamiento de la información, además de poseer una interfaz amigable con el usuario y la facilidad de crear procesos y flujos de actividades con solo arrastrar operadores[11].

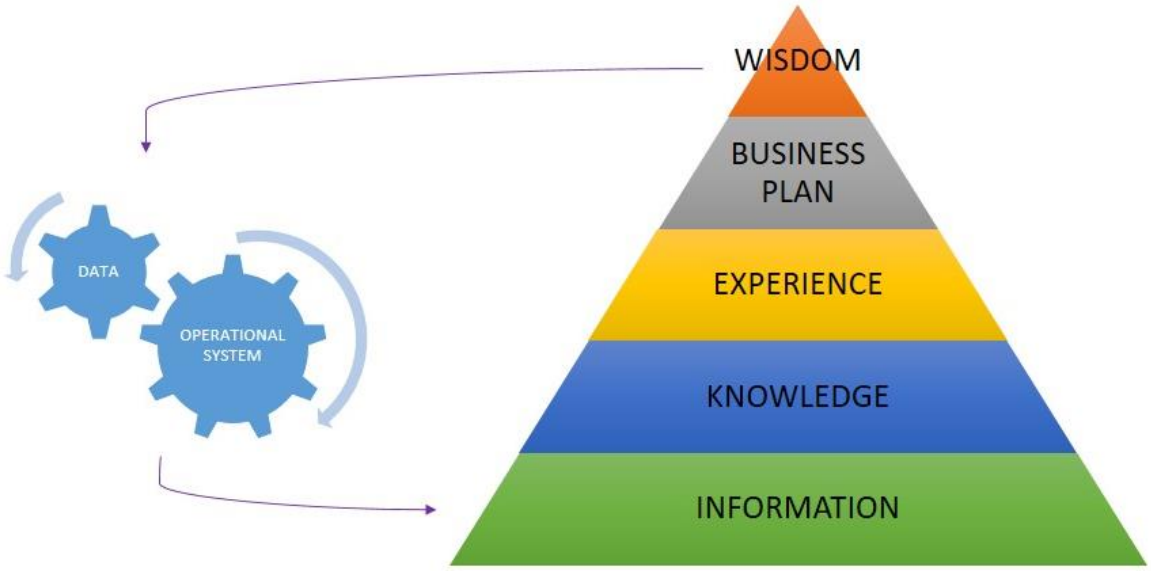
[11] HOFMANN, Markus y KLINKENBERG, Ralf. RapidMiner: Data Mining Use Cases and Business Analytics Applications. ed. 1. Florida: CRC Press, 2013. 525 p.

4 MARCO REFERENCIAL

4.1 MARCO TEÓRICO

Los datos se convierten en activos de la compañía por la importancia que tienen en los planes estratégicos y en las acciones que determinan el éxito. En consecuencia, los datos de baja calidad tienen un impacto grave para la empresa, si no se identifican y se corrigen desde el inicio pueden contaminar los sistemas y con ello aumentar los costos, desprestigiar la empresa y causar análisis imprecisos y malas decisiones[12].

Figura 1. Flujo de los datos de una organización.



Fuente: ECKERSON, Wayne. Data quality and the bottom line. Chatsworth, 101 Communications LLC, 2002. 33 p.

El tratamiento de los datos que se obtienen durante cierto tiempo está regido a las leyes del país donde se quieren manejar, en Colombia, se maneja la Ley de transparencia de datos y del derecho al acceso de la información pública nacional. Esta ley exige a las entidades responsables, mantener datos de calidad como lo dice el parágrafo 1 del artículo 9[13].

Para el cumplimiento de la ley las organizaciones registran su información en el repositorio www.datos.gov.co. En esta página se encuentra toda la información referente al funcionamiento de las organizaciones públicas de Colombia, como lo

[12] ECKERSON, Wayne. Data quality and the bottom line. Chatsworth, 101 Communications LLC, 2002. 33 p.
[13] COLOMBIA. CONGRESO DE LA REPÚBLICA. Ley 1712. (06, marzo, 2014). Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones. MINTIC. Bogotá, D.C., 2014. p. 1-14.

es información de contratos, documentación, formularios, mapas, historias o recursos.

4.1.1 Big Data. Es un término utilizado para referirse al aumento del volumen de datos que son difíciles de almacenar, procesar y analizar con las tecnologías de bases de datos tradicionales[14].

Actualmente no existe una manera clara de definir qué tan grande debe ser una muestra de datos para llamarlo Big Data, sin embargo, para efectos académicos se dice que[15]:

- 10GB de ficheros con datos procesables en Excel o R se considera Small Data.
- Entre 10GB y 1024GB (1TB) requiere una base de datos especializada.
- Más de 1TB se considera Big Data, aunque se habla de PB o EB. Se manejan con bases de datos distribuidas y es necesario un almacenamiento en múltiples ordenadores.

Big Data posee cuatro características principales: volumen, variedad, velocidad y veracidad, se le conoce como el 4V. Son las siguientes[16]:

- Volumen: Se debe gestionar y procesas grandes cantidades de datos.
- Velocidad: Se debe procesar datos a alta velocidad.
- Variedad: Se debe validar múltiples tipos de datos estructurados y no estructurados.
- Veracidad: Se debe validar la corrección de la mayoría de datos.

En la actualidad, la tendencia de Big Data está involucrando todas las áreas en cada empresa, en especial al área de tecnología, están cambiando las herramientas a utilizar y las metodologías de análisis, lo que puede dar oportunidad a mayor oferta laboral[17].

4.1.2 Datos abiertos. Big Data se alimenta de los datos abiertos que estén a su alcance. Para el término datos abiertos existen varias definiciones, una de ellas es que son todos los datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona[18]. Cabe aclarar que open data no se refiere

---

[14] TARGIO, Ibrahim, et al. The rise of "big data" on cloud computing: Review and open research issues. En: Sciencedirect. Julio, 2014. vol. 47, pag. 98-115.
[15] HERNÁNDEZ-PÉREZ, Tony. En la era de la web de los datos: primero datos abiertos, después datos masivos. En: El profesional de la información. Julio, 2016. vol. 25, no. 4, p. 517-525.
[16] SEVILLANO, Felipe. Big Data. En: Dialnet. 2015. no. 395, p. 71-86.
[17] E-skills UK, Big data analytics: adoption and employment trends, 2012-2017. En: VOCED. Noviembre, 2013. 22 p.
[18] Open Knowledge Foundation, Manual de los datos abiertos. 1 ed. Argentina: Open Knowledge Foundation, 2012. 62 p.

únicamente a los datos digitales, sino también a los impresos, ya que existe una enorme cantidad de información impresa en archivos y bibliotecas creada en ausencia de tecnología o situaciones que no aplicaba tal uso[19].

Según el manual de Open Knowledge Foundation[20], las características fundamentales de los datos abiertos son las siguientes:

- Disponibilidad y acceso: Mediante internet preferiblemente, se brinda la oportunidad de disponer de la misma información a todo el mundo, además que preferiblemente se permita su modificación. Esto fortalece la transparencia en diferentes sectores, como es el económico y político.
- Reutilización: Además de una redistribución, combinándose con otros archivos.
- Participación universal: Cualquier persona del mundo puede utilizar la información.
- Tratamiento en tiempo real: Normalmente a través de la nube.
- Formatos abiertos[21].

Un ejemplo que aplica el concepto de Open Data es el siguiente: "La película Amanecer 2 es un ejemplo de cómo interpretar los sentimientos a través del análisis de datos, analizando los tweets generados en la promoción de la película. La campaña de marketing fue a nivel mundial y estaba basada en las redes sociales. Los movimientos en Twitter se analizaron con Social sentiment index de IBM."[22].

Una ventaja que brinda la implementación de Open Data es invitar a la ciudadanía a participar en los proyectos públicos vigentes en la ciudad, como en la opinión en la toma de decisiones, colaboración en diferentes planteamientos y debates y el diseño, prestación y evaluación de los servicios que se brindan.

Otras ventajas son:

- Generación de riqueza para las empresas, ya que poseen información para sus análisis para la creación de nuevos servicios o productos de consumo.
- Socializar con la ciudadanía el método que utilizan respecto a las entidades públicas, buscando una mejor percepción de los servicios que se prestan.
- Desarrollar nuevos servicios que complementen los ya creados[23].

---

[19] FERRER, Antonia y SÁNCHEZ, Enrique. Open data, big data: ¿hacia dónde nos dirigimos?. En: Anuario ThinkEPI. Febrero, 2012. vol. 7, p. 150-156.
[20] Open Knowledge Foundation, Manual de los datos abiertos. 1 ed. Argentina: Open Knowledge Foundation, 2012. 62 p.
[21] GARRIGA, Marc. ¿Datos abiertos? Sí, pero de forma sostenible. En: El profesional de la información. Mayo, 2011. vol. 20, no. 3, p. 298-303.
[22] FERRER. Op. cit., p. 150-156.
[23] GARRIGA. Op. cit., p. 298-303.

Una de las razones más importantes de que los datos sean abiertos es buscar una propiedad, la interoperabilidad, ¿qué es esto? es una habilidad que nos permite la integración de diferentes bases de datos, lo cual permite a diversos sistemas y organizaciones trabajar juntos[24].

4.1.3 Data Quality. No es un secreto que estamos en la era de la información, las organizaciones trabajan con todo lo que esté a su alcance respecto a datos, todo lo que se pueda recopilar es importante, con un análisis posterior a esa información se genera conocimiento para la toma de decisiones, ya sean proyectos, nuevos servicios, productos o correcciones del funcionamiento de la organización[25].

La recolección de los datos se realiza de todas las fuentes que están al alcance, lo que se va a obtener como resultado es información en múltiples formatos. Luego de guardar todos esos datos que se han obtenido hay que evaluar ¿Qué tan confiables son esos datos? Para responder esta pregunta es necesario hablar de la calidad de los datos.

Se dice que los datos son de alta calidad cuando son "aptos" para su uso en su intención operativa y la toma de decisiones. Igualmente se define la calidad como la conformidad de las normas que se han establecido[26].

La calidad de los datos se define por 5 propiedades[27]:

- Relevancia: Consiste en que los datos almacenados puedan ser utilizados en diferentes sectores y no en uno, el hecho que con los datos no sea posible realizar los estudios requeridos representa un gasto de dinero adicional y tiempo el complementar los datos para su uso.
- Exactitud: Los datos nunca van a ser 100% precisos, dependiendo de la situación y el sector en que se maneje la información será de utilidad o no.
- Oportunidad: Los datos deben ser accesibles de manera oportuna.
- Comparabilidad: Consiste en combinar diferentes bases de para facilitar el uso de los datos en los análisis, modelado y estimaciones estadísticas.
- Completitud: Consiste en que no hayan elementos vacíos en los registros o registros sin ningún dato.

[24] Open Knowledge Foundation, Manual de los datos abiertos. 1 ed. Argentina: Open Knowledge Foundation, 2012. 62 p.
[25] LOSHIN, David. The practitioner's guide to data quality improvement. 1 ed. Burlington: Morgan Kaufmann, 2010. 432p.
[26] HERZOG, Thomas; SCHEUREN, Fritz y WINKLER, William. Data Quality and Record Linkage Techniques. 1 ed. New York: Springer-Verlag New York, 2007. 234 p.
[27] HAUG, Anders; ZACHARIASSEN, Frederik y VAN LIEMPD, Dennis. The costs of poor data quality. En: Journal of Industrial Engineering and Management. Enero, 2011. vol. 4, no. 2, p. 168-193.

4.1.4 Arquitectura orientada a servicios. Es una arquitectura contractual que ofrece y consume servicios web. Se conforma de 3 entidades: proveedores de servicios que son los que ofrecen el servicio detalladamente, demandantes de servicios que los localizan según el interés correspondiente, y registro de servicios[28].

Una forma que existe para el uso de la arquitectura orientada a servicios es REST. Se define como una arquitectura que utiliza diferentes estándares como HTTP, XML, URL y HTML para el uso de web services, enfocándose en los recursos del sistema[29].

Consiste en que un cliente realiza una solicitud a un servidor mediante una URI, este la procesa y retorna una respuesta al cliente. La transferencia de los recursos de realiza mediante el estándar HTTP, ya que nos ofrece los métodos GET, PUT, POST y DELETE, son las operaciones que podemos realizar, consisten en:

- GET: Se utiliza para obtener un recurso del servidor.
- PUT: Se utiliza para actualizar un recurso del servidor.
- POST: Se utiliza para crear un recurso en el servidor.
- DELETE: Se utiliza para borrar un recurso del servidor.

4.2 MARCO CONCEPTUAL

Big Data: Volúmenes masivos y complejos de información estructurada y no estructurada que requiere de métodos computacionales para extraer conocimiento[30].

Datos abiertos: Son todos los datos que se pueden utilizar y publicar análisis en base a ellos libremente, siempre y cuando no sean alterados y se publiquen de la misma manera en la que se encontraron[31].

Arquitectura orientada a servicios: Arquitectura contractual que ofrece y consume servicios web. Se conforma de 3 entidades: proveedores de servicios que son los que ofrecen el servicio detalladamente, demandantes de servicios que los localizan según el interés correspondiente, y registro de servicios[32].

---

[28] ERL, Thomas. Service-Oriented Architecture: A Field Guide to Integrating XML and Web Services. 1 ed. New Jersey: Prentice Hall PTR, 2004. 541 p.
[29] CABRERA, Yandy. Transferencia de estado representacional (REST): estilo de arquitectura para sistemas distribuidos de hipermedia. En: Serie científica de la universidad de las ciencias informáticas. Julio, 2013. vol. 6, no. 7.
[30] ARCILA, Carlos; BARBOSA, Eduar y CABEZUELO, Francisco. Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística. En: El profesional de la información. Julio, 2016. vol. 25, no. 4, p. 623-631.
[31] Open Knowledge Foundation, Manual de los datos abiertos. 1 ed. Argentina: Open Knowledge Foundation, 2012. 62 p.
[32] ERL. Op. cit.

Data Quality: Son aquellos datos que cumplen los estándares y normas establecidas y son aptos para el uso libre de cualquier persona u organización[33].

MongoDB: Es un sistema de gestión de bases de datos diseñado para desarrollar aplicaciones web e infraestructura de internet con la posibilidad de escalamiento[34].

JAVA: Es un lenguaje de programación orientado a objetos diseñado para ser pequeño, simple y portátil a través de plataformas y sistemas operativos[35].

[33] HERZOG, Thomas; SCHEUREN, Fritz y WINKLER, William. Data Quality and Record Linkage Techniques. 1 ed. New York: Springer-Verlag New York, 2007. 234 p.
[34] BANKER, Kyle. MongoDB in action. 2 ed. New York: Manning Publications Co., 2012. 312 p.
[35] LEMAY, Laura y CADENHEAD, Rogers. Sams Teach Yourself Java 2 in 21 Days. 3 ed. Indianapolis: SAMS Publishing, 2002. 736 p.

5 METODOLOGÍA

5.1 METODOLOGÍA DE INVESTIGACIÓN

La metodología que se aplica al proyecto proporciona una mejor formación, ofreciendo técnicas para la recolección de datos, herramientas para realizar mejores investigaciones, pensamiento disciplinado y más objetivo hacia el campo a explorar. Se encuentra que la metodología es útil en varios campos de la ciencia.

La metodología para la investigación de la información es el siguiente[36]:

- Definir el problema de investigación: El problema principal de la investigación es que se desconoce si los datos alojados en el repositorio www.datos.gov.co cumple con los estándares de calidad de los mismos.
- Revisión de la literatura: Se recopila información de diferentes fuentes acerca del problema y con la ayuda del director de grado se válida que la información es relevante.
- Desarrollo de hipótesis de trabajo: Una vez que la información encontrada es validada, se realiza una hipótesis de cómo se aborda el problema de la investigación.
- Preparar el diseño de la investigación: Se define las fuentes donde se va a recopilar la información de las métricas a trabajar. Las búsquedas se realizaron en diferentes Journal, por ejemplo: IEEE explore, CiteSeerX, DBPL: Computer Science Bibliography y Google Scholar.
- Coleccionar datos: Una vez definidas las fuentes de información se realizan las búsquedas con base al diseño de investigación planteado.
- Análisis de los datos: La información recopilada se categoriza según las métricas y definiciones generales del proyecto, por ejemplo: Data Quality.
- Interpretación y reporte: En el proyecto no aplica esta fase, ya que el proyecto no tiene una finalización con la investigación. Con base a la información recopilada se va a realizar un prototipo para la evaluación de las métricas.

5.2 METODOLOGÍA DE DESARROLLO DE SOFTWARE

La metodología de software para la codificación del proyecto es la metodología espiral. La metodología espiral posee un enfoque cíclico para incrementar gradualmente el nivel de detalle e implementación del prototipo, al tiempo que disminuye su grado de riesgo[37].

