Importancia del *big data* en un gestor documental para las entidades públicas de Colombia*

[Artículos de investigación]

*Wilson Camilo Vargas Guzmán***

*Ana Gabriela Moreno Cadena****

*Angélica Marcela Oñate Escalante*****

*Maritza Sanabria Hivon******

Recibido: 29 de abril de 2020

Revisado: 10 de agosto de 2020

Aceptado: 13 de agosto de 2020

---

* Artículo de resultado de investigación

** Universidad de Educación Superior de Celaya A. C. Doctorando en Administración. Semillero de investigación de la Corporación Universitaria Minuto de Dios. Bogotá, Colombia. Correo electrónico: wvargas@uniminuto.edu.co y arkanotot@gmail.com. ORCID: https://orcid.org/0000-0002-8461-619X. CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000790982

*** Corporación Universitaria Minuto de Dios. Especialista en Gerencia de Proyectos. Semillero de investigación de la Corporación Universitaria Minuto de Dios. Bogotá, Colombia. Correo electrónico: ana.moreno-c@uniminuto.edu.co y gabimoreno.cadena@gmail.com. ORCID: https://orcid.org/0000-0002-0687-1929. CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001814023

**** Corporación Universitaria Minuto de Dios. Especialista en Gerencia de Proyectos. Semillero de investigación de la Corporación Universitaria Minuto de Dios. Bogotá, Colombia. Correo electrónico: angelica.onate@uniminuto.edu.co y am_onate@hotmail.com. ORCID: https://orcid.org/0000-0001-5423-9146. CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001813823

***** Corporación Universitaria Minuto de Dios. Especialista en Gerencia de Proyectos. Semillero de investigación de la Corporación Universitaria Minuto de Dios. Bogotá, Colombia. Correo electrónico: hivon.sanabria@uniminuto.edu.co e hivonmaritza26@gmail.com. ORCID: https://orcid.org/0000-0001-6175-4613. CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001813788

Cómo citar este artículo:

Vargas Guzmán, W. C., Moreno Cadena, A. C., Oñate Escálate, A. M., Sanabria Hivon, M. (2020). Importancia del *big data* en un gestor documental para las entidades públicas de Colombia. *Signos, Investigación en sistemas de gestión, 13*(1). https://doi.org/10.15332/24631140.6345

Resumen

Este artículo se centra en indagar cómo en las entidades públicas los sistemas de gestión de documentos electrónicos de archivo permiten identificar si el *big data* contribuye a satisfacer las necesidades de gestión de información a partir del análisis y la interpretación de las evidencias encontradas. Se utilizó una metodología de tipo exploratorio cualitativo basada en la revisión de literatura sobre el tema. De acuerdo con los resultados obtenidos, se concluye que en Colombia no se ha desarrollado un proyecto que implemente *big data* en gestores documentales y administre la información que contienen. Se plantea la posibilidad de desarrollar una herramienta para el análisis, recuperación, organización, transformación y almacenamiento ágil y oportuno de información.

Palabras clave: recuperación, información, *big data*, sistema de gestión de documentos electrónicos de archivo, integración de datos.

Importance of big data in an electronic archive document management system for public entities in Colombia

Abstract

This article focuses on investigating how in public entities, archival electronic document management systems make it possible to identify whether big data contributes to meeting information management needs

based on the analysis and interpretation of the evidence found. A qualitative exploratory methodology was used, based on a review of the subject literature. According to the results obtained, it may be concluded that in Colombia there is no project implementing big data in document managers to manage the information contained therein. The possibility of developing a tool for the analysis, recovery, organization, transformation, and agile and opportune storage of information is raised.

Keywords: recovery, information, big data, electronic archive document management system, data integration.

Importância dos macrodados (big data) em um sistema de gerenciamento de documentos para entidades públicas na Colômbia

Resumo

Este artigo tem como foco pesquisar como, nos órgãos públicos, os sistemas de gerenciamento de documentos eletrônicos de arquivo tornam possível identificar se os macrodados contribuem para atender às necessidades de gerenciamento de informações com base na análise e interpretação das provas encontradas. Uma metodologia de tipo exploratório qualitativo foi utilizada a partir da revisão da bibliografia especializada sobre o assunto. De acordo com os resultados obtidos, conclui-se que na Colômbia não há nenhum projeto que implemente macrodados em gerentes de documentação e gerencie as informações neles contidas. Coloca-se a possibilidade de desenvolver uma ferramenta para a análise, recuperação, organização, transformação e armazenamento ágil e oportuno das informações.

Palavras-chave: recuperação, informação, macrodados, sistema de gerenciamento eletrônico de documentos para arquivamento, integração de dados.

