OPEN DATA Y *BIG DATA*: HERRAMIENTAS DE *SOFTWARE* PARA CIUDADES INTELIGENTES (CASO DE ESTUDIO)

5

Este caso de estudio forma parte del proyecto de investigación Derecho y Big Data, en el que participan dos grupos de investigación de la Universidad Católica de Colombia: a) Investigación en Derecho Público y TIC, perteneciente a la Facultad de Derecho; b) Software Inteligente y Convergencia Tecnológica (GISIC), adscrito a la Facultad de Ingeniería. Además, cuenta con la participación de dos colaboradores externos: la Universidad de Texas, en Estados Unidos, y el Departamento Nacional de Planeación (DNP); esto último, a partir de un convenio celebrado entre la Universidad y la entidad del Estado.

Las ciudades inteligentes o *smart cities* buscan optimizar sus procesos con el objetivo de mejorar la calidad de vida de sus ciudadanos, mediante la recopilación de datos y utilizando dispositivos de *software* y *hardware* —por ejemplo, cámaras de video, sensores y *smartphones*—. De esta manera, el Gobierno puede identificar los posibles problemas que se encuentran en alguna sociedad en particular e incluso adelantarse a situaciones que se pueden convertir en caos (Anaya, 2017).

El término *big data*, o 'datos masivos', indica grandes volúmenes de datos, de toda variedad, que se procesan a grandes velocidades con el fin de lograr un potencial de valor incalculable sobre ellos. La cantidad de datos generados pueden provenir del sector público o privado; es decir, conjuntos de datos pueden

convertirse en *big data* independientemente del sector macroeconómico, lo cual es una gran ventaja por cuanto el término es transversal a cualquier empresa y sociedad. Por su parte, *open data*, o 'datos abiertos', abarca todos los datos generados por entidades gubernamentales con el fin de ser utilizados, reutilizados y redistribuidos libremente por cualquier persona natural o jurídica, independientemente del propósito o beneficio que se tenga destinado (Vetrò *et al.*, 2016). *Open data* también engloba los datos que se almacenen en medios físicos —como los documentos impresos— y que deban ser mantenidos en esta forma por normatividad de la empresa o del Estado.

Aporte de *big data* y *open data* a ciudades inteligentes

Open data nace a partir de la ejecución de los procesos operativos y estratégicos en las entidades gubernamentales. El área de tecnología es normalmente la encargada de consolidar los datos que se van a compartir al público, ya que por su naturaleza misional debe gestionar datos para la elaboración de conjuntos de datos que son compartidos a diferentes instancias públicas y privadas. En este caso, debe publicar y compartir los datos en algún repositorio en internet destinado a datos abiertos.

Las áreas de tecnología deben tener la habilidad de identificar cuándo *open data* se torna *big data*, pues gestionar grandes volúmenes de información en varias empresas es engorroso para la elaboración de informes: el procesamiento tarda mucho tiempo y el almacenamiento es inmanejable, ya que alcanza los límites de los servidores. Por tal motivo, es crucial identificar mediante métricas cuándo el conjunto de datos forma parte de *big data* con el fin de garantizar una gestión adecuada de los datos. A continuación se presentan algunas características que ayudarán a las entidades gubernamentales a identificar cuándo *open data* se convierte en *big data*:

- Hasta 10 GB de datos en cualquier formato de archivo (p. e., Excel, Word, Power Point, etc.) se definen como *small data*.
- Entre 10 GB y 1024 GB (1 terabyte) de datos, normalmente almacenados en un gestor de datos, se les conoce como *base de datos*; por ejemplo, MySQL, SQL, ORACLE, entre otros.

- De 1 terabyte de datos en adelante se define como *big data*; para este tipo de volúmenes de datos normalmente se requieren bases de datos distribuidas y con gran capacidad de almacenamiento en los servidores.

Contexto en Colombia

En Colombia actualmente existen grandes volúmenes de información generados a partir de múltiples fuentes de datos y que normalmente evidencian una baja calidad en su contenido. Por ejemplo, una fuente de datos es el repositorio nacional de información *www.datos.gov.co*, utilizado por las entidades públicas para compartir en internet información financiera, contractual, logística y administrativa, entre otras, con el fin de dar cumplimiento a la Ley de Transparencia y Acceso a la Información Pública; sin embargo, este portal no suministra los datos con métricas de completitud, coherencia y exactitud, lo cual afecta directamente la calidad de la información, por ejemplo, al limitar la manera de identificar problemas de corrupción y otros.