Otras ventajas de la metodología espiral son:

[36] KOTHARI, C. Research Methodology. 2 ed. New Delhi: New Age International (P) Limited, Publishers, 2004. 418 p.
[37] BOEHM, Barry. Spiral Development: Experience, Principles, and Refinements Spiral Development Workshop February 9, 2000. Los Angeles, Software Engineering Institute, 2000. 37 p. CMU/SEI-2000-SR-008.

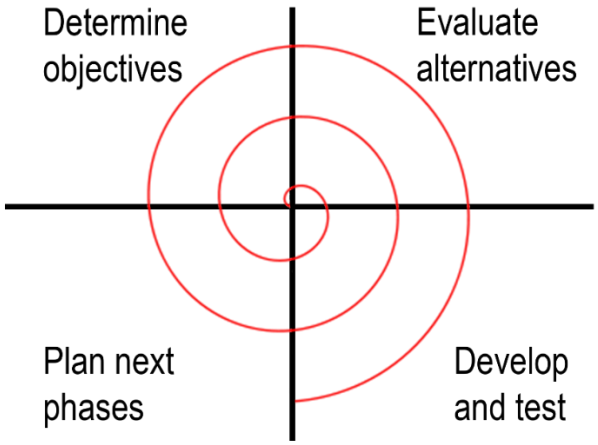
- Permite añadir funcionalidades adicionales en la ejecución o después de finalizar el proyecto[38], lo que aporta escalabilidad a la herramienta.
- Ofrece una etapa de planeación, lo que beneficia al proyecto en el análisis y codificación de las métricas[39].

Una ventaja de la metodología espiral frente a las metodologías secuenciales es que si los requerimientos no están bien definidos y hay posibilidad que haya cambios, estos no van a afectar el desarrollo del proyecto de manera significativa[40].

Una ventaja de la metodología espiral frente a las metodologías ágiles es el enfoque a la calidad de software, mientras la metodología ágil prioriza la entrega del software. Otra ventaja de la metodología en espiral es la adaptación a los proyectos medianos, mientras las metodologías ágiles van enfocadas a soluciones de gran tamaño[41].

El modelo de la metodología espiral es el siguiente[42]:

Figura 2. Modelo espiral.



Fuente: Los autores.

[38] ALSHAMRANI, Adel y BAHATTAB, Abdullah. A Comparison Between Three SDLC Models Waterfall Model, Spiral Model, and Incremental/Iterative Model. En: International Journal of Computer Science Issues. Enero, 2015. vol. 12, no. 1, p. 106-111.
[39] NABIL, Ali y GOVARDHAN, A. A Comparison Between Five Models Of Software Engineering. En: International Journal of Computer Science Issues. Septiembre, 2010. vol. 7, no. 5, p. 94-101.
[40] ISAIAS, Pedro y ISSA, Tomayess. High Level Models and Methodologies for Information Systems. 1 ed. New York: Springer, 2015. 145 p.
[41] JAVANMARD, Mahdi y ALIAN, Maryam. Comparison between Agile and Traditional software development methodologies. En: Science Journal (CSJ). Mayo, 2015. vol. 36, no. 3, p. 1386-1394.
[42] BOEHM, Barry. Spiral Development: Experience, Principles, and Refinements Spiral Development Workshop February 9, 2000. Los Angeles, Software Engineering Institute, 2000. 37 p. CMU/SEI-2000-SR-008.

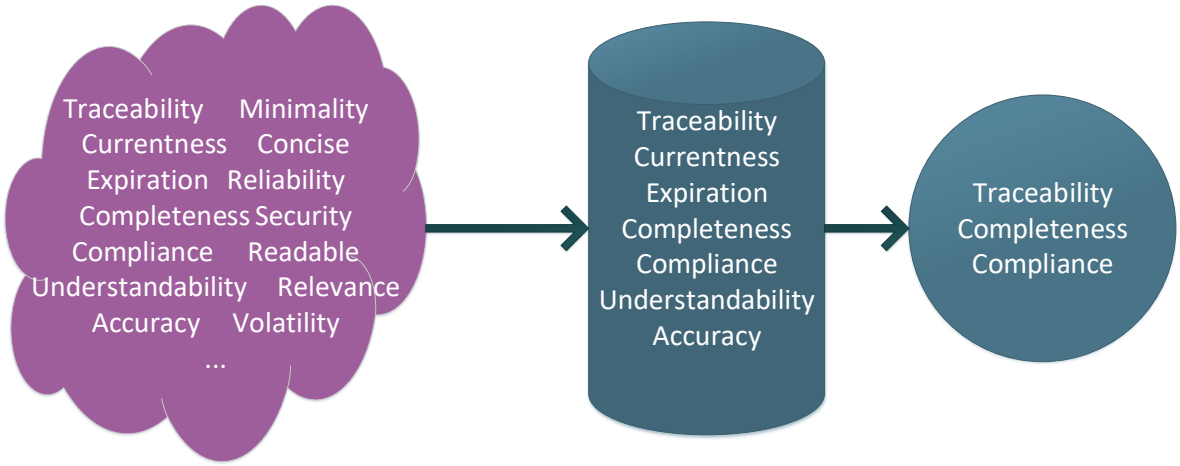
Las fases que comprende son las siguientes:

- Plantear objetivos: En esta fase se determina el módulo a codificar según el desarrollo del sistema.
- Evaluación de las alternativas y riesgos identificados: En esta fase de analizan las alternativas para la codificación y los riesgos presentes en ellas. Al escoger una alternativa se deciden estrategias alternativas para mitigar los riesgos.
- Desarrollo y pruebas: Comprende la codificación y testeo del módulo desarrollado con pruebas unitarias y pruebas de aceptación.
- Planeación siguiente fase: En esta fase se plantea los planes de desarrollo, integración y pruebas.

6 ANÁLISIS DE LAS MÉTRICAS DE CALIDAD

En calidad de datos existe un conjunto de métricas que cada día crece más[43], según un estudio del Politécnico de Torino[44] se tomaron los resultados obtenidos y publicados por medio de un artículo para la aplicación de las métricas que el politécnico trabajó. En el artículo se encuentran 3 métricas accesibles para su evaluación, por lo tanto se procede a analizar una a una con base a la fórmula presentada.

Figura 3. Métricas calidad de datos.



Fuente: VETRÒ, Antonio, et al. Open Data Quality Measurement Framework: Definition and Application to Open Government Data. En: ScienceDirect. Abril, 2016. vol. 33, no. 2, p. 325-337.

6.1 TRAZABILIDAD

La trazabilidad indica la presencia o ausencia de metadatos asociados con el proceso de creación y actualización de un conjunto de datos[45]. La expresión matemática para la evaluación de la trazabilidad es la siguiente:

$$tc=2s+dc \quad (1)$$

$$tu=lu+du \quad (2)$$

tc: Traza de creación
s: Fuente
dc: Fecha de creación
tu: Traza de actualización

[43] BATINI, Carlo, et al. Methodologies for Data Quality Assessment and Improvement. En: ACM Computing Surveys. Julio, 2009. vol. 41, no. 3, p. 16:1-16:52.
[44] VETRÒ, Antonio, et al. Open Data Quality Measurement Framework: Definition and Application to Open Government Data. En: ScienceDirect. Abril, 2016. vol. 33, no. 2, p. 325-337.
[45] Ibid., p. 334.

lu: Lista de actualización
du: Fecha de actualización

Los sistemas que cumplen con las métricas de calidad de datos, especialmente con trazabilidad pueden traer grandes beneficios a las empresas cuando se utilizan en condiciones adecuadas, por ejemplo: control de procesos, optimización de procesos y mejor comercialización[46].

6.2 COMPLETITUD

La completitud se entiende como la medida en que los datos están completos y son de amplitud y profundidad suficiente para la tarea en la que son utilizados[47]. La expresión matemática para la evaluación de la completitud es la siguiente[48]:

$$ncl=nr*nc \quad (3)$$

$$pcc=\left(1-\frac{ic}{ncl}\right)*100 \quad (4)$$

$$pcpr=\left(1-\frac{nir}{nr}\right)*100 \quad (5)$$

ncl: Numero de celdas
nr: Número de filas
nc: Número de columnas
pcc: Porcentaje de filas completas
ic: Número de celdas incompletas
pcpr: Porcentaje de filas completas
nir: Número de filas incompletas

La completitud de los datos es importante, ya que a la hora de realizar procesos de análisis es necesario tener integridad y disponibilidad de toda la información para que las acciones de análisis sean exhaustivas y puedan brindar su máximo potencial[49].

6.3 CONFORMIDAD

La conformidad es la capacidad de los datos para adherirse a las normas, convenciones o reglamentos en las leyes y prescripciones similares en relación de funcionalidad, confiabilidad, usabilidad, eficiencia, mantenibilidad, efectividad,

[46] FROSCH, Stina. The importance of data quality and traceability in data mining. Applications of robust methods for multivariate data analysis. A case-study conducting the herring industry. Lyngby: Danish Institute for Fisheries Research, Department of Seafood Research & The Technical University of Denmark, 2006. Reporte anual 2006.
[47] PIPINO, Leo; LEE, Yang y WANG, Richard. Data Quality Assessment. En: Communications of the ACM. Abril, 2002. vol. 45, no. 4, p. 211-218.
[48] VETRÒ, Antonio, et al. Open Data Quality Measurement Framework: Definition and Application to Open Government Data. En: ScienceDirect. Abril, 2016. vol. 33, no. 2, p. 325-337.
[49] CHAPMAN, Arthur y SPEERS, Larry. Principles of data quality. 1 ed. Copenhagen: Global Biodiversity Information Facility, 2005. 58 p.

productividad, seguridad y satisfacción[50]. La expresión matemática para evaluar la conformidad es la siguiente[51]:

$$psc=\frac{ns}{nsc}*100 \quad (6)$$

$$egmsc=s+dc+c+t+0.2(d+id+pb+cv+l) \quad (7)$$

psc: Porcentaje de columnas estandarizadas
ns: Número de columnas con estándar asociado
nsc: Número de columnas estandarizadas
egmsc: Completitud según el estándar EGMS
s: Fuente
dc: Fecha de creación
c: Categoría
t: Título
d: Descripción (Opcional)
id: Identificador (Opcional)
pb: Publicación (Opcional)
cv: Cobertura (Opcional)
l: Lenguaje (Opcional)

El cumplimiento es importante en calidad de datos, ya que representa una forma de control de calidad enfocado a eliminar el error mediante unos procesos de producción en las bases de datos[52]. El parágrafo 1 del artículo 9 de la ley de datos abiertos se refiere a la forma de publicación de los conjuntos, tal que facilite el uso y compresión de ellos a las personas y que permita verificar su calidad[53], el cumplimiento facilita este proceso con la eliminación del error en los datos.

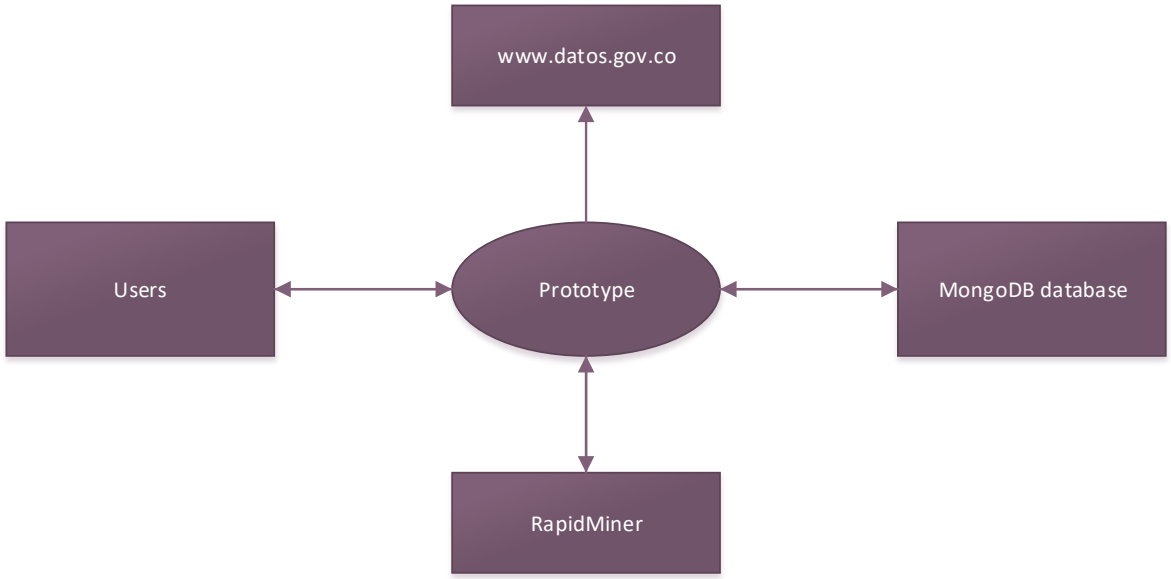
[50] VANÍČEK, Jiří. Software and data quality. En: Czech Academy of Agricultural Sciences. Febrero, 2006. vol. 52, no. 3, p. 138-146.
[51] VETRÒ, Antonio, et al. Open Data Quality Measurement Framework: Definition and Application to Open Government Data. En: ScienceDirect. Abril, 2016. vol. 33, no. 2, p. 325-337.
[52] VEREGIN, H. Data quality parameters. En: Geographical information systems. 1999. vol. 1, no. 12, p. 177-189.
[53] COLOMBIA. CONGRESO DE LA REPÚBLICA. Ley 1712. (06, marzo, 2014). Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones. MINTIC. Bogotá, D.C., 2014. P. 1-14.

7 DISEÑO

7.1 DIAGRAMA DE CONTEXTO

El prototipo interactúa con 4 sistemas diferentes (véase la figura 4). www.datos.gov.co provee los datos con la que va a trabajar el prototipo, estos datos se almacenarán en la base de datos no relacional MongoDB, ya que brinda flexibilidad, rendimiento y permite al prototipo ser escalable para evaluar cualquier conjunto de datos[54]. En RapidMiner se realizarán diferentes acciones respecto al análisis y representación de la información extraída. Finalmente, el usuario visualiza y analiza los resultados que provee el prototipo.

Figura 4. Diagrama de contexto



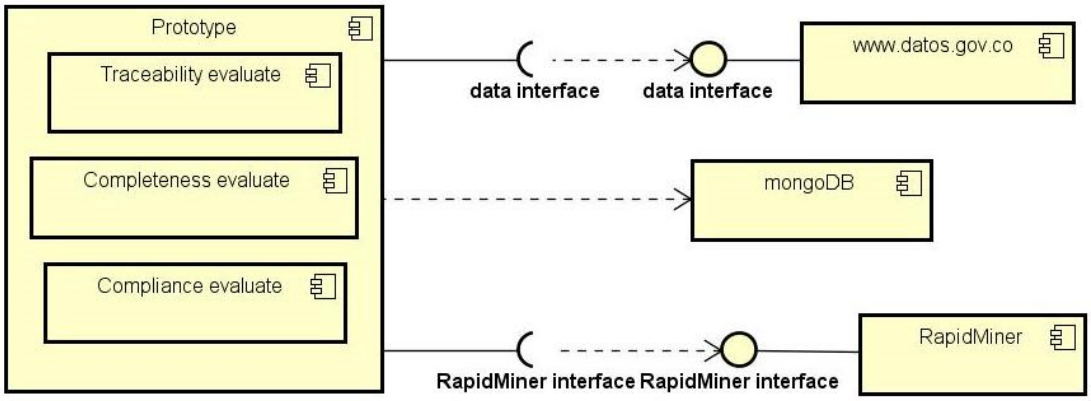
Fuente: Los autores.

7.2 DIAGRAMA DE COMPONENTES

Se observa la estructura del prototipo (véase la figura 5). Está compuesto por 3 componentes, la evaluación de la trazabilidad, completitud y conformidad. El prototipo consume la interfaz ofrecida de www.datos.gov.co para la extracción de los datos, que se guardan en la base de datos MongoDB. El prototipo también consume la interfaz de RapidMiner para el análisis de los datos, esto genera unos resultados que se presenta al usuario.

[54] BANKER, Kyle. MongoDB in action. 2 ed. New York: Manning Publications Co., 2012. 312 p.