Introducción

En Colombia, la tecnología ha evolucionado vertiginosamente y ha impactado la producción documental y el manejo de la información a través de diferentes herramientas, dentro de las cuales se pueden mencionar los sistemas de gestión de documentos electrónicos (SGDE) y los sistemas de gestión de documentos electrónicos de archivo (SGDEA), también conocidos como *gestores documentales*. Teniendo en cuenta el volumen exponencial de la información y el crecimiento de datos generados en las entidades públicas, se puede considerar que los archivos —independientemente de su soporte— son parte fundamental de toda empresa, puesto que almacenan la memoria de la organización. Por consiguiente, las entidades del Estado están obligadas a disponer de su información, así como a desarrollar programas tecnológicos que les permitan recuperarla, identificarla, soportarla, almacenarla y protegerla.

De acuerdo con lo anterior, se identifica la necesidad de administrar adecuadamente grandes volúmenes de datos. Al consultar fuentes primarias, se encontró que el desarrollo de herramientas tecnológicas es una preocupación cotidiana en cada sector de la economía y el desarrollo social. Aunque las tecnologías diseñan programas o herramientas, necesitan trabajo colaborativo para ejecutar este tipo de labores, en este caso, la gestión documental.

En este orden de ideas, se considera necesario que en los gestores documentales se incorpore la tecnología del *big data*, de una forma que permita a estos procesar, identificar, almacenar, analizar y describir enormes cantidades de datos (estructurados, no estructurados y semiestructurados). El objetivo de este estudio fue analizar la importancia del *big data* como una solución para el análisis, la transformación, la recuperación y la conservación de la información a partir de la interpretación de las evidencias encontradas.

Importancia del *big data*

Según Puyol (2015), "el término *big data* fue usado por primera vez en un artículo de los investigadores de la NASA Michael Cox y David Ellsworth. Ambos afirmaron que el ritmo de crecimiento de los datos empezaba a ser un problema para los sistemas informáticos actuales. Esto se denominó el problema del *big data*" (p. 1).

Según el Documento Conpes 3920, Política Nacional de Explotación de Datos (*big data*), en Colombia "desde hace aproximadamente veinte años se identificó la necesidad de emplear las Tecnologías de la Información y las Comunicaciones para aumentar la eficiencia en el desarrollo de los procesos y la gestión gubernamental" (Departamento Nacional de Planeación, 2018, p. 3). Esto se debe a que las reglas para la gestión de documentos electrónicos y la conformación de expedientes fueron definidas recientemente por el Archivo General de la Nación mediante el Acuerdo 003 de 2015.

A principios de la década del 2000, la inmensa cantidad de datos generados por diversas fuentes, principalmente derivados de la masificación de los dispositivos móviles, hicieron que los medios tradicionales de almacenamiento y procesamiento creados en los años 80 para la recolección selectiva de datos estructurados resultaran insuficientes. Por esto, la primera definición que varios autores otorgaron al término *big data* fue "conjunto de datos cuyo tamaño va más allá de la capacidad de captura, almacenado, gestión y análisis de las herramientas de base de datos" (López, 2012, p. 3).

En 2005 estuvo disponible al público el primer *software* diseñado específicamente para atender los retos de almacenar y explotar los datos digitales que se encontraban en audios, videos y textos no estructurados. La aparición de nuevas formas de almacenamiento, procesamiento,

análisis y visualización permitió superar el reto tecnológico que dio origen al término *big data*. Según el Departamento Nacional de Planeación (2018):

> Ahora, el reto consiste en definir las condiciones para aprovechar los datos como insumo central de la economía digital, impulsado desde el sector público, así como mitigar los riesgos que puedan derivarse de la explotación de datos, para garantizar la protección de los ciudadanos en este contexto. Por lo expuesto, actualmente la explotación de datos corresponde a la generación de valor social y económico mediante el aprovechamiento para la creación de nuevos bienes, servicios, procesos, así como para el mejoramiento de los existentes. (p. 27)

La administración pública en Colombia inició la investigación e implementación del *big data*. El Departamento Administrativo Nacional de Estadística, por ejemplo, adoptó una política interna para promover proyectos que integren *big data* de acuerdo con los avances mundiales. Para empezar, se implementó la estrategia *smart data*, que implica el uso riguroso de nuevas fuentes de información y el acceso a las grandes bases de datos con fines estadísticos, con el objetivo de desarrollar programas que beneficien a los ciudadanos (Perfetti, 2017). En todos los casos, los investigadores coinciden en que la información está disponible para producir bienestar y el ser humano es el único capaz de detectar su potencial y proyectar su uso.