En el sector privado hay un gran interés por el uso de datos de gobierno para diferentes propósitos; por ejemplo, para realizar seguimiento sobre los procesos contractuales que involucran dinero o para revisar el comportamiento de indicadores de educación. Sin embargo, en Colombia hay desafíos que aún se deben abordar antes de utilizar los datos abiertos. El primero de ellos es el aseguramiento de la calidad de los datos suministrados por las entidades gubernamentales, en términos de completitud, coherencia, oportunidad y consistencia. El segundo desafío es la claridad con la que se debe presentar la información, ya que, debido al volumen que se puede manipular, normalmente se cae en el error de querer mostrar olas de información, lo que, en consecuencia, colige desinformar a los ciudadanos. El tercer desafío es la importancia que deben darle las mismas entidades gubernamentales, ya que el adecuado uso de datos abiertos trae asociados valores agregados como la transparencia y optimización de procesos y recursos del Estado.

El tratamiento de los datos abiertos que se gestiona en las entidades gubernamentales está regido por la "Ley de Transparencia de Datos y del Derecho al Acceso de la Información Pública Nacional", la cual exhorta a publicar y compartir los datos derivados de la gestión pública para que sean alcanzables, utilizados,

reutilizados y redistribuidos por la sociedad colombiana. Para el cumplimiento de la ley, las entidades gubernamentales publican y comparten sus datos en la página *www.datos.gov.co* o en cualquier portal en internet que cumpla los estándares necesarios para la aplicación de la ley. En estos portales se encuentran conjuntos de datos que son clasificados de acuerdo con los procesos derivados de la gestión pública; por ejemplo, datos relacionados con procesos financieros, contables, contractuales, a excepción de información que es confidencial o secreto de Estado.

Contexto internacional

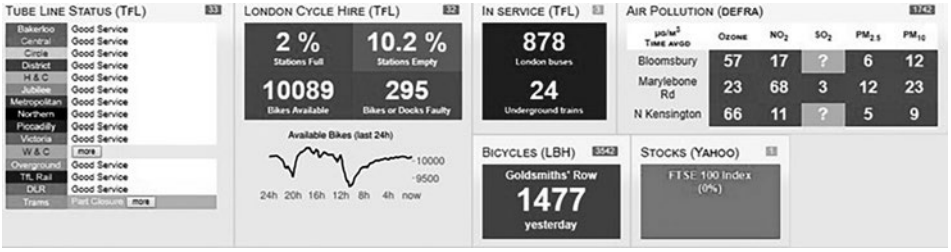
La economía mundial se ha centrado en los datos; en consecuencia, quienes tengan las capacidades de extraer el máximo beneficio de sus datos tendrán el poder en términos políticos, sociales, culturales y, especialmente, económicos. En los últimos años, un número creciente de gobiernos ha comenzado a abrir sus datos. "Este movimiento llamado gobierno abierto ha resultado en el lanzamiento de numerosos portales de datos abiertos y de infraestructuras que tienen como objetivo proporcionar un punto único de acceso a datos del gobierno y explorar sus consecuencias" (Máchová y Lnénicka, 2017, p. 21).

Los reportes en línea y tableros de indicadores en las ciudades miden el progreso de las ciudades utilizando métricas urbanas y mejorando sus estrategias de operación a medida que se disponga de nuevos datos, mediante la toma de decisiones informadas. Un ejemplo de ello es lo realizado en ciudades como Seúl, Chicago, Nueva York, Londres, entre otras, "donde se evidencia una mejora en la transparencia y la responsabilidad con la ciudad por parte del gobierno" (Martin y Begani, 2016).

Por otra parte, en los países desarrollados, el éxito de los proyectos relacionados con ciudades inteligentes está intrínsecamente relacionado con la existencia de grandes volúmenes de datos que podrían ser procesados para alcanzar sus objetivos. Ciudades o países como "Vancouver, Portland (Oregon), San Francisco (2009), Nueva York (2012) y Nueva Zelanda (2011) tienen sitios web donde publican datos sin requerir licencia facilitando su acceso y explotación por parte de terceros, generalmente empresas privadas y sin ánimo de lucro.