Figura 5. Diagrama de componentes

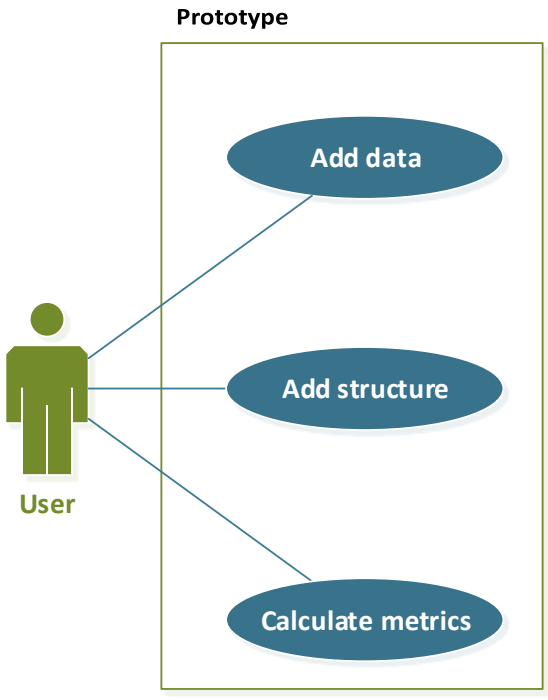


Fuente: Los autores.

7.3 DIAGRAMA DE CASOS DE USO

El diagrama de casos de uso muestra los módulos que posee el sistema. Se representa los 3 módulos que posee el prototipo (véase la figura 6). Igualmente representa que el usuario tiene acceso a cualquiera de los 3 módulos.

Figura 6. Diagrama de casos de uso

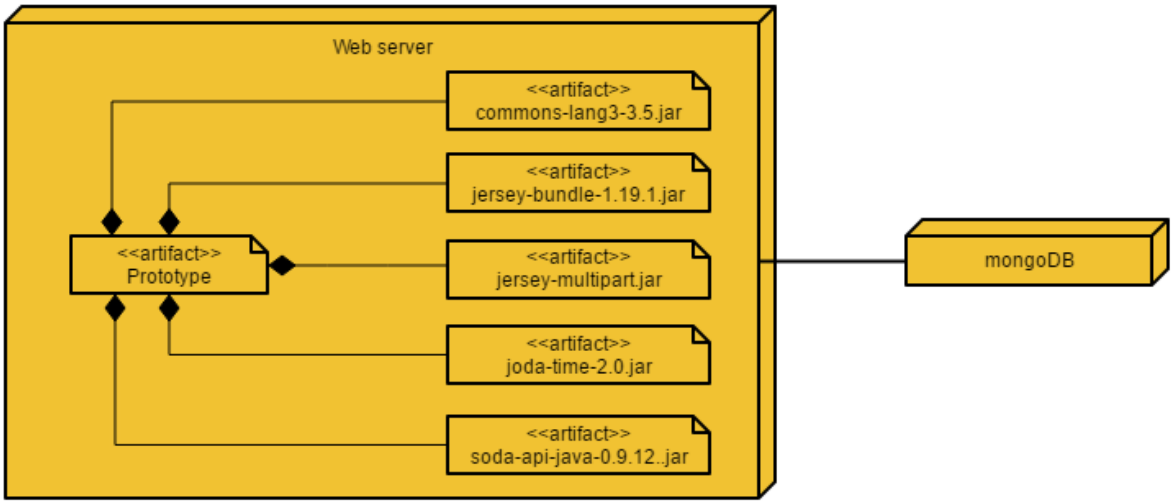


Fuente: Los autores.

7.4 DIAGRAMA DE DESPLIEGUE

El diagrama de despliegue contempla el prototipo alojado en un servidor web junto a los componentes necesarios para su ejecución (véase la figura 7).

Figura 7. Diagrama de despliegue.



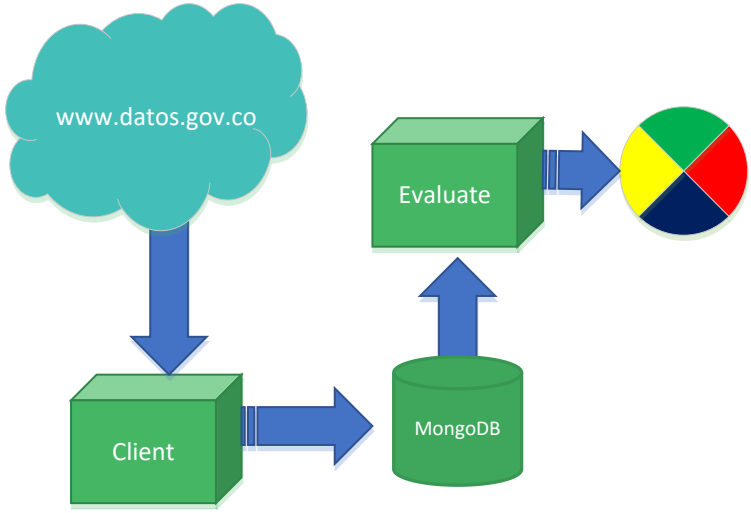
Fuente. Los autores.

7.5 MOCKUP ARQUITECTURA

La propuesta de la arquitectura se presenta en la figura (véase la figura 8). El proceso inicia con la extracción de los datos del repositorio www.datos.gov.co mediante un cliente utilizando JAVA como lenguaje de programación. Esta información se almacena en una base de datos no relacional, se decidió almacenar en MongoDB, esta permite almacenar los datos sin tener en cuenta la estructura de la base de datos[55]. Toda la información es dirigida hacia RapidMiner. Esta permite diseñar procesos de minería de datos cajas que representan módulos o actividades llamados procesos, esto con el fin de realizar flujos de datos o control sin programación[56]. Estos son destinados para el análisis y procesamiento de la información mediante gráficos y estadísticas.

[55] BANKER, Kyle. MongoDB in action. 2 ed. New York: Manning Publications Co., 2012. 312 p.
[56] HOFMANN, Markus y KLINKENBERG, Ralf. RapidMiner: Data Mining Use Cases and Business Analytics Applications. ed. 1. Florida: CRC Press, 2013. 525 p.

Figura 8. Mockup de la arquitectura



Fuente: Los autores.

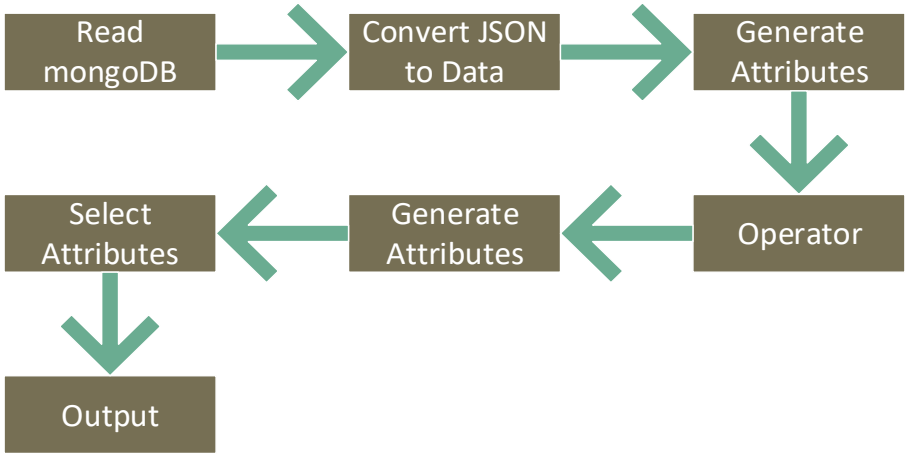
8 IMPLEMENTACIÓN

8.1 TRAZABILIDAD

Para la evaluación de la trazabilidad se utiliza la fórmula número 1 y número 2 ya descritas en el apartado de análisis de cada una de las métricas. La evaluación se realiza en el software RapidMiner, siendo la entrada los registros existentes en la base de datos previamente extraídos del repositorio www.datos.gov.co, y la salida el valor en porcentaje de la métrica.

$$tc=2s+dc \quad (1)$$
$$tu=lu+du \quad (2)$$

Figura 9. Proceso de evaluación de la métrica de trazabilidad.

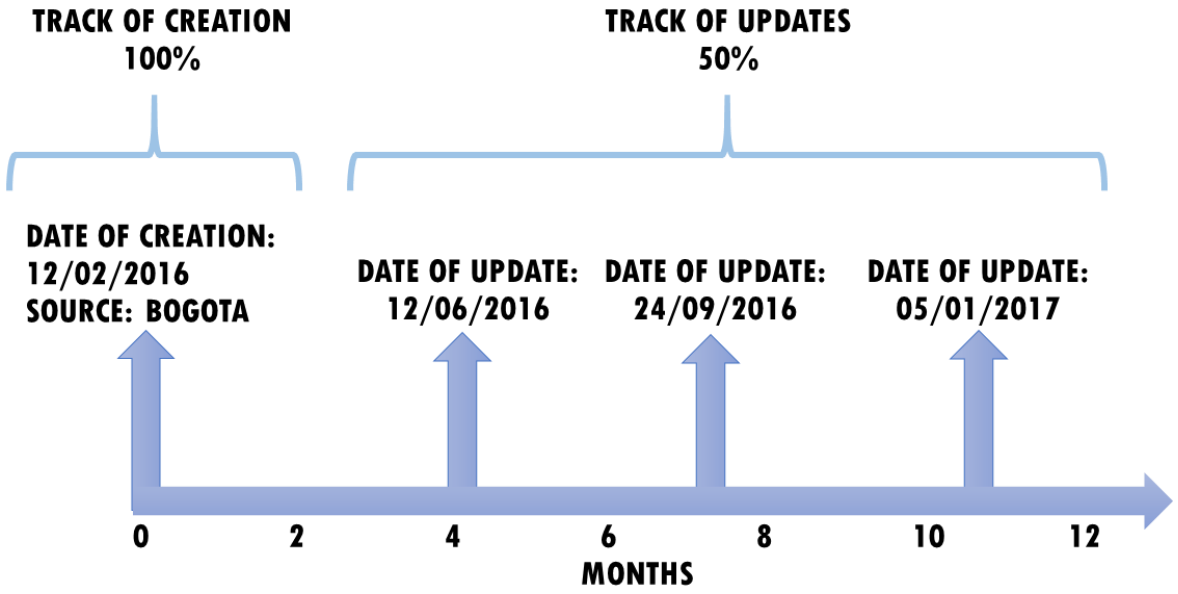


Fuente: Los autores.

Se representa el proceso para el cálculo de la métrica (véase la figura 9). Inicialmente RapidMiner consulta la base de datos donde se encuentran los registros, el resultado de la consulta es un formato JSON con toda la información. Para que RapidMiner pueda utilizar estos campos es necesario pasar el JSON a una estructura que entienda RapidMiner, éste proceso se realiza en la conversión de JSON a datos. Una vez están todas las filas y columnas en RapidMiner se procede a obtener todos los datos necesarios para la aplicación de la fórmula, en la primera generación de atributos se obtiene si existe la fuente de la información, fecha de creación, fecha de actualización, lista de actualización. Una vez se obtiene la información para cada conjunto de datos se aplica la formula. En el operador se realiza un promedio de los datos obtenidos según cada fórmula para finalmente en la última generación de atributos obtener el porcentaje. Se seleccionan los campos de salida para publicarlos en el servidor e imprimir.

8.1.1 Resultados y análisis. Para la evaluación de la métrica de trazabilidad se escogieron aleatoriamente 6 conjuntos de datos y se ingresaron en el prototipo. Los identificadores para cada conjunto de datos en el repositorio son 3piv-wxdz, vu85-kh7n, c594-32w3, fhr6-myxs, 5eqb-e4gs e igux-a5yx, correspondientes a las categorías de agricultura, ciencia, justicia, organismos de control, seguridad y defensa, y transporte. Después de realizar el proceso de evaluación en el prototipo se obtienen los resultados (véase la figura 10).

Figura 10. Resultados de la métrica de trazabilidad.



Fuente: Los autores.

La trazabilidad se mide con dos fórmulas, una mide la traza de los datos en su creación, y otra el seguimiento que se le ha hecho hasta la fecha de evaluación, la conjunción de ambos es el resultado de la métrica de trazabilidad.

La traza de creación representa la presencia de metadatos asociados con la creación de los conjuntos de datos. Se encontró que el 100% de los metadatos asociados con la creación de los conjuntos de datos evaluados son existentes, lo cual es positivo, ya que si cualquier persona desea obtener la información referente a las fechas de creación y quién subió la información la puede obtener.

La traza de actualización representa la presencia de metadatos asociados con la actualización de los conjuntos de datos. Se encontró que solamente el 50% de los metadatos asociados con la actualización de los conjuntos de datos evaluados poseen información y el 50% no lo tiene, se observó que el factor común que tienen los conjuntos de datos es la presencia de la fecha de actualización del conjunto de datos y la ausencia de la lista de actualizaciones realizadas.

8.2 COMPLETITUD

La evaluación de la métrica de completitud se conforma de las siguientes fórmulas:

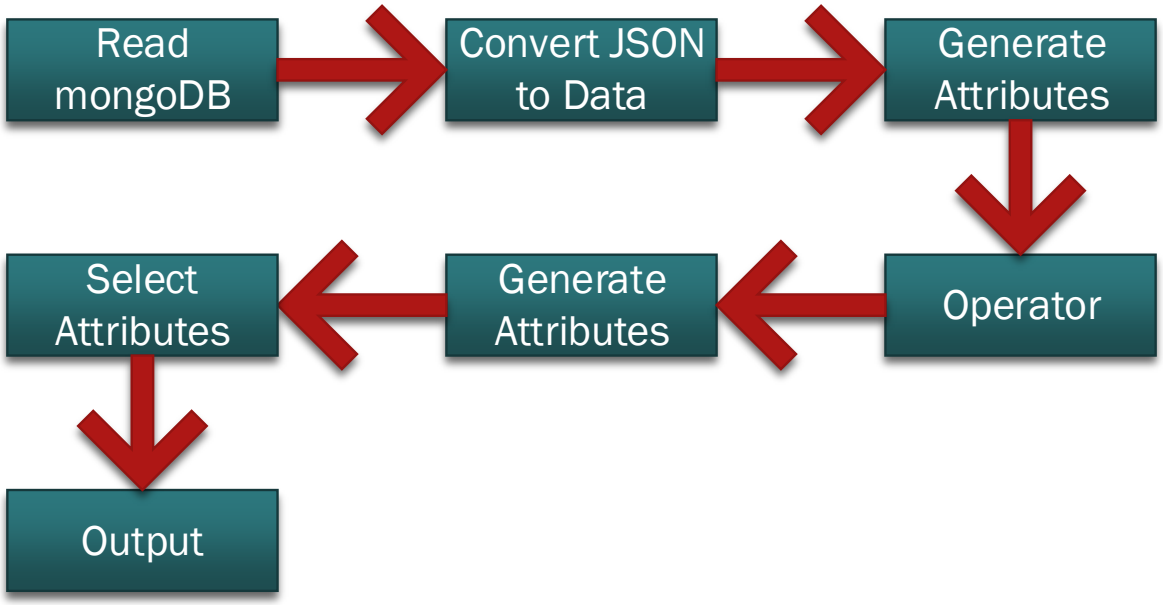
$$ncl = nr*nc \quad (3)$$

$$pcc = \left(1-\frac{ic}{ncl}\right)*100 \quad (4)$$

$$pcpr = \left(1-\frac{nir}{nr}\right)*100 \quad (5)$$

La evaluación se realiza en el software RapidMiner, siendo la entrada los registros existentes en la base de datos previamente extraídos del repositorio www.datos.gov.co, y la salida el valor en porcentaje de la métrica.

Figura 11. Proceso de evaluación de la métrica de completitud.



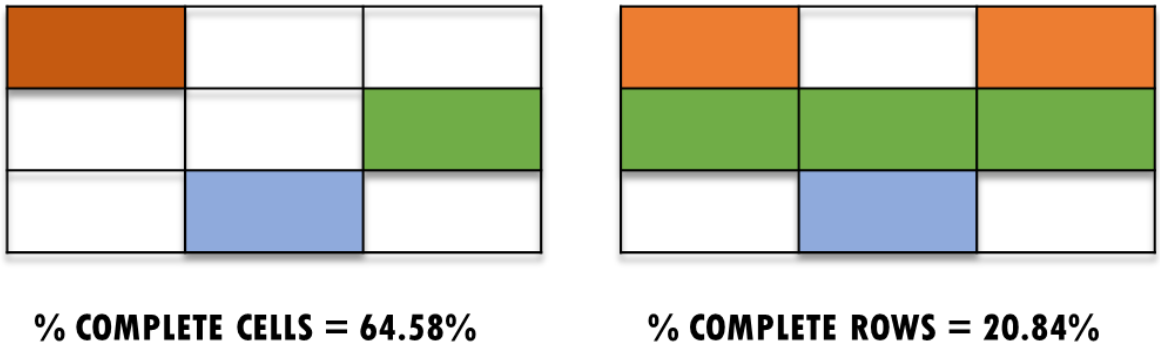
Fuente. Los autores.