En Colombia, estas herramientas de recuperación y análisis de la información ya se encuentran en el mercado; en la actualidad, existen varias empresas que ofrecen el *big data* para el análisis y procesamiento de datos. El primer paso se dio a comienzos de marzo de 2017 con el lanzamiento del Centro de Excelencia en Big Data y Data Analytics (Alianza CAOBA), el primero en el país de su género. Aunque hoy en día existen diversas empresas que prestan este servicio, no se conoce su

número exacto. Por consiguiente, se puede afirmar que, si bien en el mercado hay algunos sistemas de gestión de documentos y modelos de *big data*, no se identifican registros exactos sobre empresas que fusionen las actividades y permitan un análisis en conjunto para un fin en común como la preservación.

Se puede afirmar que el primer paso para un correcto uso del *big data* en la gestión documental es la centralización e ingesta de información. Un ejemplo es el proyecto IDEZar (López et ál., 2004), con el que se buscó integrar los datos del Ayuntamiento de Zaragoza, España. La idea principal de esta herramienta fue facilitar el acceso y la explotación de información clave para generar una visión homogénea y simplificar nuevos diseños, servicios y aplicaciones. Para esto, lo primero que hicieron sus creadores fue identificar cada característica de los datos que se deben procesar para que todos tuvieran un mismo formato.

En este ejemplo, los datos y las directrices considerados para organizar la información se basaron en cinco componentes: nombres geográficos, unidades administrativas, redes de transporte, identificadores de propiedad y parcelas catastrales. Cada tema fue dividido en diferentes subtemas que ordenaron de manera más adecuada la información. Con los problemas identificados, establecieron como elemento principal un modelo de datos de referencia urbanos acompañado de procesos y herramientas de integración de información y servicios de red para su explotación. La implementación de un proceso de limpieza, alineamiento, aumentación y referencia permitió crear un modelo que procesara correctamente la información para tenerla en un mismo formato.

La experiencia y los hallazgos de este ejemplo se pueden implementar en la parte inicial del modelo aquí propuesto. En primer lugar, se debe hacer una correcta ingesta, integración y clasificación de la información, para luego realizar una búsqueda eficaz dentro del modelo *big data*. El

siguiente paso es usar tecnología *data lake*, que permite trasladar la información procesada a un gran lago para su posterior almacenamiento en la nube; de esta forma, la tecnología *big data* podrá realizar su tarea específica, que es la búsqueda fácil y oportuna de la información.

A través de un mapeo sistemático, el *data lake* permite entender a la perfección el significado e identificar los problemas a enfrentar. Sobre el uso de esta herramienta y su implementación en el diseño y la transformación de datos, Aucancela et ál. (2018) afirman que:

> La gestión de un *data lake* es muy importante para las estrategias de datos empresariales, ya que responden mejor a las realidades de los datos actuales: volúmenes y variedades, mayores expectativas de los usuarios y la rápida globalización de las economías. (p. 58)

Ahora bien, "la importancia del *big data* radica no solo en la facilidad de manejo del volumen de datos (como el nombre *big data* podría hacer presuponer), sino en la variedad de los mismos" (Duque-Jaramillo y Villa-Enciso, 2017, p. 5). Por consiguiente, esta tecnología es vital para el correcto funcionamiento e implementación de la herramienta, ya que además de recuperar y analizar la información, permitirá realizar una correcta gestión documental, donde prime la preservación y el almacenamiento de los documentos que hacen parte de una organización.

Se debe tener en cuenta la importancia que tiene en la actualidad el *big data*. Para sacar más provecho de las características de esta herramienta, es necesario entender que "a pesar de que el término *big data* se asocia principalmente con cantidades de datos exorbitantes, se debe dejar de lado esta percepción, pues *big data* no va dirigido solo a gran tamaño, sino que abarca tanto volumen como variedad de datos y velocidad de acceso y procesamiento" (Hernández-Leal et ál., 2017, p. 3).

De acuerdo con lo anterior, es importante mencionar que *big data* se define como una gran cantidad de datos, sean o no estructurados, que pueden ser recuperados para obtener información. Se trata también de una tendencia que se está imponiendo en el mundo para el análisis de grandes volúmenes de datos que, justamente por su gran volumen, no pueden ser tratados en los *software* habituales y deben ser procesados en una herramienta que los organice en el menor tiempo posible con los resultados esperados.

En Colombia, gracias a los avances tecnológicos y el desarrollo de plataformas digitales, "hoy es posible hacer más de 320 trámites en línea [...] el 78 % de las empresas y el 50 % de los ciudadanos afirman haber interactuado con plataformas del Gobierno en línea" (Minuto30.com, 2013). Por consiguiente, se deduce que el modelo *big data* que se implemente en cualquier entidad beneficiará a todos los interesados, puesto que permitirá recuperar de forma oportuna cualquier tipo de información.

El uso del *big data* en la archivística es muy importante, dado que un sistema de gestión de documentos electrónicos contiene todos los datos y la información de una organización, de acuerdo con su estructura funcional y los servicios que presta al ciudadano, y su función básica siempre será el almacenamiento, procesamiento y recuperación de información (Colmenares, 2016, p. 12).