La figura 1 presenta un ejemplo de un tablero de indicadores que permite ver en tiempo real la información de Londres de manera resumida; así, se le permite a cualquier ciudadano hacer seguimiento de datos relacionados con el transporte y el aire, entre otros; consecuentemente, este tipo de información hace posible que cualquier ciudadano o empresa tome decisiones de acuerdo con su situación particular. En detalle, el tablero de indicadores presentado parte con los indicadores del estado de las líneas del metro, donde la mayor parte de ellas se encuentran en funcionamiento. Luego se disponen indicadores de la afluencia de bicicletas en las estaciones, incluyendo aquellas disponibles para utilizar. Posteriormente se registra la cantidad de buses y trenes disponibles para la prestación del servicio. Por último, se muestran indicadores de polución.

Figura 1. Tablero de indicadores de Londres, Inglaterra



Como caso de éxito, en la ciudad de Bandung, una de las más grandes de Indonesia, se propuso desarrollar una aplicación de monitoreo urbano a través de un tablero que permitiera resumir en tiempo real la condición climática, de transporte, económica, de salud y energía. El sistema de arquitectura utiliza sensores de red, que consiste en nodos con sensores que tienen la función de capturar las condiciones de la ciudad, como la temperatura, el nivel de polución del aire, el nivel de agentes contaminantes del agua y temas relacionados con el tráfico. La conclusión de este proyecto internacional se basa en el éxito alcanzado, por haber puesto en marcha el tablero de instrumentos de la ciudad inteligente y proporcionado información a los ciudadanos (Suakanto *et al.,* 2013, pp. 1-2).

Big data y sus desafíos en las ciudades inteligentes

Big data hace referencia a la manipulación de grandes volúmenes de datos (Khan *et al.,* 2017; Kacfah, Cullot y Nicolle, 2015); sin embargo, hasta el momento ni las ciencias de datos, ni la industria, ni la academia han logrado concluir cuál es la cantidad mínima de datos para otorgar tal nombre. Dos factores que podrían ayudar a determinar si el conjunto de datos forma parte de *big data* son la utilización de herramientas informáticas especiales para la administración de datos y la infraestructura física para procesarlos (Sivarajah *et al.*, 2017; Oussous *et al.*, 2017).

Según lo anterior, por ejemplo, un conjunto de datos que se almacenan en hojas de cálculo de Excel no podría ser considerado como *big data*, ya que el *software* utilizado forma parte de la *suite* de ofimática que un computador normalmente tiene instalado. En este escenario se evidencia capacidad limitada de almacenamiento y procesamiento. En específico, *big data* determina una utilización avanzada de recursos físicos, tales como servidores capaces de gestionar grandes volúmenes de datos de manera eficiente y por largos periodos (Khan *et al.*, 2017). Nuevos métodos matemáticos y tecnologías también son un factor fundamental para la limpieza, el análisis y el procesamiento de datos, con propósitos múltiples:

- Suministrar información a través de indicadores que facilitan la toma de decisiones en las empresas.
- Permitir un análisis detallado de los datos, llevando a las empresas a ser más competitivas en la personalización de servicios.
- Generar nuevas estrategias de mercadeo a partir del descubrimiento de nueva información.

Tipos de datos

Existe una clasificación de datos de acuerdo con su organización interna, definida con el fin de poder guardar, consultar, actualizar y eliminar información. De esta manera, las herramientas tecnológicas pueden gestionar adecuadamente los datos y cualquier operación que se desprenda de estos. A continuación se presenta la clasificación en mención:

- *Estructurados*. Hace referencia a los datos normalizados que se almacenan en bases de datos relacionales; por ejemplo, Oracle, MySQL o Sybase. Esto no aplica a datos que se encuentran en hojas de cálculo, ya que, por su volumen, no forman parte de *big data.*
- *No estructurados.* Son aquellos datos que carecen de organización estructural o que no se encuentran normalizados. Por ejemplo, datos en formato de audio, video y texto.
- *Semiestructurados.* Los datos semiestructurados no se ajustan a estándares estrictos; un ejemplo es el lenguaje XML (Extensible Markup Language), que contiene etiquetas que pueden ser legibles por cualquier máquina (Gandomi y Haider, 2015).