Se muestra el proceso para el cálculo de la métrica (véase la figura 11). Inicialmente RapidMiner consulta la base de datos donde se encuentran los registros, el resultado de la consulta es un formato JSON con toda la información. Para que RapidMiner pueda utilizar estos campos es necesario pasar el JSON a una estructura que entienda RapidMiner, éste proceso se realiza en la conversión de JSON a datos. Una vez están todas las filas y columnas en RapidMiner se procede a obtener todos los datos necesarios para la aplicación de la fórmula, en la primera generación de atributos se calcula el número de celdas incompletas para cada registro y el número de columnas. En el operador se realiza la suma de todas las celdas incompletas y se calcula el número de registros existentes. Finalmente en la

última generación de atributos se aplican la fórmula 3 para hallar el número total de celdas y después la fórmula 4 y 5 que comprende el resultado de la métrica. Se seleccionan los campos de salida para publicarlos en el servidor e imprimir.

8.2.1 Resultados y análisis. Para la evaluación de la métrica de completitud se escogieron aleatoriamente 6 conjuntos de datos y se ingresaron en el prototipo. Los identificadores para cada conjunto de datos en el repositorio son 3piv-wxdz, vu85-kh7n, c594-32w3, fhr6-myxs, 5eqb-e4gs e igux-a5yx, correspondientes a las categorías de agricultura, ciencia, justicia, organismos de control, seguridad y defensa, y transporte. Después de realizar el proceso de evaluación se obtuvieron los resultados (véase la figura 12).

Figura 12. Resultados de la métrica de completitud.



Fuente: Los autores.

La primera fórmula representa el porcentaje de celdas que poseen datos no nulos y que aportan información a los registros que pertenezcan. El 64.58% de las celdas poseen un valor, mientras el 35.42% son celdas vacías no válidas. El valor que nos aporta el cálculo representa que alrededor del 6/10 de los campos poseen un dato, sin embargo, el hecho que existan campos vacíos da lugar a información incompleta o poco certera para determinadas situaciones, lo que es un factor negativo y para mejorar.

La falta de datos genera dudas en la fuente, porque la información computarizada es un reflejo de los mismos[57]. La integridad en conjunto con la calidad de los datos se hacen presentes en la ley 1712 de 2014[58] que dice: “Principio de la calidad de la información. Toda la información de interés público que sea producida, gestionada y difundida por el sujeto obligado, deberá ser oportuna, objetiva, veraz, completa,

[57] RUTHBERG, Zella y POLK, William. Report of the Invitational Workshop on Data Integrity. Gaithersburg: National Institute of Standards and Technology, 1999. Reporte número 500-168.
[58] COLOMBIA. CONGRESO DE LA REPÚBLICA. Ley 1712. (06, marzo, 2014). Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones. MINTIC. Bogotá, D.C., 2014. p. 1-14.

reutilizable, procesable y estar disponible en formatos accesibles para los solicitantes e interesados en ella, teniendo en cuenta los procedimientos de gestión documental de la respectiva entidad.”.

La segunda fórmula representa el porcentaje de filas que poseen datos no nulos y que aportan información. El 20.84% de las filas poseen un valor, mientras el 79.16% son filas vacías no válidas. Estos valores aseguran que la información es incompleta y que existen campos que las organizaciones prefieren no llenar. Esto se genera por la falta de regulación de la entidad de los datos que se están ingresando. Igualmente no es posible realizar estadísticas, promedios o análisis con la información con los vacíos que presentan los repositorios, lo que pone en tela de juicio el cumplimiento de la ley.

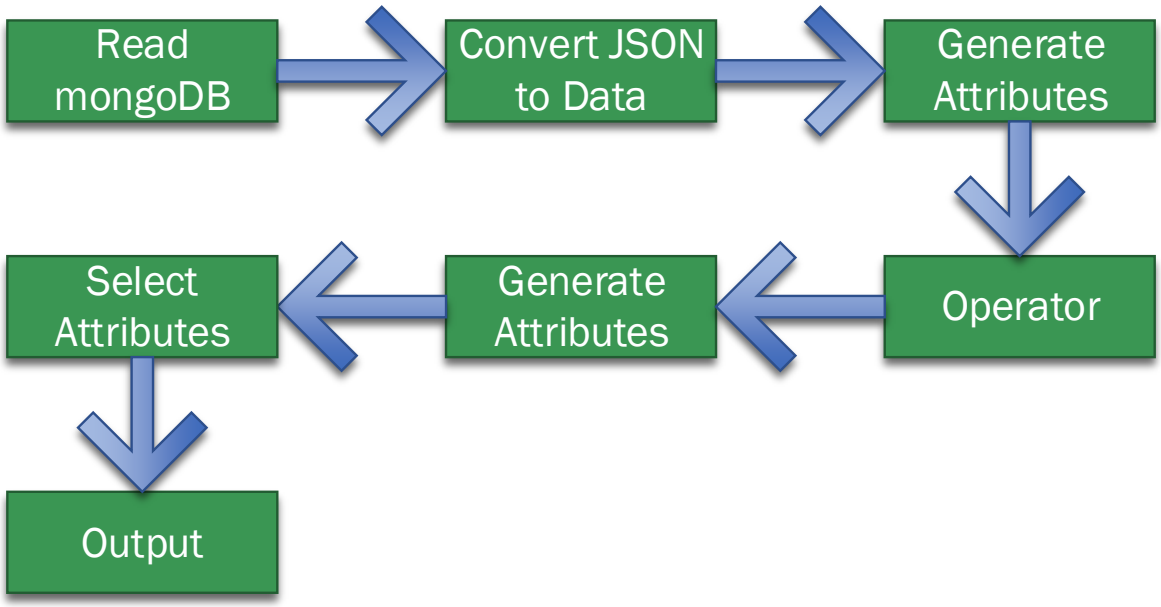
8.3 CONFORMIDAD

Para la evaluación de la métrica de conformidad se utiliza la fórmula número 6 y número 7 presentadas a continuación. La evaluación se realiza en el software RapidMiner, siendo la entrada los registros existentes en la base de datos previamente extraídos del repositorio www.datos.gov.co, y la salida el valor en porcentaje de la métrica.

$$psc = \frac{ns}{nsc} * 100 \qquad (6)$$

$$egmsc = s + dc + c + t + 0.2(d + id + pb + cv + l) \qquad (7)$$

Figura 13. Proceso para el cálculo de conformidad.

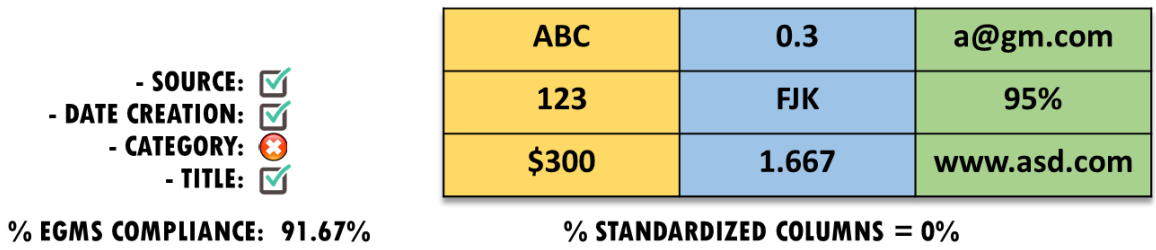


Fuente: Los autores.

Se muestra el proceso para el cálculo de la métrica de conformidad (véase la figura 13). El proceso inicia con la consulta la base de datos donde se encuentran los registros, el resultado de la consulta es un JSON con toda la información. Para que RapidMiner pueda utilizar estos campos es necesario pasar el JSON a una estructura que entienda RapidMiner, éste proceso se realiza en la conversión de JSON a datos. Una vez están todas las filas y columnas en RapidMiner se procede a obtener todos los datos necesarios para la aplicación de la fórmula, en la primera generación de atributos se calcula el número de columnas con un estándar asociado y el número de campos que cumplen con el estándar eGMS. Después en el operador se realiza un promedio para obtener el número de columnas estandarizadas y el promedio de campos que cumplen el estándar eGMS. Una vez realizado el cálculo se realiza una nueva generación de atributos para aplicar las formulas 6 y 7. Una vez obtenidos los resultados se seleccionan los campos de salida del proceso y se publican al servidor para su tratamiento.

8.3.1 Resultados y análisis. Para la evaluación se ingresaron 6 conjuntos de datos. Los identificadores para cada conjunto de datos en el repositorio son 3piv-wxdz, vu85-kh7n, c594-32w3, fhr6-myxs, 5eqb-e4gs e igux-a5yx, correspondientes a las categorías de agricultura, ciencia, justicia, organismos de control, seguridad y defensa, y transporte. Después de realizar el proceso de evaluación se obtuvieron los resultados (véase la figura 14).

Figura 14. Resultados de la métrica de conformidad.



Fuente: Los autores.

La primera fórmula representa el porcentaje de columnas estandarizadas en los conjuntos de datos. Ninguna de las columnas cumple la estandarización de columnas, lo que hace pensar que el repositorio de datos www.datos.gov.co no se preocupa por éste factor y no existe un control en el formato con el que se sube la información y la cantidad de campos vacíos.

En caso de querer realizar modelos de datos se dificulta, esto se relaciona a que hay muchos campos vacíos, trabajar con esta información puede llegar a ser una

tarea imposible. La entidad que sube esta información podría estar incurriendo en información falsa o mal digitada, lo cual podría generar problemas legales.

La segunda fórmula representa el porcentaje de los registros de metadatos que cumplen el estándar eGMS. La aplicación muestra que el 91.67% de los registros poseen los campos regidos por el estándar. Demuestra que el repositorio de datos www.datos.gov.co está enfocado en el estándar eGMS, que consiste en el registro de fuentes, fecha de creación de repositorio, categoría y título, sin embargo, hay conjuntos de datos que carecen de alguno de estos datos.

La evaluación general de calidad de datos fue de 54,515% después de aplicar las 3 métricas a los conjuntos de datos evaluados.

9 PRUEBAS

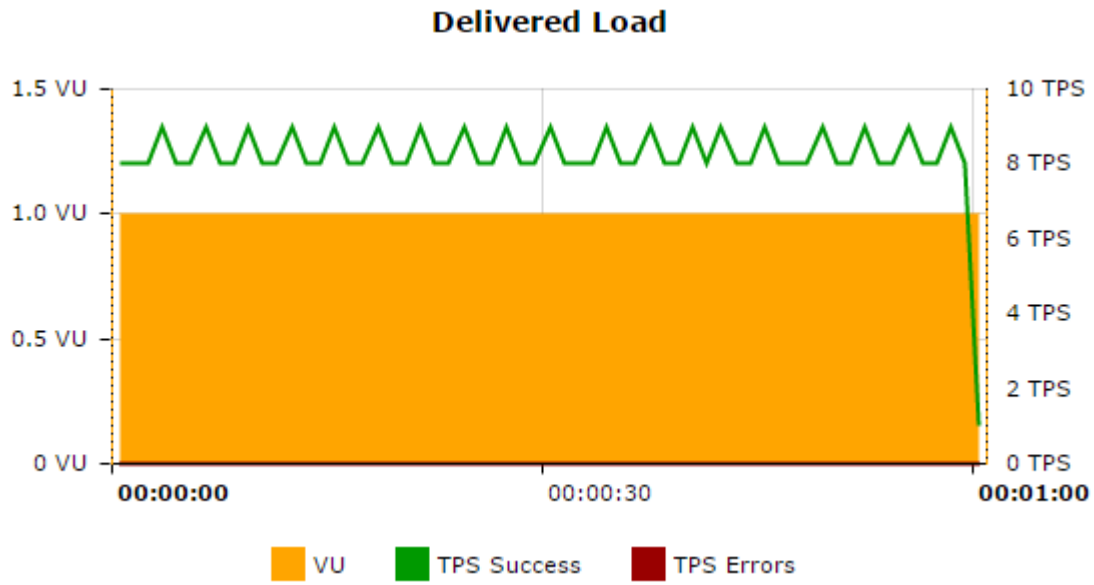
Para la realización de las pruebas se utilizó un equipo con las siguientes características:

- Microsoft Windows 10 64 bits.
- Intel core i5-4210U 1.70GHz 2.20GHz.
- 8GM RAM.
- JAVA versión 8.
- MongoDB versión 3.4.2.
- Apache Tomcat versión 8.0.
- Rapidminer versión 7.4.0.
- Apache JMeter versión 3.1.

9.1 PRUEBAS DE ESTRÉS

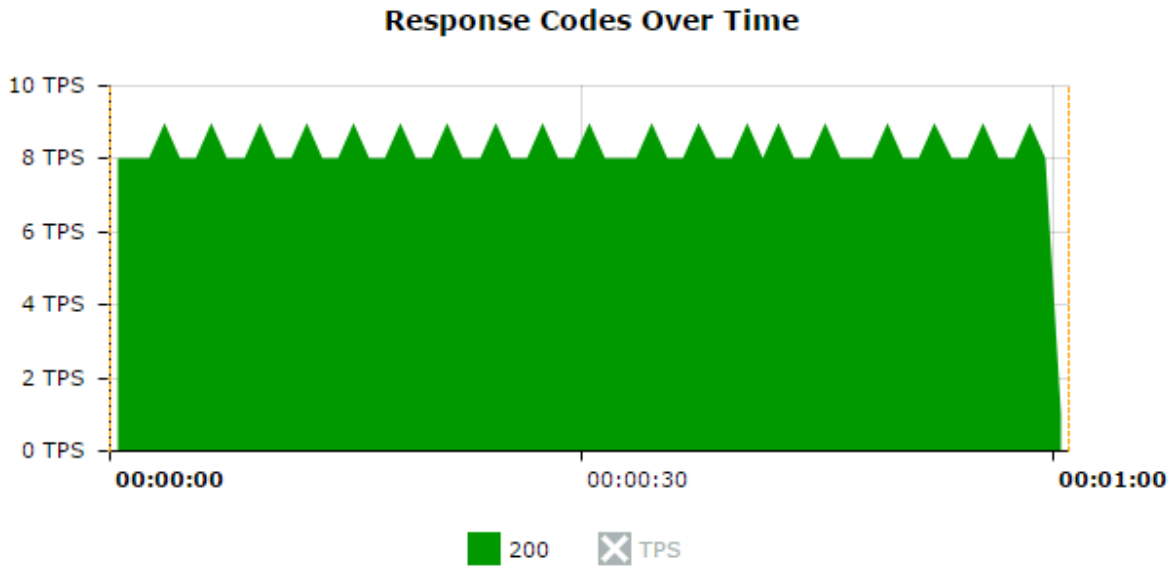
9.1.1 Módulo de insertar estructura. Se tienen 500 hilos ingresando un conjunto de datos. El tiempo de la prueba de los hilos es de 1 minuto para todos. Los resultados son los siguientes.

Figura 15. Transacciones en el servidor.



Fuente: Los autores.

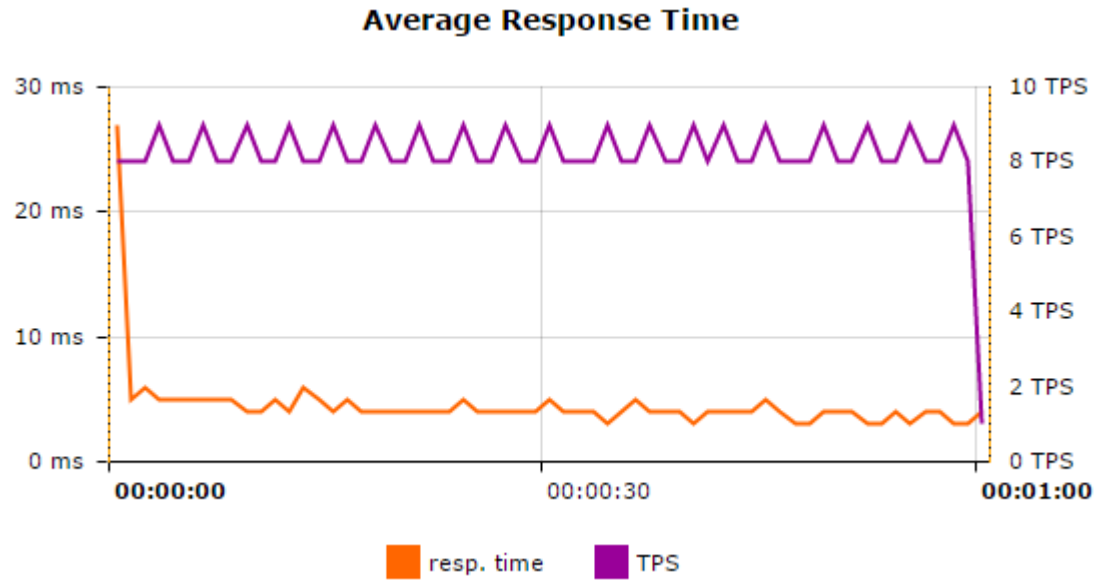
Figura 16. Retorno de las transacciones.