Ahora bien, el objetivo más importante en las organizaciones es generar cultura frente a la gestión documental y el manejo adecuado de la información, ya que esta es el área que realmente requiere un cambio estratégico. Este proceso de construir un negocio guiado por el procesamiento de la información involucra el desarrollo de habilidades que permitan no solo la captura de la data, sino también el establecimiento de relaciones con los objetivos corporativos propuestos.

Por otro lado, se evidenció que la gestión documental en Colombia supera una optimización de procesos y una correcta administración de la información. Se puede decir que ha contribuido a la transparencia en la gestión pública y privada, la conservación del patrimonio documental y cultural, y la protección del medio ambiente por medio de lineamientos de digitalización y preservación de la información.

Según la Ley 594 de 2000, Ley General de Archivos, la gestión documental está definida como el "conjunto de actividades administrativas y técnicas tendientes a la planificación, manejo y organización de la documentación producida y recibida por las entidades, desde su origen hasta su destino final, con el objeto de facilitar su utilización y conservación" (Congreso de la República, 2000, p. 2).

Se considera que los documentos son de vital importancia para el desarrollo de una entidad, dado que proporcionan la información necesaria para el correcto funcionamiento de su administración, la optimización de los procesos, el desarrollo de sus actividades y el mejoramiento continuo. Por esto, es imperativo que las entidades implementen un gestor documental o sistema de gestión de documentos electrónicos de archivo (SGDEA), que es un *software* encargado de controlar y organizar los documentos electrónicos, tanto las evidencias de la ejecución de las actividades del negocio como los documentos de valor informativo (tabla 1).

Tabla 1. Aspectos y características de un sistema de gestión de documentos electrónicos de archivo

Aspectos que debe incluir un SGDEA	Características de un SGDEA
Creación y captura de contenido y documentos	Debe proporcionar un repositorio seguro para la conservación de los documentos, tanto en producción como en gestión y trámite.

Aspectos que debe incluir un SGDEA	Características de un SGDEA
Indexación, acceso, almacenamiento y recuperación de contenidos y documentos	Debe permitir la gestión de documentos.
Edición y revisión de contenidos y documentos	No debe permitir la modificación del documento una vez culmine la etapa de gestión y se convierta en un documento de archivo.
Procesamiento de imágenes	Solo debe conservar la versión final del documento y no podrá ser modificada.
Flujo de trabajo de documentos y gestión de procesos empresariales (BPM, por su sigla en inglés)	Se prohíbe la eliminación de documentos, excepto en transferencias de un archivo a otro, si se han cumplido con el tiempo establecido en la TRD y si la disposición final es eliminación.
Distribución de documentos	Debe incluir obligatoriamente el cuadro de clasificación documental, que permite la clasificación de los documentos de la entidad según su estructura orgánico-funcional.
Repositorios de documentos	Debe soportar el establecimiento de los criterios de retención y disposición final resultantes de la valoración documental por cada una de las agrupaciones documentales, las cuales son políticas de conservación.

Fuente: elaboración con base en Rangel (2017).

El análisis de la información permite realizar búsquedas retrospectivas y recuperar cada documento cuando se necesita (Corral, 2015), mientras que la recuperación de la información se refiere al proceso para obtener de un fondo documental los documentos relacionados con determinada demanda de información por parte de un usuario (Saiz, 2013). De acuerdo con el artículo 21 de la Ley 594 de 2000, las entidades del Estado deben empezar a organizar e implementar sistemas tecnológicos para un adecuado servicio. Estas entidades empezaron a desarrollar dicha gestión con procesos de organización archivísticos para tener un orden lógico en cada serie documental.

Con la evolución de la era informática, las instituciones, empresas, científicos y personas en general se ven abocados a consultar información de forma reiterativa para solucionar diferentes conflictos en el desarrollo de su profesión o su vida cotidiana. En este contexto, los SGDEA surgen

como medios tecnológicos en los que se puede almacenar información electrónica en forma estructurada y eliminar así el ruido informacional del lenguaje natural, para recuperar la información en el momento en que se requiera, satisfacer la necesidad informacional y disminuir los tiempos de respuesta.

Metodología

El estudio fue de tipo exploratorio cualitativo, en el que se determinaron las características y la situación actual frente a la deseada. Esta investigación partió de una revisión sistemática de la literatura, que es una metodología para identificar las investigaciones y los documentos existentes más relevantes sobre un tema (Beltrán, 2005). Esta metodología permitió un acercamiento a todos los elementos de la investigación como paso previo en la construcción del modelo para la integración de la gestión documental con la tecnología *big data*.