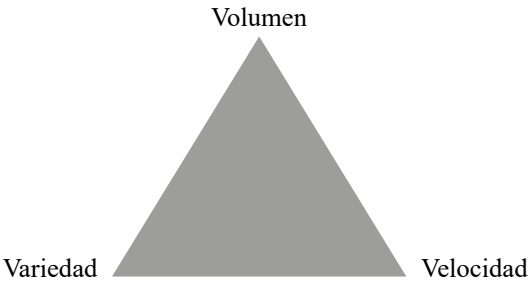
Evolución del termino

Aunque siempre han existido grandes volúmenes de datos de manera digital, el termino *big data* fue oficialmente utilizado en 1941, de acuerdo con el diccionario de inglés de *Oxford*. Posteriormente, fue consolidándose debido al crecimiento de datos que se generan en internet. Una evidencia de esto es la creación diaria de datos en fuentes tales como redes sociales, *smartphones*, carros y casos inteligentes, blogs, entre otros; 2,5 quintillones de bytes de datos son generados por estas fuentes diariamente, por lo cual *big data* es y seguirá siendo un concepto comúnmente utilizado (Actuaries, 2015).

Características

Las características de *big data* normalmente están orientadas a las tres *V*: volumen, variedad y velocidad; sin embargo, día tras día ha ido ampliando el número de características. En la figura 2 se presentan las comúnmente utilizadas de acuerdo con Actuaries (2015), Kacfah *et al.* (2015) y Oussous *et al.* (2017).

Figura 2. Características del big data, o las 3 V



a. *Volumen*. Esta característica está orientada a la utilización de grandes volúmenes de datos; por ejemplo, son los datos que el mismo internet genera diariamente (de acuerdo con la empresa International Data, 4.4 zetabytes son producidas anualmente).
b. *Variedad*. Se focaliza en los diferentes formatos en que los datos se encuentran estructurados. Normalmente, los datos almacenados no poseen un orden definido ni tampoco se encuentran disponibles para procesar inmediatamente. Como se indicó, los datos se pueden encontrar de manera: a) estructurada (almacenada en bases de datos relacionales), b) semiestructurada (correo electrónico, sensores o redes sociales) y c) sin estructurar (video, imágenes o audios).
c. *Velocidad*. Indica el flujo de información entre y hacia diferentes fuentes de datos. Esto involucra acciones tales como la creación, actualización y eliminación de datos. También incluye la oportunidad de acceder a la información, es decir, la disponibilidad y rapidez con que un usuario o sistema de información acceden a la información; por ejemplo, la rapidez con que se sube o descarga la información en una plataforma tecnológica.

Tecnológicas en big data

a. *Apache Hadoop.* Es una herramienta de código abierto que se liberó en internet en 2011; su objetivo es el procesamiento de grandes volúmenes de datos a través de sistemas distribuidos, utilizando algoritmos computaciones simples (Actuaries, 2015). El procesamiento de datos se hace mediante un procesamiento en paralelo de hasta cientos de máquinas; incluso es capaz de contemplar errores causados por daños en el *hardware*

de las máquinas, garantizando así el correcto funcionamiento de esta herramienta. Empresas dedicadas a la recopilación de información en el ámbito mundial adoptaron Hadoop para incrementar sus ingresos y ser más competitivas (como Amazon, Facebook, Twitter y Google) (Oussous *et al.*, 2017).

b. *Bases de datos no relacionales.* Son bases de datos que nacieron para gestionar *big data*, específicamente grandes volúmenes y variedad de datos. Las herramientas que se encuentran para almacenar datos son Oracle NoSQL, Apache Cassandra y Hive (Kacfah *et al.*, 2015).
c. *Data mining*. Un conjunto de métodos, algoritmos y técnicas utilizado para manipular datos; tiene como objetivo encontrar datos escondidos, los cuales no son visualmente perceptibles para el ser humano. En *big data*, la herramienta Map Reduce es comúnmente utilizada para practicar minería de datos.

Aspectos legales

Las empresas incluyen como procesos recopilar y procesar datos, con el fin de analizar los comportamientos de sus clientes; de esta manera, se pueden personalizar los productos, las ofertas y las promociones, ofreciendo aparentemente beneficios a una población específica de clientes fidelizados o potenciales. En este escenario, *big data* presenta oportunidades para las empresas, por cuanto les permite ser más creativas y competitivas a partir del análisis descriptivo y predictivo de datos, que ayudan a conocer el comportamiento de los clientes a partir de datos históricos que estos mismos han ido acumulando a partir de sus compras de bienes o servicios.