Fuente: Los autores.

Se procesaron un promedio entre 1 y 2 peticiones (véase la figura 15) donde ninguna fue errónea (véase la figura 16). Con esto se deduce que el servidor tiene una buena tolerancia para el manejo de recepción y ejecución de peticiones.

Figura 17. Tiempo de respuesta por petición.



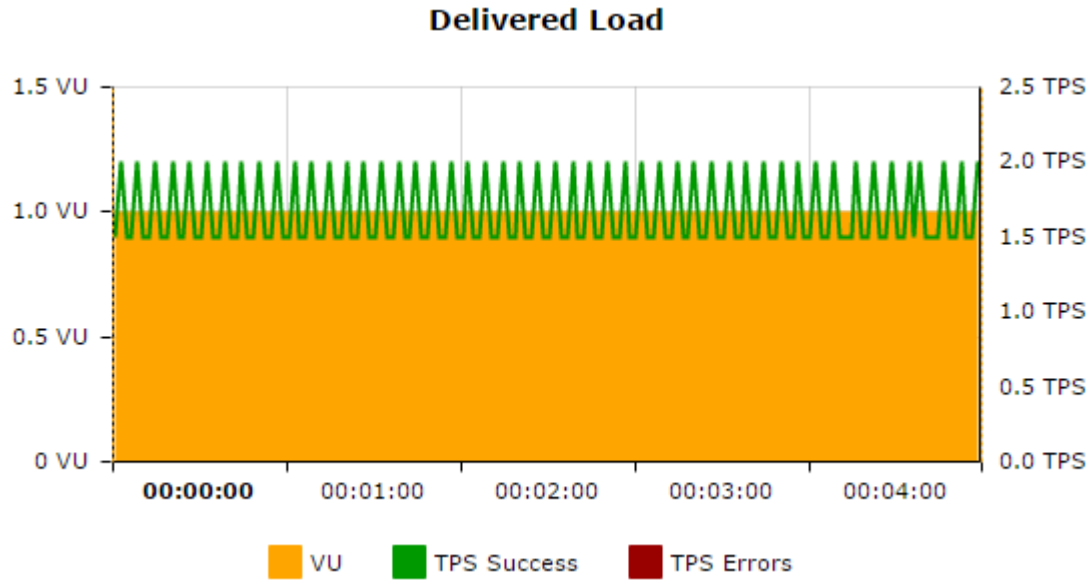
Fuente: Los autores.

Se aprecia el tiempo de respuesta según la cantidad de peticiones (véase la figura 17). El tiempo de respuesta varía entre 1 y 8 milisegundos, donde en 8 milisegundos

atendió un promedio de 8 peticiones y en 1 milisegundo atiende un promedio de 1 milisegundos.

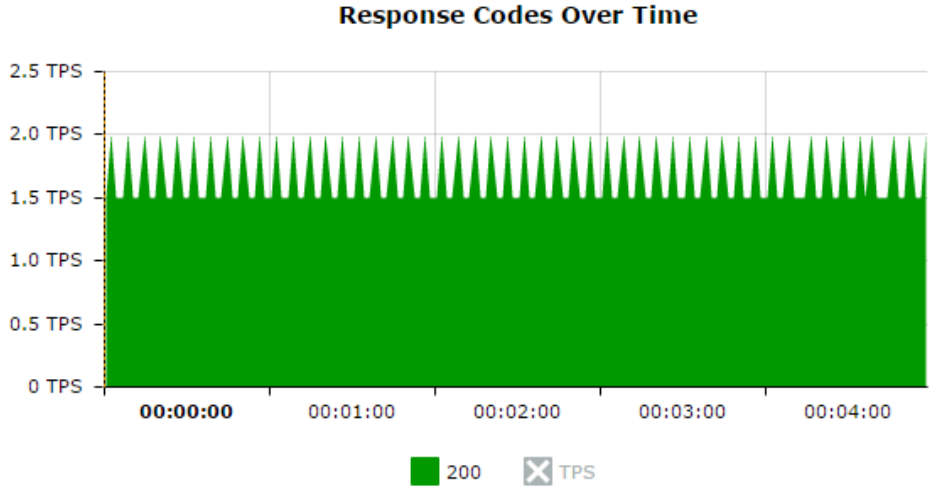
9.1.2 Módulo de insertar datos. Se tienen 500 hilos enviando 5 identificadores de conjuntos de datos, 2 códigos correctos y 3 códigos incorrectos. El tiempo de la prueba de los hilos es de 5 minutos, los resultados son los siguientes.

Figura 18. Transacciones en el servidor.



Fuente: Los autores.

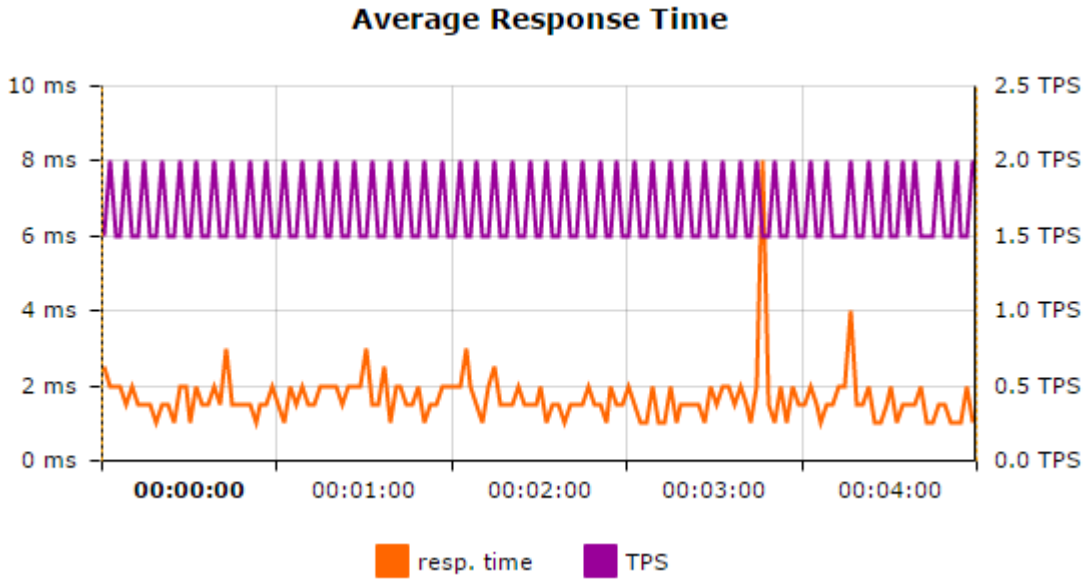
Figura 19. Retorno de las transacciones.



Fuente: Los autores.

Se procesaron un promedio entre 1 y 2 peticiones (véase la figura 18) donde ninguna fue errónea (véase la figura 19). Con esto se deduce que el servidor tiene una buena tolerancia para el manejo de recepción y ejecución de peticiones.

Figura 20. Tiempo de respuesta por petición.

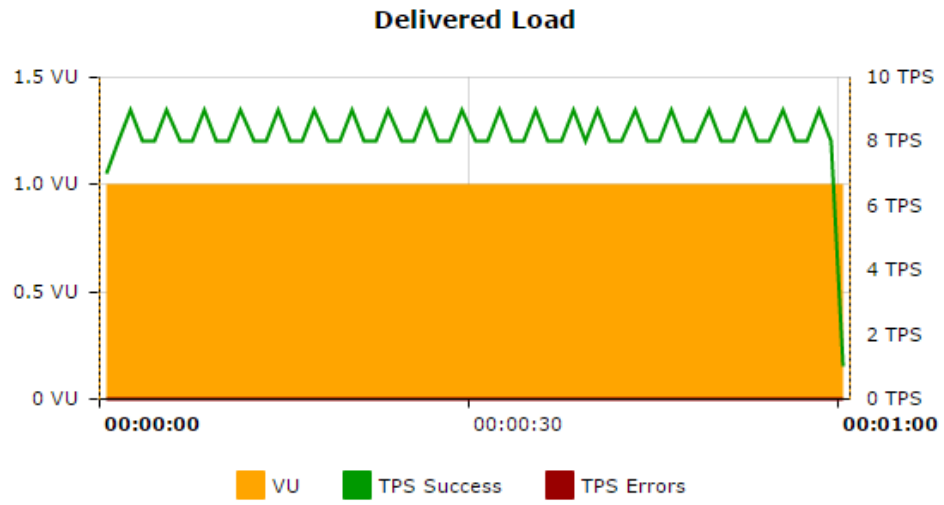


Fuente: Los autores.

Se aprecia el tiempo de respuesta según la cantidad de peticiones (véase la figura 20). El tiempo de respuesta varía entre 1 y 8 milisegundos, donde en 8 milisegundos atendió un promedio de 8 peticiones y en 1 milisegundo atiende un promedio de 1 milisegundos.

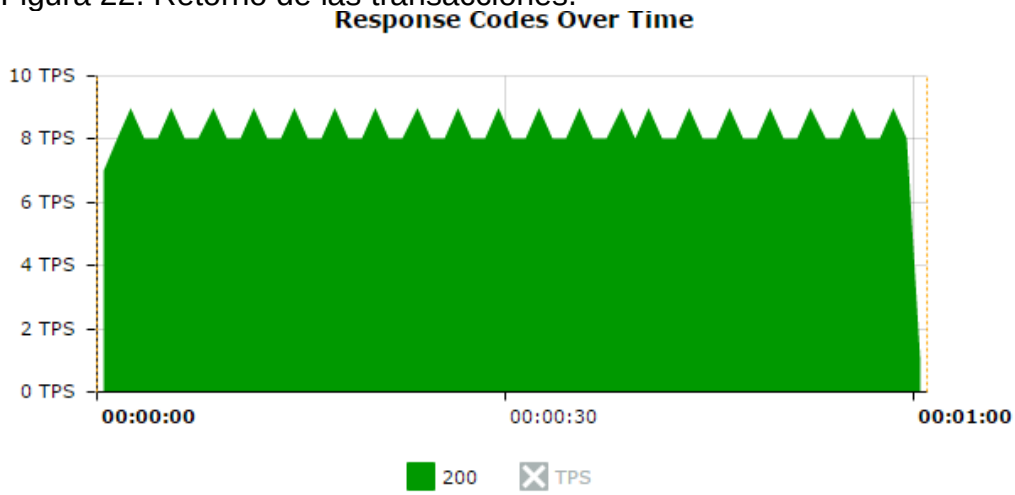
9.1.3 Módulo de cálculo de métricas. El escenario inicial fue de 500 peticiones en 1 minuto al módulo de la evaluación de las métricas con los datos existentes en la base de datos. El rendimiento fue de 500,676 peticiones por minuto.

Figura 21. Transacciones en el servidor.



Fuente: Los autores.

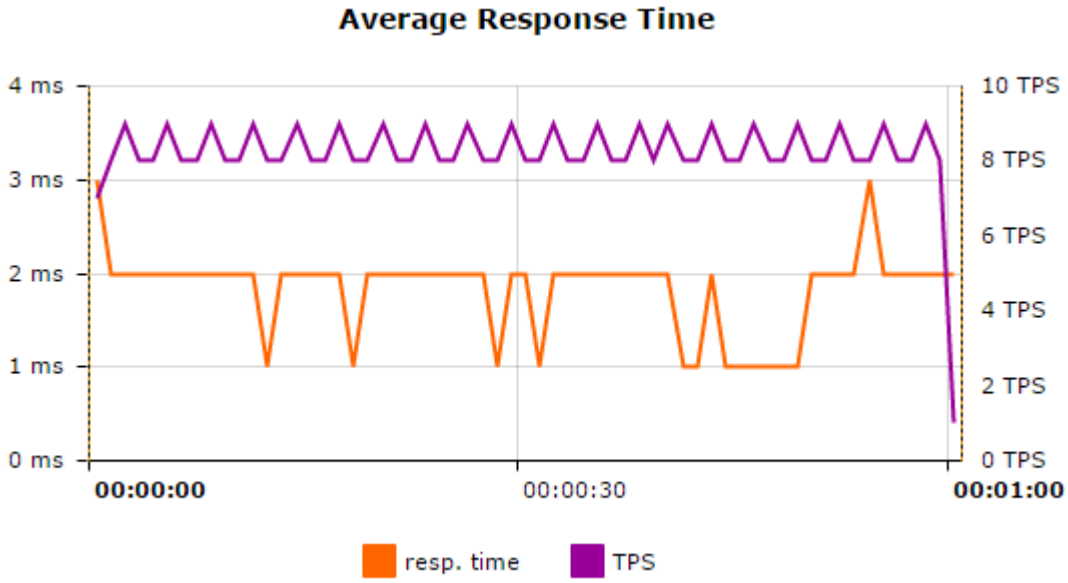
Figura 22. Retorno de las transacciones.



Fuente: Los autores.

Se procesaron un promedio entre 8 y 9 peticiones (véase la figura 21), donde ninguna fue errónea (véase la figura 22). Con esto se deduce que el servidor tiene una buena tolerancia para el manejo de recepción y ejecución de peticiones.

Figura 23. Tiempo de respuesta.



Fuente: Los autores.

Se aprecia el tiempo de respuesta según la cantidad de peticiones (véase la figura 23). El tiempo de respuesta varía entre 1 y 3 milisegundos, donde en 3 milisegundos atiende un promedio de 8 peticiones y en 1 milisegundo atiende un promedio de 2 milisegundos.

10 CONCLUSIONES

Dentro del análisis se eligen las métricas de trazabilidad, completitud y conformidad para el cálculo de la calidad de datos por la importancia y el valor agregado que dan al cumplimiento de la ley de datos abiertos, en el artículo 9 se habla de la forma y la calidad con la que se maneja la información. Otro motivo de la elección de las métricas es el proyecto de derecho mencionado en el documento, el enfoque de ellas en una compañía trae beneficios a las compañías en sus tareas, promueve análisis exhaustivos y con mejor potencial, y puede eliminar el error que pueda existir en los datos.

Para realizar las pruebas a los conjuntos de datos alojados en www.datos.gov.co se eligieron 6 conjuntos de datos al azar y se encontró que las estructuras obtenidas no son homogéneas, esto dificulta el análisis y el entendimiento de la información allí reflejada. Por ejemplo, se eligieron dos estructuras, 3piv-wxdz y c594-32w3, y se encontró que ambas tenían estructuras totalmente diferentes, incluyendo los metadatos asociados.

El repositorio www.datos.gov.co no posee una manera lógica de diferenciar los conjuntos de datos, esto debido a que los identificadores que hacen referencia a los conjuntos no tienen un orden o una estructura en las letras y números que la compone, sino una combinación aleatoria. Un ejemplo de esto pueden ser los conjuntos de datos j3bg-66aw y mpka-keq9. Aunque son de la misma entidad y poseen la misma estructura, no hay ninguna forma de conocer datos básicos del conjunto de datos, por ejemplo, la fuente.

El repositorio www.datos.gov.co no posee una buena base de información para el tratamiento público, sin embargo existen más métricas para la evaluación de calidad de datos. Se puede apreciar que los conjuntos de datos evaluados j3bg-66aw, whvw-q2qr, unsz-yhtr, 4v3y-ijrt, h6cd-ixp7 y jqhv-pcam no poseen información completa, de calidad y confiable, los estudios y análisis que se realicen posiblemente sean erróneos y alejados de la realidad, lo que significa una baja calidad en la creación de productos o prestación de servicios que tengan como fuente este repositorio. Por ejemplo, el conjunto de datos j3bg-66aw en la evaluación de trazabilidad obtuvo un resultado de 75%, conformidad de 50% y completitud 45,87%, con una evaluación general de las 3 métricas de 56,96%.

11 RECOMENDACIONES

Para los interesados en continuar el proyecto se recomienda optimizar el código JAVA de extracción y almacenamiento de los datos, organizando mejor las clases y analizando la manera con la que se realizaron estas tareas. Otra recomendación es la complementación del prototipo con la implementación de otras métricas asociadas a la calidad de los datos, por ejemplo exactitud y comprensibilidad.

Una mejora a realizar al prototipo es la compatibilidad de más formatos abiertos, como lo son CSV, PDF, TXT, RDF, TSV, XML, JPEG, DJVU, EPUB, HTML, RTF, PNG, SVG y VP9.