Una revisión sistemática de la literatura es una metodología en la que se identifican, analizan e interpretan las evidencias encontradas con base en una pregunta inicial (Kitchenham y Charters, 2007), que para el presente artículo fue: ¿es importante la implementación del *big data* como solución para el análisis y la recuperación de la información dentro de un SGDEA en las entidades públicas en Colombia? No se debe dejar de lado el mapeo sistemático que, si bien es una metodología muy similar a la ya mencionada, resalta la importancia de utilizar fuentes y estudios primarios en el área a investigar para identificar el objeto sobre un tema (Kitchenham y Charters, 2007).

El objetivo principal de esta metodología es encontrar respuestas y evidencias sobre la información que ya existe, para llegar a una conclusión en la que se menciona si en el mundo se ha desarrollado una herramienta que atienda esta necesidad, contemple estas características y contribuya en

el correcto desarrollo y funcionamiento de la gestión documental en las entidades públicas de Colombia. En la tabla 2 se pueden observar las características de la investigación con la metodología de revisión sistemática.

Tabla 2. Revisiones sistemáticas

Característica	Revisión sistemática
Pregunta de investigación	Estructurada, clara, concreta y centrada en un problema clínico bien definido
Búsqueda bibliográfica y selección de fuentes de información	Búsqueda detallada, sistemática y explícita
Selección de artículos	Selección basada en criterios explícitos. Aplicación uniforme de los criterios de selección/exclusión a todos los artículos
Valoración de la calidad de los estudios	Valoración/evaluación crítica de la calidad metodológica de los estudios
Síntesis	Basada en la calidad metodológica de los estudios. A menudo, resumen cuantificado por un estimador estadístico
Interpretación	Generalmente basada en la evidencia

Fuente: elaboración con base en Martín (2014).

La búsqueda y selección de artículos se realizó en la base de datos Redalyc y en el banco de proyectos de la Universidad Militar Nueva Granada, a partir de términos y frases cortas como, por ejemplo, “gestor documental”, “*big data* en la gestión documental en Colombia” y “gestión documental en las entidades públicas de Colombia”. Luego, se revisaron y seleccionaron los artículos más relevantes como referencias para el desarrollo del tema. Por otro lado, se realizaron consultas en las páginas de entidades públicas de Colombia relacionadas con los lineamientos en materia de gestión documental para evidenciar si existía un desarrollo o lineamiento al respecto de la implementación del *big data* en un gestor documental.

Resultados y discusión

A partir de la exploración que se realizó en los procesos de gestión documental, informativos y administrativos de algunas entidades públicas en Colombia, se evidenció que no se ha desarrollado un proyecto similar que implemente *big data* en un gestor documental y, a su vez, contribuya con el análisis, la recuperación, el almacenamiento y la preservación de la información localizada en ellos. Por otra parte, entre los resultados de la investigación se identificó que puede desarrollarse una herramienta *big data* (figura 1) y determinar los aspectos técnicos más importantes que se deben tener en cuenta para su diseño:

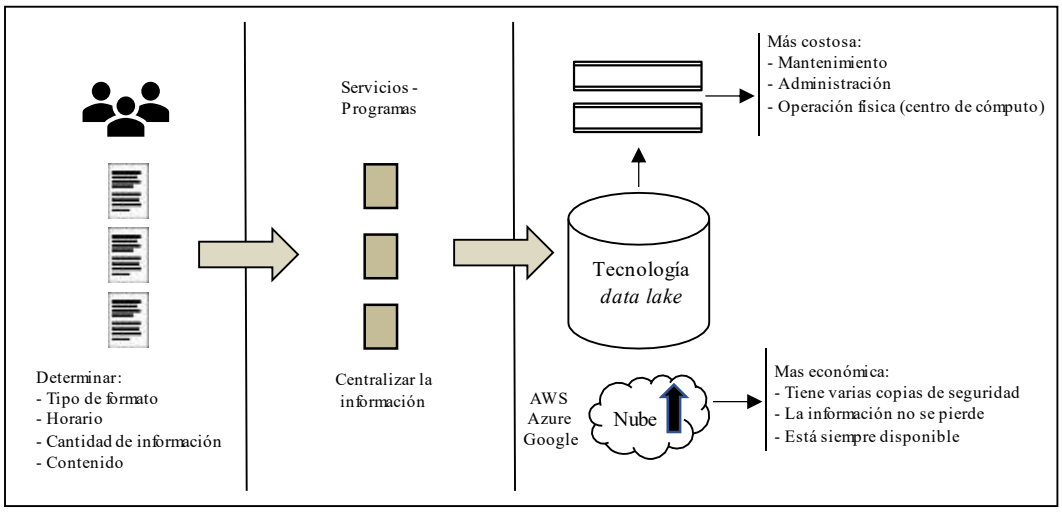
> Dentro del sector de tecnologías de la información y la comunicación, *big data* es una referencia a los sistemas que manipulan grandes conjuntos de datos. Las dificultades más habituales en estos casos se centran en la captura, almacenamiento, búsqueda, compartición, análisis y visualización. (Pérez, 2015, p. 4)

Primero, se debe determinar la fuente (formatos) de cada entidad, así como la tecnología para capturar y centralizar la información. Luego, se convierten los archivos a un formato binario llamado *BSON*, usado en el almacenamiento y transferencias de bases de datos. En este caso, se debe contratar un servidor para hacer consultas masivas y análisis de los datos de la información de la entidad. Data WareHouse es la tecnología que se puede emplear, pues permite almacenar los datos de forma jerárquica por medio de carpetas que facilitan una búsqueda rápida y eficiente a la hora de consultar la información.