La protección de los datos personales es primordial durante el proceso de recopilación, procesamiento y análisis de información, debido a que la información básica de los clientes debe ser anónima, para prevenir discriminación y abuso durante el mercadeo digital (por ejemplo, enviar datos básicos tales como nombres, direcciones y números de documento a usuarios de manera errónea, poniendo en riesgo la privacidad). En el ámbito mundial se han establecido regulaciones, leyes y acuerdos que ayudan a garantizar el manejo de datos personales y la privacidad

de las personas. A continuación se presenta la evolución de los recursos legales que han concedido a las personas el derecho a la protección de sus datos:

- La Convención para la Protección de las Personas con Respecto al Procesamiento Automático de Datos Personales; adoptada por el Consejo Europeo en 1981.
- Actualmente, el instrumento legal para la Unión Europea es la Directiva de Protección de Datos (9546/EC), establecida en 1995, que regula la protección de las personas con respecto al procesamiento de datos y el libre movimiento de estos. Esta directiva es adoptada por cada uno de los países miembros de la Unión Europea.
- E-Privacy es otro instrumento legal (2012/58/EC) definido en la Unión Europea con el fin de garantizar la privacidad de los datos en el marco de servicios que se emplean a través de las comunicaciones electrónicas. Incluye el uso adecuado de dispositivos que se encargan de recopilar y almacenar información a través de *cookies*, las cuales deben ser informadas al usuario de internet, quien a su vez es libre de decidir si autoriza o no el manejo de sus datos.
- Regulación General para la Protección de Datos (GDPR), establecida en 2015 con el fin de aumentar la cobertura de la protección de datos y mejorar las oportunidades de negocio a partir del intercambio de información entre diferentes entidades.
- Directiva de Empleo Igualitario (2000/78/EC), orientada al manejo de datos evitando la discriminación de cualquier tipo: racial, étnica, religiosa, etc.
- Directiva de Servicios y Mercancías (2004/113/EC), la cual busca la utilización adecuada de los procesos inherentes al flujo de datos.
- En Colombia, la Ley 1581 de 2012 dicta un conjunto de disposiciones generales para la protección de datos personales. Con ella se busca garantizar el derecho a conocer, actualizar y rectificar cualquier información que se haya recopilado de los ciudadanos. Adicionalmente, la ley establece principios que garantizan la calidad, el tratamiento y la legalidad de los datos.

Big data analytics

El proceso de analítica de datos es utilizado para explorar, procesar y entender los datos y las relaciones que existen entre ellos. De esta manera, las empresas pueden extraer conocimiento invaluable para validar o establecer nuevas estrategias de mercadeo. Adicionalmente, se pueden determinar patrones que afectan positiva o negativamente los procesos de negocio (Oussous *et al.*, 2017). Dentro del proceso de analítica de datos existen técnicas tales como minería de datos, visualización de datos, análisis estadísticos y aprendizaje de máquina, que están basadas en algoritmos matemáticos que apoyan el procesamiento de datos para un fin específico; estos algoritmos están soportados en arquitecturas de *hardware* y *software*.

Por otra parte, también se han establecido métodos para el procesamiento de datos; su selección depende del conjunto de datos que se van a utilizar. A continuación se presentan los métodos comunes de acuerdo con el tipo de datos que se utilizan:

- *Extracción de información.* Consiste en extraer datos estructurados a partir de datos no estructurados. Posee dos actividades:
  - *Reconocimiento de entidades (ER)*. Toma el texto y lo clasifica de acuerdo con unas entidades (categorías) predefinidas.
  - *Extraer relaciones (RE)*. Se extraen relaciones semánticas entre entidades.
- *Resumen de textos.* Resume el contenido de uno o varios documentos. Tiene dos puntos de abordaje:
  - *Extracción de resumen.* El resumen es un subconjunto de todas las unidades de texto que componen el documento. La importancia de las unidades de textos depende de la frecuencia con que se repite una frase o palabra, así como su lugar, para luego construir el resumen (conjunto de unidades salientes). La ventaja es que no se requiere entender el texto analizado.
  - *Abstracción de resumen.* Es una técnica que consiste en analizar el texto semánticamente; luego se analizan y se incorporan algunas otras palabras a partir del estudio semántico utilizando técnicas de

lenguaje natural (NLP). Generalmente, este tipo de resúmenes son más coherentes que los resúmenes de extracción.