REFERENCIAS

ALSHAMRANI, Adel y BAHATTAB, Abdullah. A Comparison Between Three SDLC Models Waterfall Model, Spiral Model, and Incremental/Iterative Model. En: International Journal of Computer Science Issues. Enero, 2015. vol. 12, no. 1, p. 106-111.

ARCILA, Carlos; BARBOSA, Eduar y CABEZUELO, Francisco. Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística. En: El profesional de la información. Julio, 2016. vol. 25, no. 4, p. 623-631.

AZUMAH, Kenneth y QUARSHIE, Henry. Towards Higher Quality Data: Impact of Perception of Data Quality on IT Investment - Ghana. En: International Journal of Emerging Trends in Computing and Information Sciences. Enero, 2013. vol. 3, no. 12, p. 1614-1621.

BANKER, Kyle. MongoDB in action. 2 ed. New York: Manning Publications Co., 2012. 312 p.

BATINI, Carlo, et al. Methodologies for Data Quality Assessment and Improvement. En: ACM Computing Surveys. Julio, 2009. vol. 41, no. 3, p. 16:1-16:52.

BOEHM, Barry. Spiral Development: Experience, Principles, and Refinements Spiral Development Workshop February 9, 2000. Los Angeles, Software Engineering Institute, 2000. 37 p. CMU/SEI-2000-SR-008.

CABRERA, Yandy. Transferencia de estado representacional (REST): estilo de arquitectura para sistemas distribuidos de hipermedia. En: Serie científica de la universidad de las ciencias informáticas. Julio, 2013. vol. 6, no. 7.

CHAPMAN, Arthur y SPEERS, Larry. Principles of data quality. 1 ed. Copenhagen: Global Biodiversity Information Facility, 2005. 58 p.

COLOMBIA. CONGRESO DE LA REPÚBLICA. Ley 1712. (06, marzo, 2014). Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones. MINTIC. Bogotá, D.C., 2014. P. 1-14.

ECKERSON, Wayne. Data quality and the bottom line. Chatsworth, 101 Communications LLC, 2002. 33 p.

ERL, Thomas. Service-Oriented Architecture: A Field Guide to Integrating XML and Web Services. 1 ed. New Jersey: Prentice Hall PTR, 2004. 541 p.

E-skills UK, Big data analytics: adoption and employment trends, 2012-2017. En: VOCED. Noviembre, 2013. 22 p.

FERRER, Antonia y SÁNCHEZ, Enrique. Open data, big data: ¿hacia dónde nos dirigimos?. En: Anuario ThinkEPI. Febrero, 2012. vol. 7, p. 150-156.

FROSCH, Stina. The importance of data quality and traceability in data mining. Applications of robust methods for multivariate data analysis. A case-study conducting the herring industry. Lyngby: Danish Institute for Fisheries Research, Department of Seafood Research & The Technical University of Denmark, 2006. Reporte anual 2006.

GARRIGA, Marc. ¿Datos abiertos? Sí, pero de forma sostenible. En: El profesional de la información. Mayo, 2011. vol. 20, no. 3, p. 298-303.

HAUG, Anders; ZACHARIASSEN, Frederik y VAN LIEMPD, Dennis. The costs of poor data quality. En: Journal of Industrial Engineering and Management. Enero, 2011. vol. 4, no. 2, p. 168-193.

HERNÁNDEZ-PÉREZ, Tony. En la era de la web de los datos: primero datos abiertos, después datos masivos. En: El profesional de la información. Julio, 2016. vol. 25, no. 4, p. 517-525.

HERZOG, Thomas; SCHEUREN, Fritz y WINKLER, William. Data Quality and Record Linkage Techniques. 1 ed. New York: Springer-Verlag New York, 2007. 234 p.

HOFMANN, Markus y KLINKENBERG, Ralf. RapidMiner: Data Mining Use Cases and Business Analytics Applications. ed. 1. Florida: CRC Press, 2013. 525 p.

ISAIAS, Pedro y ISSA, Tomayess. High Level Models and Methodologies for Information Systems. 1 ed. New York: Springer, 2015. 145 p.

JAVANMARD, Mahdi y ALIAN, Maryam. Comparison between Agile and Traditional software development methodologies. En: Science Journal (CSJ). Mayo, 2015. vol. 36, no. 3, p. 1386-1394.

KOTHARI, C. Research Methodology. 2 ed. New Delhi: New Age International (P) Limited, Publishers, 2004. 418 p.

LEMAY, Laura y CADENHEAD, Rogers. Sams Teach Yourself Java 2 in 21 Days. 3 ed. Indianapolis: SAMS Publishing, 2002. 736 p.

LOSHIN, David. The practitioner's guide to data quality improvement. 1 ed. Burlington: Morgan Kaufmann, 2010. 432p.

NABIL, Ali y GOVARDHAN, A. A Comparison Between Five Models Of Software Engineering. En: International Journal of Computer Science Issues. Septiembre, 2010. vol. 7, no. 5, p. 94-101.

Open Knowledge Foundation, Manual de los datos abiertos. 1 ed. Argentina: Open Knowledge Foundation, 2012. 62 p.

PIPINO, Leo; LEE, Yang y WANG, Richard. Data Quality Assessment. En: Communications of the ACM. Abril, 2002. vol. 45, no. 4, p. 211-218.

RUTHBERG, Zella y POLK, William. Report of the Invitational Workshop on Data Integrity. Gaithersburg: National Institute of Standards and Technology, 1999. Reporte número 500-168.

SEVILLANO, Felipe. Big Data. En: Dialnet. 2015. no. 395, p. 71-86.

TARGIO, Ibrahim, et al. The rise of "big data" on cloud computing: Review and open research issues. En: Sciencedirect. Julio, 2014. vol. 47, pag. 98-115.

VANÍČEK, Jiří. Software and data quality. En: Czech Academy of Agricultural Sciences. Febrero, 2006. vol. 52, no. 3, p. 138-146.

VEREGIN, H. Data quality parameters. En: Geographical information systems. 1999. vol. 1, no. 12, p. 177-189.

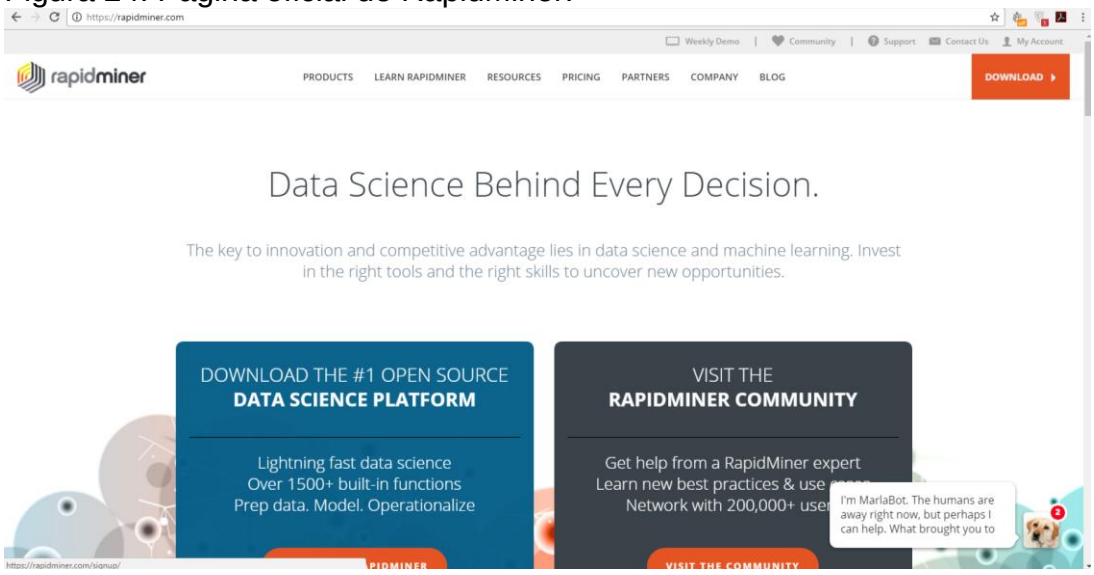
VETRÒ, Antonio, et al. Open Data Quality Measurement Framework: Definition and Application to Open Government Data. En: ScienceDirect. Abril, 2016. vol. 33, no. 2, p. 325-337.

ANEXOS

ANEXO A. MANUAL DE INSTALACIÓN

Las 2 herramientas que se utilizaron para la ejecución del proyecto son RapidMiner y MongoDB, a continuación se presentará la guía de instalación de las 2 herramientas y su integración para el funcionamiento.
En primer lugar se realizará la instalación de RapidMiner, para ello se ingresa a la página oficial www.rapidminer.com.

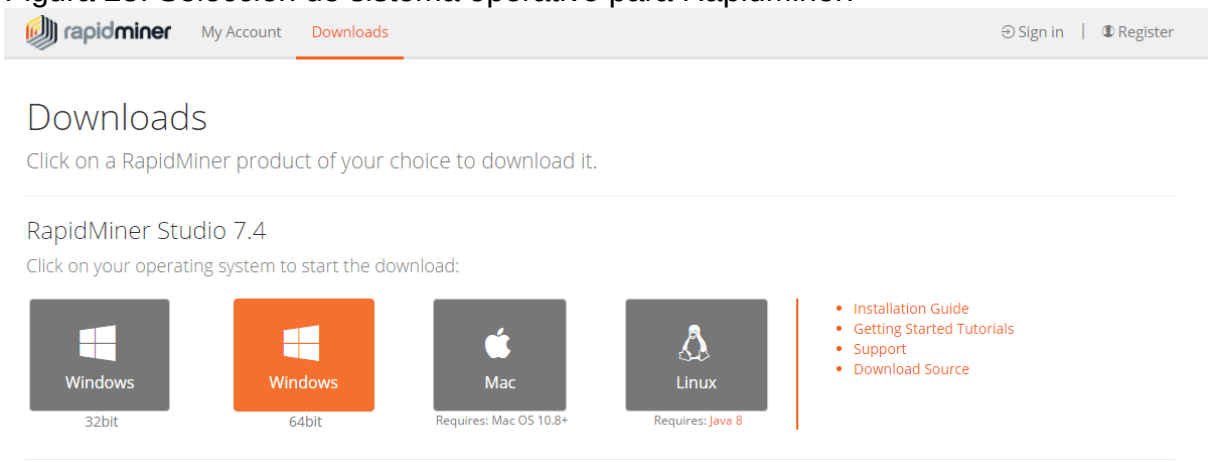
Figura 24. Página oficial de RapidMiner.



Fuente: Los autores.

A continuación se dirige a descargas y selecciona el sistema operativo con el que trabaja.

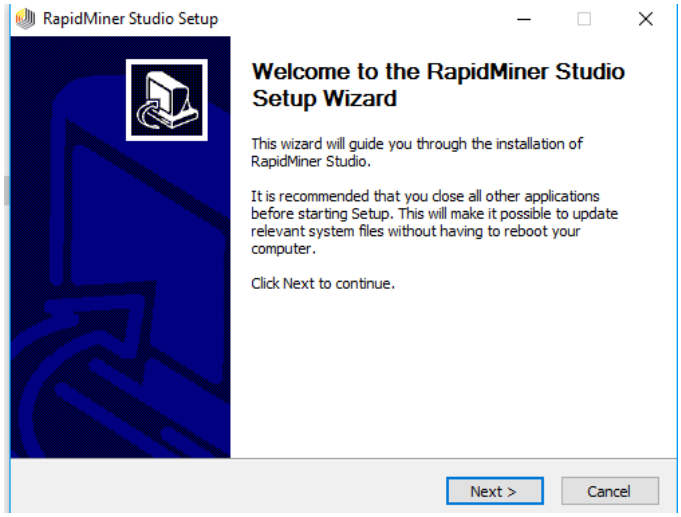
Figura 25. Selección de sistema operativo para RapidMiner.



Fuente: Los autores.

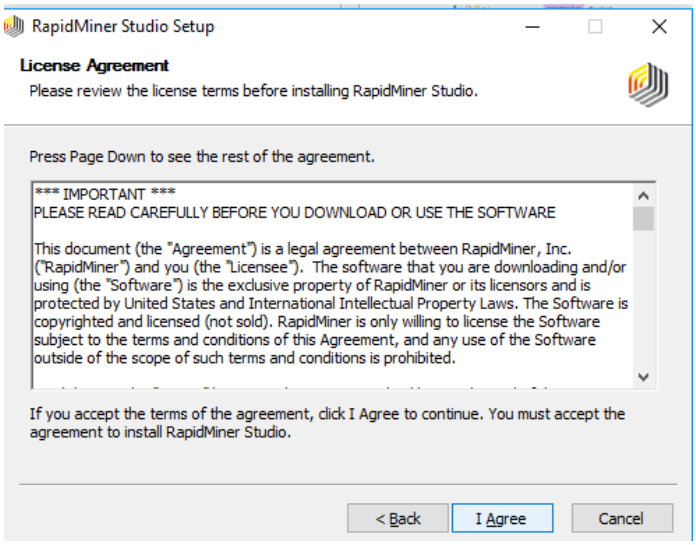
Luego que se descarga el instalador se procede a ejecutar. A continuación se presentan las ventanas de instalación.

Figura 26. Instalación de RapidMiner.



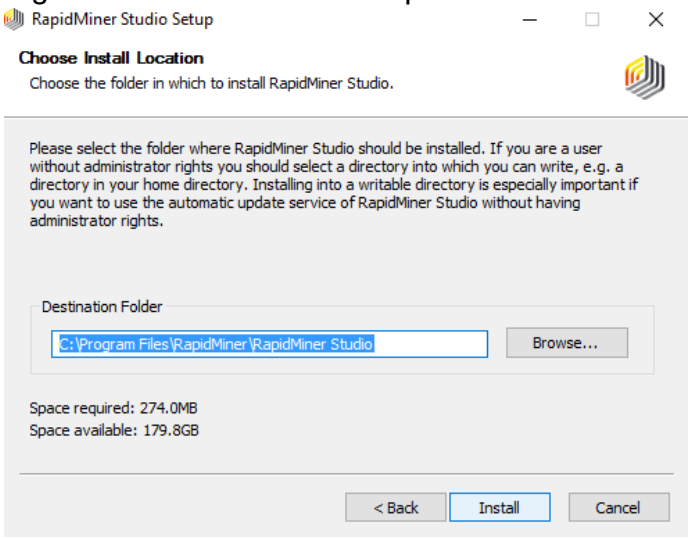
Fuente: Los autores.

Figura 27. Instalación de RapidMiner.



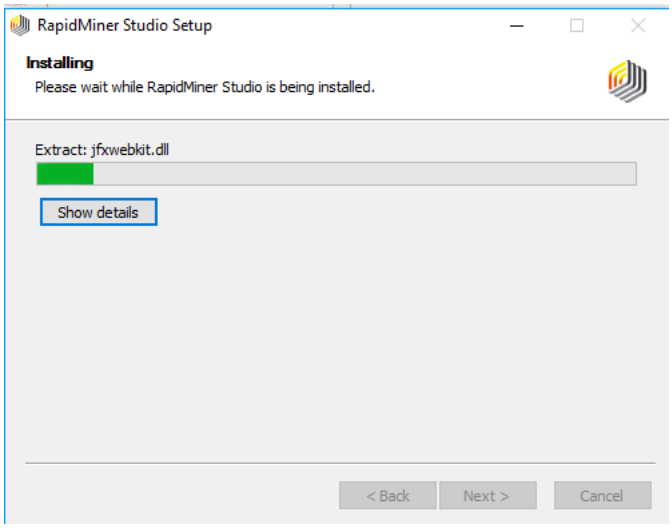
Fuente: Los autores.

Figura 28. Instalación de RapidMiner.



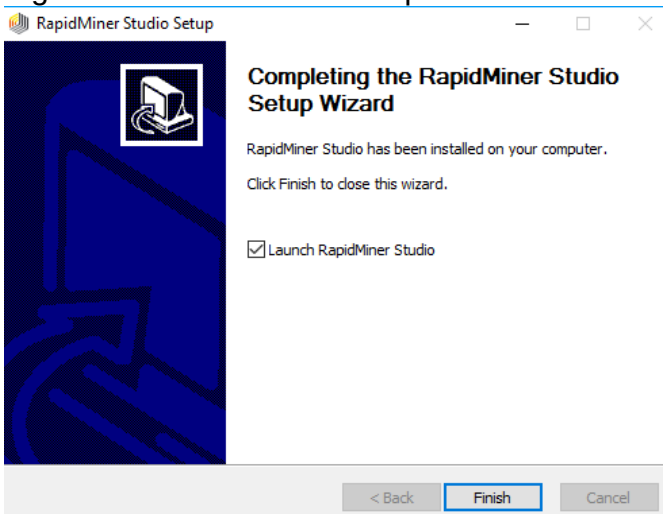
Fuente: Los autores.