Es importante tener en cuenta que con este modelo se busca analizar y recuperar la información, con el fin de organizarla, transformarla y almacenarla de manera ágil y oportuna, mejorando los tiempos de

respuesta de acuerdo con las consultas. Para ello, proponemos la tecnología *data lake*.

Figura 1. Flujo de los datos en las herramientas



Fuente: elaboración propia.

Un *data lake* es un repositorio de datos de bajo costo que permite almacenar datos estructurados, no estructurados y semiestructurados. La tecnología que permite la implementación de un *data lake* es Hadoop, una plataforma de *software* que permite escribir con facilidad y ejecutar aplicaciones que procesan ingentes cantidades de datos, lo que obliga a los analistas de datos a investigar su implementación (Aucancela et ál., 2018). Un *data lake* es un enorme repositorio de datos en bruto, que contiene datos multiestructurados para realizar los análisis necesarios, incluso antes de que se definan. Su propósito es tener una plataforma donde se pueden realizar rutinas de preparación de datos y creación de perfiles en el sistema. Funciona como un almacenamiento de datos operacional, que facilita el acceso a la información antes de su almacenamiento.

Según Aucancela et ál. (2018), la arquitectura de un *data lake* es "plana centrada en los datos, que almacena grandes volúmenes de datos en varios formatos, es decir, datos sin procesar [...] Básicamente, en un *data lake* los

datos ingresan por procesamiento en lotes o procesamiento en tiempo real. Cada entidad de datos en una laguna está asociada con un identificador único y un conjunto de metadatos extendidos" (p. 55). Estos forman un catálogo de datos que es utilizado por los consumidores (científicos de datos, analistas de negocios) para crear esquemas específicos de acuerdo con sus necesidades. En este caso, el *data lake* ejerce un rol de proveedor, es decir, proporciona datos y análisis como servicio (DaaS).

Un *data lake* tiene como función la ingesta de información basada en una herramienta que permita hacerlo antes de su almacenamiento definitivo (gobernanza de los datos). En este punto se clasifica la información y se eliminan los elementos que no sirvan para la entidad; así, se pueden establecer las reglas necesarias para identificar y clasificar la información de manera correcta según los filtros determinados. Esta herramienta permite retener y almacenar la información de manera temporal para poder hacer uso de ella, de forma que el usuario o la entidad puedan tomar decisiones y procesar la información en un formato estandarizado.

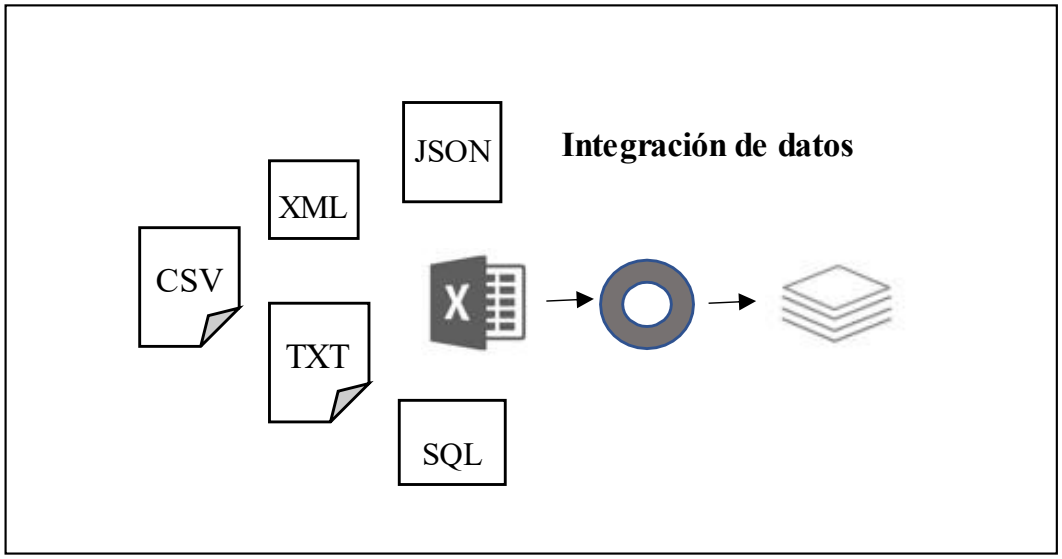
El acceso a la información dentro del *data lake* se puede realizar por medio de consultas o extracciones. Es importante contar con usuarios experimentados en el manejo de la herramienta y parámetros de seguridad que permitan un manejo adecuado, para no poner en riesgo la información tratada.