- *Respuesta a preguntas.* Suministra respuestas a preguntas utilizando lenguaje natural. Apples Siri o IBM Watson son un ejemplo.
  - Estos sistemas dependen de técnicas desarrolladas para lenguaje natural (NLP).
  - Existen tres categorías de preguntas a respuestas automáticas:
    - *Information retrieval (IR).* Se clasifica el tipo de pregunta, se analiza la información buscando en pasajes u oraciones que se precargan en el sistema y se determinan las respuestas candidatas.
    - *Basado en conocimiento.* Depende del contexto, la medicina, el turismo, etc.
    - *Híbrido*. Es la composición entre las dos categorías; es decir, se precargan las preguntas y respuestas, y al mismo tiempo se entrena al sistema con la información suministrada por el experto.
- *Análisis de sentimientos (opinion mining).* Consiste en definir qué tipo de opinión existe en una persona hacia una empresa específica, de acuerdo con su producto, servicio, etc. Existen tres tipos de análisis:
  - *Basado en documento.* Se lee todo un documento y se determina si tiene una posición positiva o negativa frente a la empresa.
  - *Basado en sentencias.* Se analiza una sentencia en particular y se determina el tipo de sentimiento.
  - *Basado en aspectos.* Se reconocen todos los sentimientos de una persona o conjunto de personas frente a una empresa, y se determina hacia qué parte de la empresa se tienen esos sentimientos; por ejemplo, un producto, servicio, área, etc.
- *Analítica de audio.* Consiste en analizar discursos, ligando sonidos con palabras. Normalmente, aplica a formatos de audio y video.
- *Analítica de video*. Consiste en analizar contenido de video; sin embargo, las técnicas aún no se encuentran en desarrollo, sobre todo cuando se quiere analizar video en tiempo real, debido a la gran cantidad de video. Por ejemplo, un segundo de alta definición de video equivale a 2000 páginas de texto. Si se quisiera analizar Youtube, sería una tarea titánica debido a que 100 horas de video son subidas cada minuto. Hay varias

ramas en las que se puede desarrollar análisis de video; por ejemplo, en la de mercadeo, específicamente en la situación en la que se requiera analizar cuántas personas ingresaron a una tienda, el género, la raza, la edad, el tiempo en que estuvieron, los patrones de comportamiento y el tiempo en una fila.

- *Analítica de redes sociales.* Se analizan datos estructurados y no estructurados producidos a partir de los eventos que generan los usuarios de redes sociales como Facebook, Twitter, LinkedIn, entre otras.

Desafíos

Desde sus múltiples perspectivas, *big data* ha venido desarrollándose debido a las necesidades y los beneficios que representa; sin embargo, aún existen desafíos por abordar y resolver desde diferentes áreas de conocimiento; los más sobresalientes son desafíos de datos, desafíos de procesos y desafíos de *software* y *hardware* (Sivarajah *et al.*, 2017; Bertot y Choi, 2013).

Desafíos de datos

Están orientados al aseguramiento de la calidad de los datos; específicamente seguridad, integridad, veracidad, velocidad, disponibilidad, volatilidad, gobernabilidad de los datos, privacidad, entre otros. El desarrollo de estándares que apoyen el aseguramiento de la calidad de los datos va a permitir que agencias, investigadores, científicos, compañías del sector privado y público, entre otros, gestionen datos de manera adecuada, con el fin de brindar análisis de resultados sin errores y con una alta calidad. De este modo, se impacta positivamente la toma de decisiones en las empresas (Bertot y Choi, 2013).

Desafíos de procesos

En este tipo de desafíos se busca investigar y solucionar técnicas para capturar datos, integrar datos a partir de sistemas de información, transformar datos, seleccionar el método adecuado para el análisis de información, etc. En línea con Sivarajah *et al.* (2017), los métodos comúnmente utilizados son:

- *Análisis descriptivo.* Describe la situación actual de una situación perteneciente al modelo de negocio analizado. Esta descripción permite abordar análisis de manera evidente, explicando patrones de comportamiento y excepciones.
- *Análisis predictivo.* Su objetivo es predecir comportamientos y posibilidades futuras del negocio a través de algoritmos y modelos estadísticos.
- *Análisis inquisitivo.* Se encarga de probar o certificar datos del negocio; por ejemplo, analizar factores de riesgo.
- *Análisis prescriptivo*. Su objetivo es optimizar y evaluar cómo el negocio puede mejorar sus niveles de servicio y al mismo tiempo disminuir sus gastos.