Figura 29. Instalación de RapidMiner.



Fuente: Los autores.

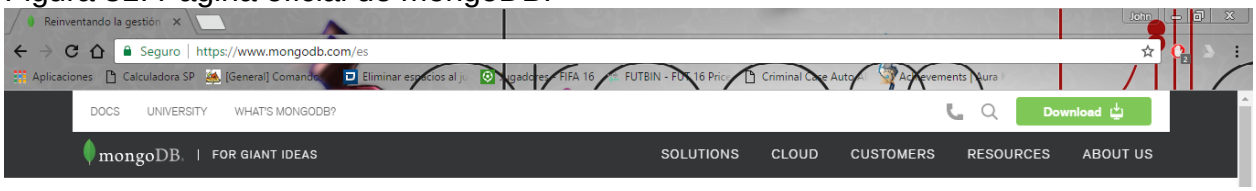
Figura 30. Instalación de RapidMiner.



Fuente: Los autores.

Con ello queda instalada la primera aplicación. Para instalar MongoDB se dirige a la página oficial: www.mongodb.com/es

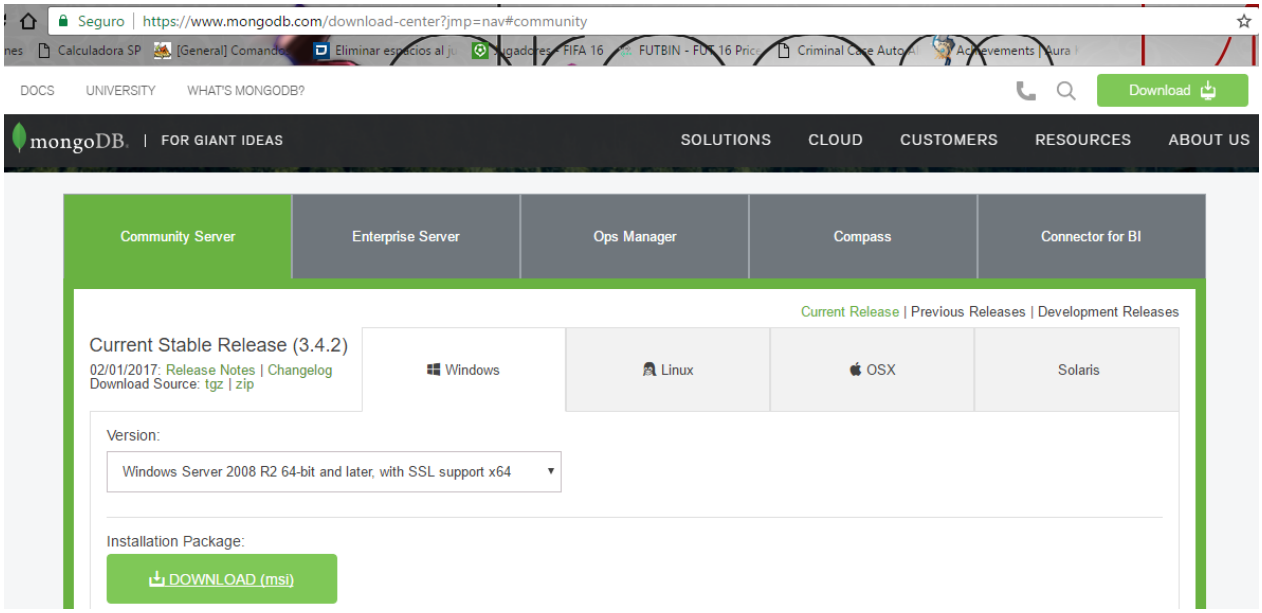
Figura 31. Página oficial de mongoDB.



Fuente: Los autores.

Se dirige a descargas y selecciona el sistema operativo con el que trabaja.

Figura 32. Selección de sistema operativo para mongoDB.



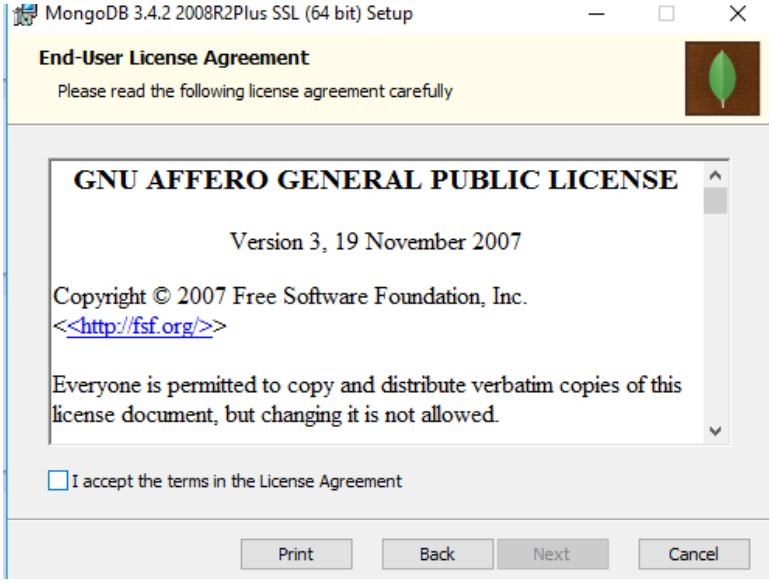
Fuente: Los autores.

Una vez descargado el archivo se procede a ejecutar.

Figura 33. Instalación de mongoDB

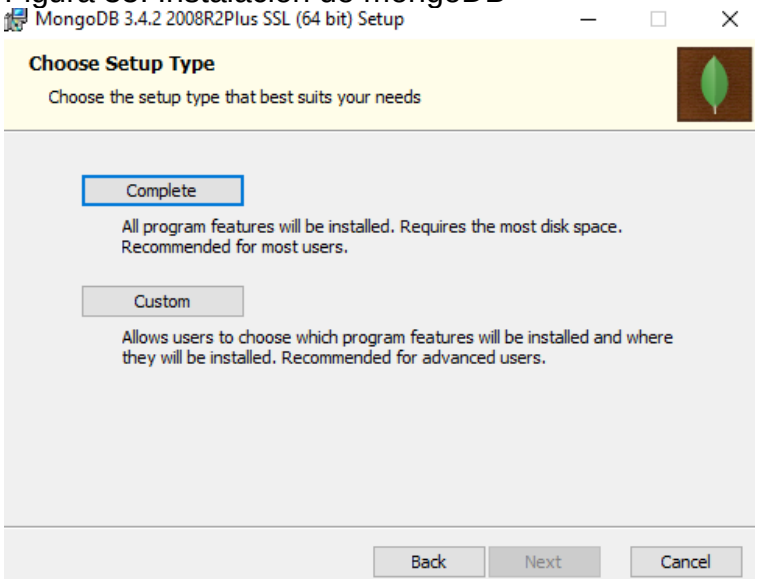
Fuente: Los autores.

Figura 34. Instalación de mongoDB.



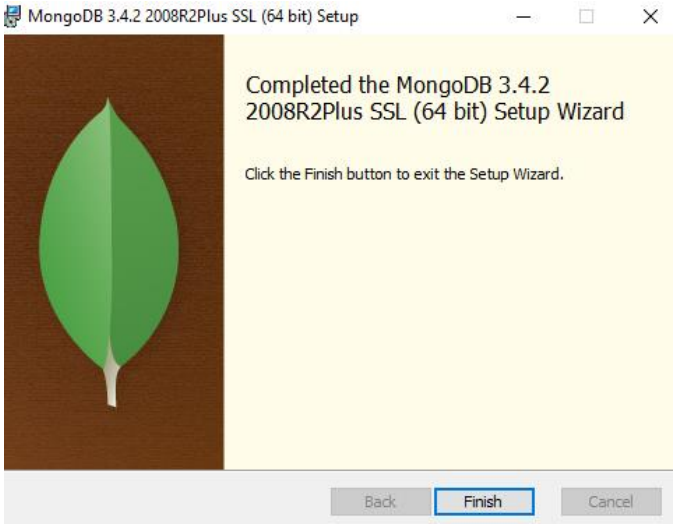
Fuente: Los autores.

Figura 35. Instalación de mongoDB



Fuente: Los autores.

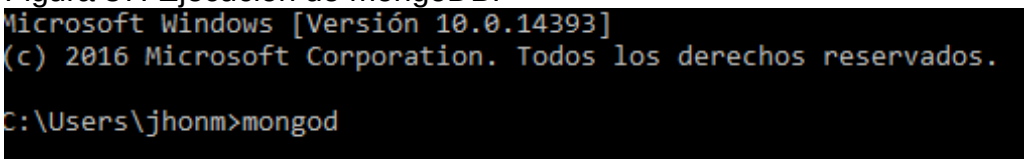
Figura 36. Instalación de mongoDB.



Fuente: Los autores.

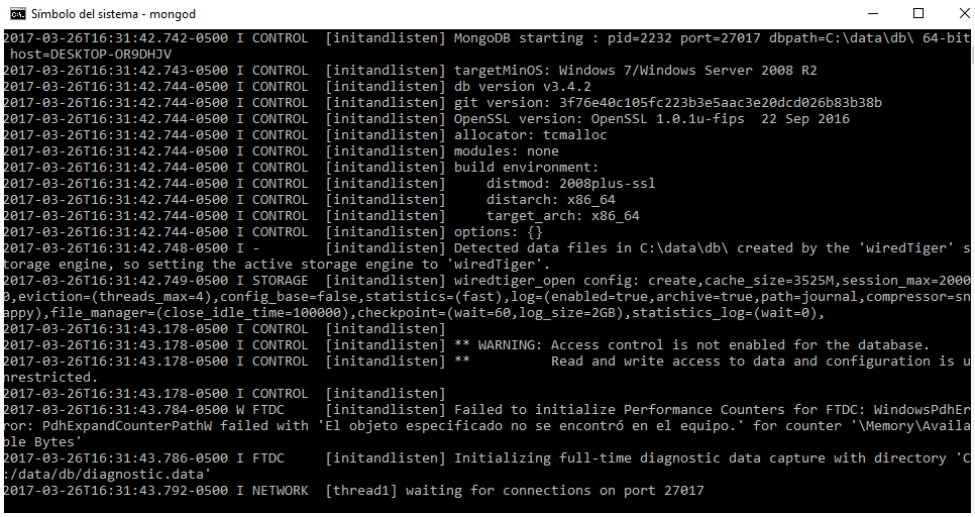
Una vez tenemos instalado MongoDB se abre una consola de comandos, una vez allí se ejecuta el comando “mongod” para iniciar la aplicación.

Figura 37. Ejecución de mongoDB.



Fuente: Los autores.

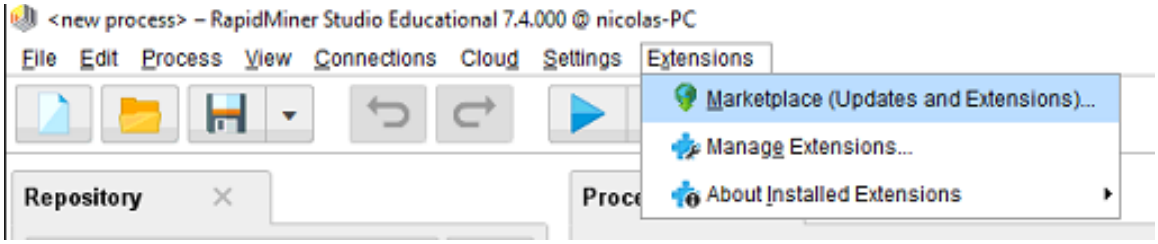
Figura 38. Inicio de mongoDB.



Fuente: Los autores.

El paso final es integrar las 2 aplicaciones. Abrimos RapidMiner y vamos extensiones/Marketplace.

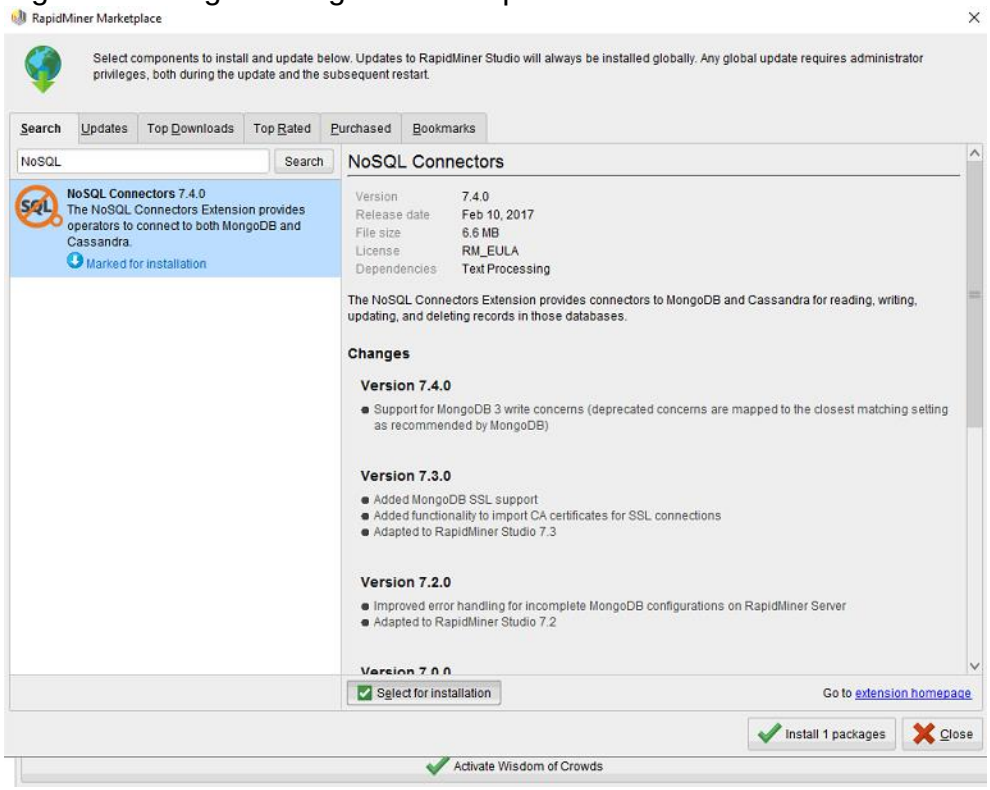
Figura 39. Integración de las 2 herramientas.



Fuente: Los autores.

Esto abrirá una nueva ventana, allí se busca la extensión NoSQL Connectors y la seleccionamos para la instalación.

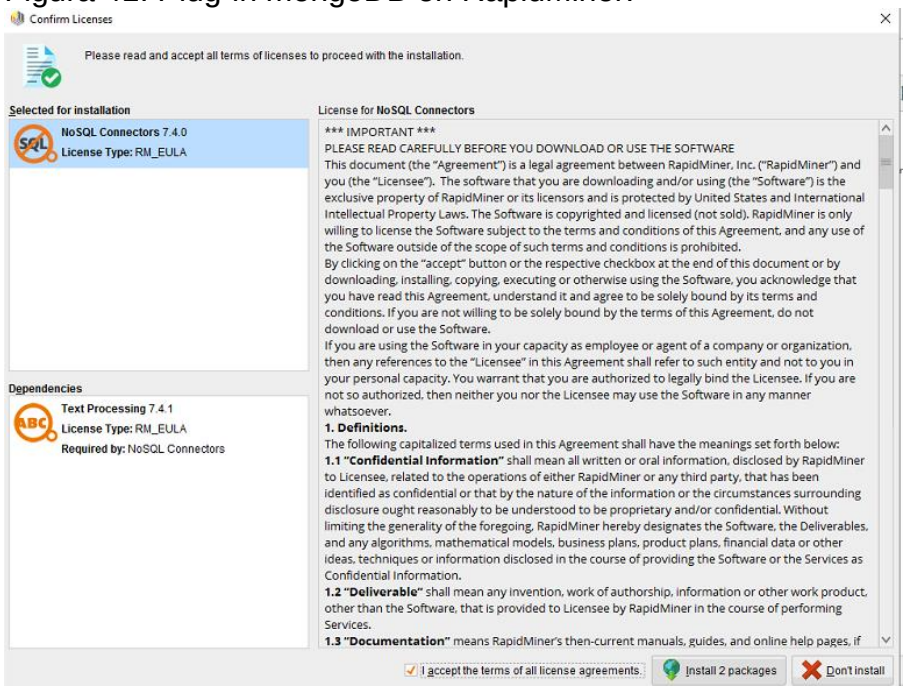
Figura 40. Plug-in mongoDB en RapidMiner.