Centralizadores e ingesta de la información

También conocida como integración de datos, la estrategia ETL (*extract, transform, load*) permite la extracción de uno o varios sistemas de fuentes, su transformación (reformateo y limpieza) y su carga (al *data lake* u otra tecnología escogida). Las herramientas para lograr su integración suelen tener una serie de desafíos (figura 2) que requieren la definición de parámetros. Dado que se encontrará información dispersa e incompatible,

las fases de diseño son importantes a la hora de su implementación (Rayón, 2016).

Figura 2. Integración de datos



Fuente: IntelDig (2018).

Almacenamiento en la nube

Después de transformar la información y usar la tecnología *data lake*, se necesita un espacio virtual que permita almacenarla. Esto se puede solucionar mediante el almacenamiento en la nube, "que permite acceder los documentos desde cualquier lugar a través de la red" (Morales, 2014, p. 275).

El almacenamiento en la nube es una forma económica y de fácil acceso, mantenimiento y manejo. Es un nuevo modelo tecnológico que consiste en almacenar muchos datos en varios servidores virtuales que son administrados por terceros. Estos proveedores gestionan y operan grandes centros de datos y los usuarios finales compran o alquilan espacio dentro de sus servidores según sus necesidades. Los operadores de dichos centros virtualizan los recursos de acuerdo con los requerimientos de los usuarios

para que, a su vez, ellos puedan utilizar el servicio para el almacenamiento de sus datos (Goyas y Vargas, 2014).

La principal ventaja del almacenamiento en la nube es su disponibilidad desde cualquier medio con acceso a la red. Por ejemplo, con una conexión a internet se puede tener acceso a los archivos para descargarlos o modificarlos. No se requiere de un dispositivo físico para el almacenamiento y transporte de la información y esta se aloja en un sitio remoto (Díaz y Cleves, 2014). Por otro lado, la principal desventaja de este sistema es la seguridad de los datos almacenados, ya que un servicio público no puede garantizar la seguridad de la información y, si no se cuenta con una conexión veloz, el uso de la información en la nube puede ser difícil.

Conclusiones

Esta investigación permitió analizar el avance y los aportes de las tecnologías de la información, sobre todo en lo concerniente a su uso, aprovechamiento y evolución. Se analizaron diferentes fuentes de información con temáticas sobre el mejoramiento de la gestión documental en entidades públicas, que pueden servir para el desarrollo de trabajos posteriores en el área. Se evidenció que la tecnología está abarcando todos los ambientes del ser humano y se ha convertido en parte fundamental para el desarrollo de sus actividades.

Las técnicas del *big data* son una nueva oportunidad de negocio para quienes buscan generar oportunidades, retos y nuevas propuestas en esta área. El *big data* permite el almacenamiento de grandes cantidades de datos con cinco criterios: volumen, variedad, velocidad, veracidad y valor, que optimizan los procesos, garantizan la veracidad de la información y facilitan el acceso a ella.

El primer paso para un correcto uso del *big data* en la gestión documental es la centralización e ingesta de información. La integración de datos y la fusión o implementación de estas herramientas son consideradas totalmente viables en Colombia, como una combinación de tecnologías que debe ser desarrollada por un grupo de colaboradores expertos que hagan de esta una herramienta innovadora para responder a las necesidades descritas en este documento.

La organización de la información mediante estas tecnologías permite un manejo local para brindar información clave, clara y oportuna a la ciudadanía, en un tiempo de respuesta remoto, lo que demuestra que estas tecnologías traen grandes beneficios a la hora de conservar la información (Beltrán, 2005). En Colombia, la gestión documental se desarrolla en todas las actividades donde se requiere el manejo de documentación con valor importante. El *big data* puede ser la solución en los sistemas de gestión electrónica de archivo (SGDEA) de entidades públicas al aumentar la eficiencia en el manejo de la información y las comunicaciones.

Referencias

Aucancela, M., Naranjo, J. y Betún J. (2018). Mapeo sistemático de literatura de un *data lake*. *Revista mktDescubre*, *11*, 55-66. http://revistas.espoch.edu.ec/index.php/mktdescubre/article/download/153/158/

Beltrán, O. (2005). Revisiones sistemáticas de la literatura. *Revista Colombiana de Gastroenterología*, *20*(1), 60-68. http://www.scielo.org.co/pdf/rcg/v20n1/v20n1a09.pdf

Colmenares, J. F. (2016). *Implementación de big data en las organizaciones como estrategia de aprovechamiento de la información para incorporara a la cadena de valor del negocio* [tesis de grado]. Universidad Militar Nueva Granada.