Fuente: Los autores.

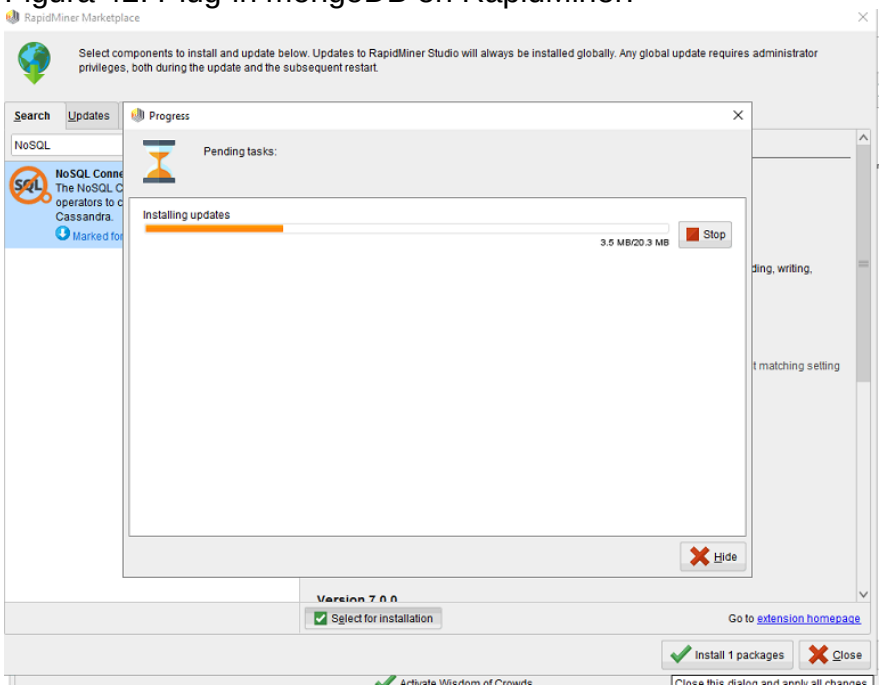
Aceptamos los términos y condiciones y empezará la instalación del complemento.

Figura 41. Plug-in mongoDB en RapidMiner.



Fuente: Los autores.

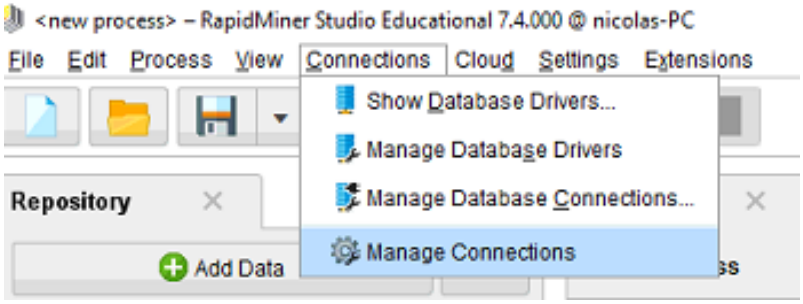
Figura 42. Plug-in mongoDB en RapidMiner.



Fuente: Los autores.

Una vez finalizada la instalación será necesario reiniciar RapidMiner. Luego de ello, vamos a conexiones y entramos a administrar conexiones.

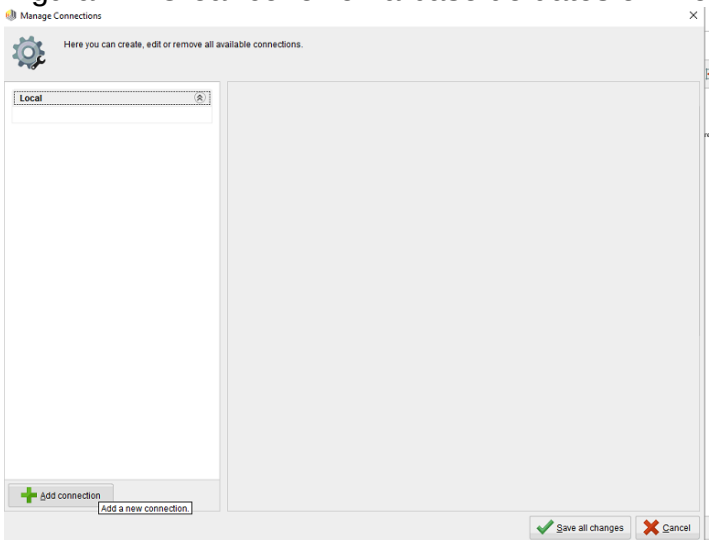
Figura 43. Crear conexión a base de datos en mongoDB.



Fuente: Los autores.

Una vez allí seleccionamos añadir conexión.

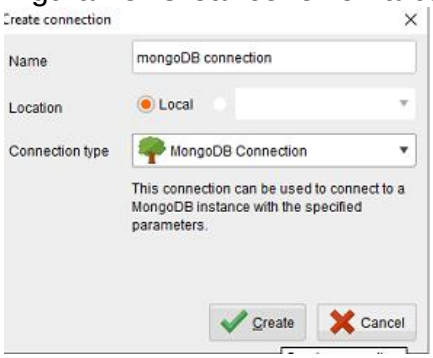
Figura 44. Crear conexión a base de datos en mongoDB.



Fuente: Los autores.

Digitamos los datos requeridos y seleccionamos el tipo de conexión de mongoDB.

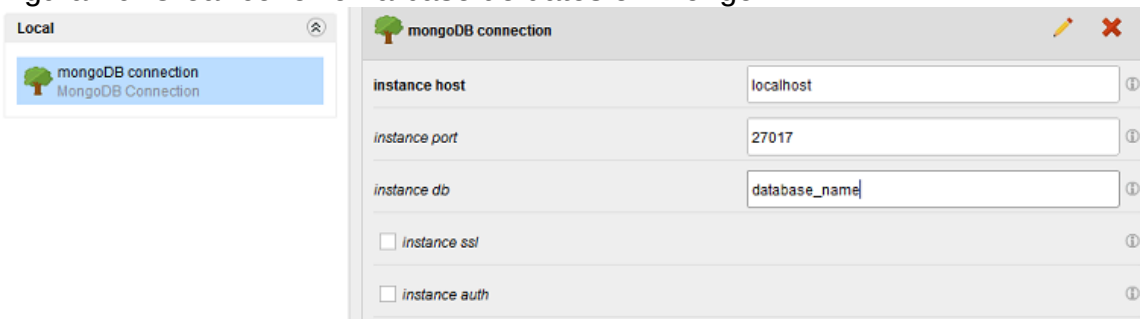
Figura 45. Crear conexión a base de datos en mongoDB.



Fuente: Los autores.

Se crea la conexión y la elegimos. Será requerido el nombre de la base de datos.

Figura 46. Crear conexión a base de datos en mongoDB.



Fuente: Los autores.

Una vez digitados se guardan los cambios y queda integrado MongoDB en RapidMiner.

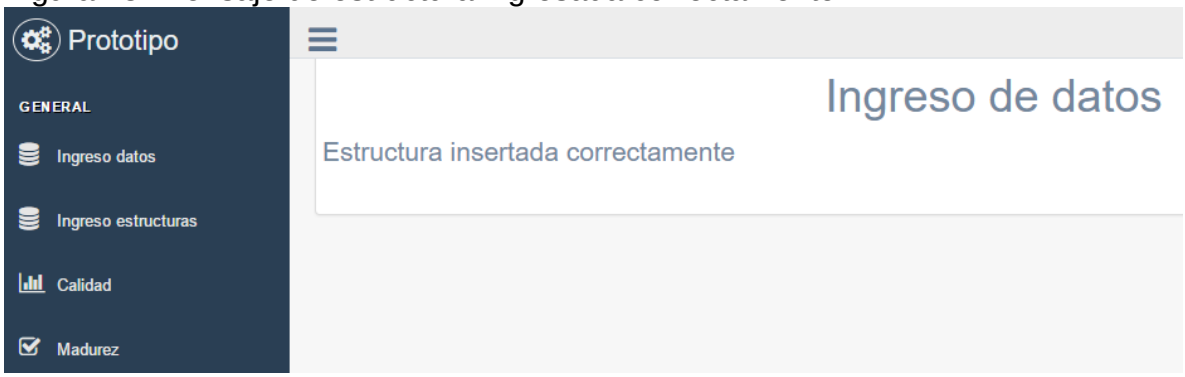
ANEXO B. MANUAL DE USUARIO

El primer módulo del prototipo es el ingreso de estructuras. En este módulo se guarda la estructura del repositorio de datos que ingrese el usuario. El módulo espera dos variables (véase la figura 47), la primera corresponde al ID del conjunto de datos a añadir. Lo que ocurrirá será que la estructura se almacenará y ahora todos los conjuntos de datos que posean esa estructura serán válidos para el prototipo. La segunda variable es la dirección del repositorio donde está el conjunto de datos. La API que maneja la plataforma www.datos.gov.co también es usada en otros países, por lo tanto es válida para cada repositorio que haga uso de la misma. Al hacer clic en enviar aparecerá un recuadro fue añadida exitosamente (véase la figura 48) y otro si ya existe la estructura en el prototipo (véase la figura 49).

Figura 47. Pantalla de inicio de ingresar estructuras.

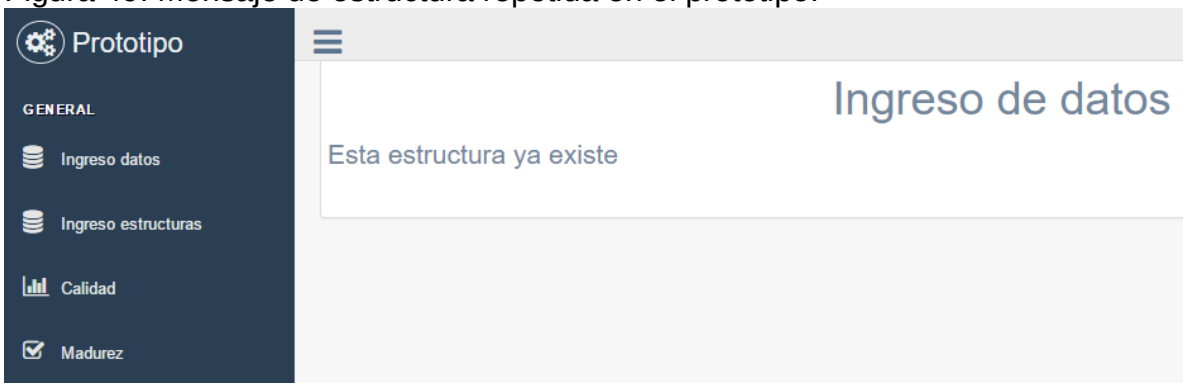
Fuente: Los autores.

Figura 48. Mensaje de estructura ingresada correctamente.



Fuente: Los autores.

Figura 49. Mensaje de estructura repetida en el prototipo.



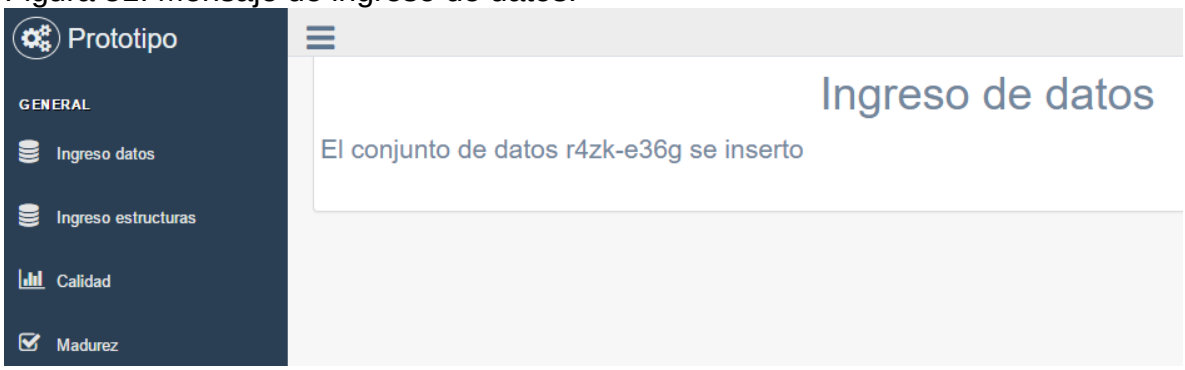
Fuente: Los autores.

En la vista del siguiente módulo (véase la figura 50) permite ingresar todos los datos a los cuales se les calculara las métricas señaladas en el proyecto. Este módulo permite ingresar uno o más conjuntos de datos, los cuales deben ser ingresados separados por comas, por ejemplo: r4zk-e36g,asdf-13k2,opoj-1234. También se ingresa el dominio al que pertenece el conjunto de datos. Si los datos no existen en la base de datos son añadidos y el prototipo informa al usuario que el proceso fue exitoso (véase la figura 51), en el caso que ya exista información de ese repositorio se retorna un mensaje de error informando de ello (véase la figura 52), y si no existe la estructura del conjunto de datos en el prototipo se informa la inexistencia y se muestran las estructuras válidas hasta el momento (véase la figura 53).

Figura 50. Módulo de ingresar datos.

Fuente: Los autores.

Figura 51. Mensaje de ingreso de datos.

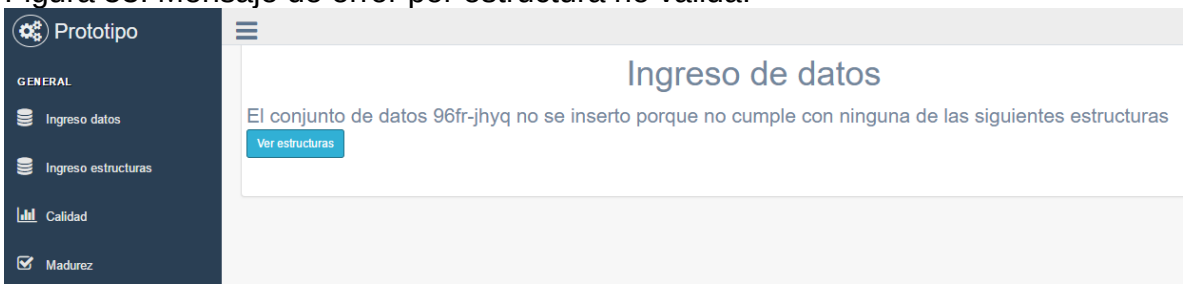


Fuente: Los autores.

Figura 52. Mensaje de error por datos existentes.

Fuente: Los autores.

Figura 53. Mensaje de error por estructura no válida.



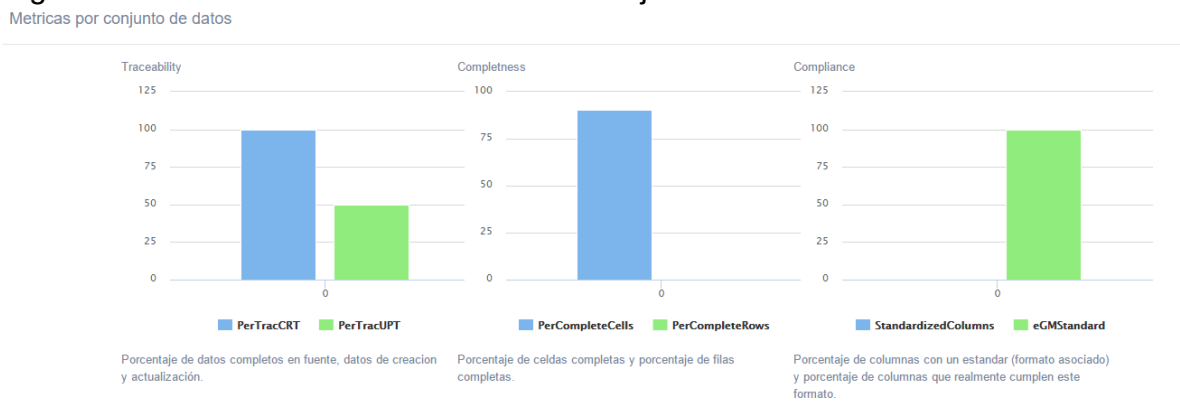
Fuente: Los autores.

En el último modulo se visualiza los gráficos que representan los cálculos de las métricas (véase la figura 54). Los tres que se visualizan al inicio son el cálculo de todos los conjuntos de datos. Si se desea ver el cálculo para un único conjunto de datos ya ingresado se debe hacer clic sobre el ID en la lista ubicada a la izquierda y se desplegara una ventana con el cálculo individual (véase la figura 55).

Figura 54. Módulo de cálculo de las métricas de calidad.

Fuente: Los autores.

Figura 55. Vista del cálculo individual del conjunto de datos.



Fuente: Los autores.