Congreso de la República de Colombia. (2000, 14 de julio). Ley 594. *Por medio de la cual se dicta la Ley General de Archivos y se dictan otras disposiciones*. Diario Oficial 44093. http://www.secretariasenado.gov.co/senado/basedoc/ley_0594_2000.html

Corral, A. M. (2015, 2 de marzo). *¿Qué es el análisis documental?* DOKUTEKANA. https://archivisticafacil.com/2015/03/02/que-es-el-analisis-documental/

Departamento Nacional de Planeación. (2018, 17 de abril). *Documento Conpes 3920. Política Nacional de Explotación de Datos (big data)*. https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3920.pdf

Diaz, R. y Cleves, J. (2014). *Almacenamiento en la nube* [tesis de grado]. Universidad Piloto de Colombia. http://repository.unipiloto.edu.co/handle/20.500.12277/2969

Duque-Jaramillo, J. y Villa-Enciso, E. (2017). *Big data:* desarrollo, avance y aplicación en las organizaciones de la era de la información. *Revista CEA*, *2*(4), 27-45. https://doi.org/10.22430/24223182.169

Goyas, M. y Vargas, J. (2014). *Almacenamiento en la nube* [tesis de grado]. Escuela Superior Politécnica del Litoral. https://www.dspace.espol.edu.ec/retrieve/102294/D-84369.pdf

Hernández-Leal, E., Duque-Méndez, N. y Moreno-Cadavid, J. (2017). *Big data*: una exploración de investigaciones, tecnologías y casos de aplicación. *TecnoLógicas*, *20*(39). http://hdl.handle.net/20.500.12622/1020

IntelDig. (2018). *Integración de datos: problemas y técnicas de integración*. https://www.tecnologias-informacion.com/integracion.html

Kitchenham, B. A. y Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*. *Version 2.3*. *EBSE Technical Report*. *EBSE-2007-1*. https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf

López, D. (2012). *Análisis de las posibilidades de uso de big data en las organizaciones* [tesis de grado]. Universidad de Cantabria. https://repositorio.unican.es/xmlui/bitstream/handle/10902/4528/TFM%20-%20David%20L%C3%B3pez%20Garc%C3%ADaS.pdf?sequence=1

López, F. J., Álvarez, P. y Muro-Medrano, P. R. (2004). *IDEZar: Procesos, herramientas y modelos urbanos aplicados a la integración de datos municipales procedentes de fuentes heterogéneas.* Universidad de Zaragoza. https://www.idee.es/resources/presentaciones/JIDEE06/ARTICULOS_JIDEE2006/articulo25.pdf

Martín, R. H. (2014). *La búsqueda bibliográfica, pilar fundamental de la medicina basada en la evidencia: evaluación multivariante de las enfermedades nutricionales y metabólicas* [tesis de doctorado]. Universidad Miguel Hernández. http://dspace.umh.es/bitstream/11000/1639/1/Tesis_Helena_VFI.pdf

Minuto30.com. (2013, 16 de agosto). *Big data, una herramienta para tomar decisiones empresariales y políticas.* https://www.minuto30.com/ciencia-tecnologia/big-data-una-herramienta-para-tomar-decisiones-empresariales-y-politicas/175046/

Morales, M. A. (2014). ¿Cuán efectivo es el almacenamiento en la nube? *Revista APEC*, *30*, 264-276. https://issuu.com/apecpr/docs/revista-apec-volumen-30-2014_99390bd5623acf

Pérez, M. (2015). *Big data. Técnicas, herramientas, y aplicaciones*. Alfaomega Grupo Editor.

Perfetti, M. (2017, 8 de noviembre). *IV Conferencia Global de big data para las estadísticas oficiales* [video]. https://www.dane.gov.co/files/noticias/mensaje_director_bigdata_2017.mp4

Puyol, J. (2015). *Aproximación jurídica y económica del big data*. Tirant lo Blanch.

Rangel P. E. (2017). *Guía de implementación de un sistema de gestión de documentos electrónicos de archivo - SGDEA*. Archivo General de la Nación https://www.archivogeneral.gov.co/sites/default/files/Estructura_Web/5_Consulte/Recursos/Publicacionees/ImplementacionSGDEA.pdf

Rayón, A. (2016, 18 de diciembre). Tecnologías de ingesta de datos en proyectos "*big data*" en tiempo real [blog]. Deusto Data. https://blogs.deusto.es/bigdata/tecnologias-de-ingesta-de-datos-en-proyectos-big-data/

Saiz, J. (2013). *Recuperación documental. Archivo de empresa*. https://archivoempresa.wordpress.com/recuperacion-documental/