Desafíos de *software* y *hardware*

Las arquitecturas de *software* y *hardware* requeridas para obtener una plataforma que soporte *big data* son importantes con el fin de garantizar la recopilación, el procesamiento y el almacenamiento de información. Estos dos tipos de arquitecturas son dependientes, ya que si alguna de estas carece de las tecnologías mínimas requeridas para la gestión de *big data,* podría convertirse en una plataforma obsoleta. La capacidad computacional debe ser robusta, pues manipular *big data* así lo requiere. Esto es conocido como *super computing,* y es lo mínimo requerido para procesar largos conjuntos de datos, aplicaciones y visualizaciones para el análisis de datos. Adicionalmente, el almacenamiento de estos datos implica preservarlos por largo tiempo y aplicando políticas de gestión de datos.

Open data y sus desafíos en *smart cities*

Open data o datos abiertos

Open data es una tendencia mundial en la que los gobiernos se comprometen a publicar y compartir los datos derivados de la gestión pública, con el fin de contribuir a la transparencia como medio de confianza entre los ciudadanos y sus gobernantes. Esto también trae como consecuencia que empresas del sector privado y personas naturales puedan usar, reutilizar y distribuir los datos a través

de aplicaciones en línea (Datos.bcn.cl, 2017). La única condición en la mayoría de leyes que invitan a la aplicabilidad de los datos abiertos es que estos no sean alterados y se publiquen de la misma manera en la que se tomaron del repositorio de internet.

La definición del término *open data* puede variar entre autores y fuentes; sin embargo, todos coinciden en que son los datos originados de la gestión pública, publicados y compartidos a través de un repositorio de datos abiertos en internet, con el fin de que estos sean utilizados, reutilizados y redistribuidos libremente por cualquier persona natural o jurídica. El objetivo de *open data* es poner a disposición de la sociedad los datos del gobierno, para que estos se conviertan en información a través de *software* en internet que genere valores agregados para la toma de decisiones y la optimización de procesos, y de esta manera, se mejore la calidad de vida de los ciudadanos. Por tal razón, es crucial concientizar a los gobiernos mundiales y a las instituciones privadas acerca de los cambios positivos que esta tendencia trae para la humanidad, ya que transforma los esquemas tradicionales de los gobiernos, independientemente de la región en la que se encuentren ubicados (Repositorio.cepal.org, 2016).

El concepto de *open data* en el Gobierno colombiano se basa en que los datos generados por las entidades públicas pertenecen a la sociedad, dado que han sido financiados y recopilados con dinero público y, por tanto, deben estar disponibles para personas naturales o jurídicas, independientemente de su propósito e interés (Herramientas.datos.gov.co, 2016).

Una ventaja que brinda la implementación de *open data* en cualquier país es la posibilidad de que la ciudadanía se involucre en el análisis de datos, con el fin de mejorar indicadores de transparencia y eficiencia en la ejecución de procesos gubernamentales. A partir de lo anterior, también se obtiene:

- Generación de mayores ingresos para las empresas privadas y las personas naturales, ya que poseen información para analizar y generar estrategias que apoyen la creación de nuevos servicios o productos orientados a mejorar la calidad de vida de los ciudadanos. Por ejemplo, la ciudad de Seúl ha creado *startups* que ayudan a mejorar los procesos relacionados con el transporte público.

- Socialización de datos que permiten mantener informado a todo un país acerca de las decisiones que se toman y se ejecutan en las entidades gubernamentales.
- Implementar nuevas aplicaciones de *software* que complementen procesos ya creados, para de esta manera monetizar el nuevo conocimiento descubierto.

En conclusión, los datos abiertos buscan la interoperabilidad entre sistemas de información y personas, con lo cual se genera independencia entre fuentes de datos, repositorios de información, variedad de información y su destino.

Características

Según la Open Knowledge Foundation (s. f.), los datos abiertos son caracterizados a través de los siguientes aspectos:

a. *Uso.* Indica la disponibilidad de los datos en tiempo real y cuando estos se requieran, utilizando internet como medio para su acceso. Los desafíos que se deben resolver para esta característica, en términos de estandarización, son los formatos, la ambigüedad, el descubrimiento y la representación de los datos.
b. *Reutilización y modificación.* Consiste en la utilización y transformación de los datos obtenidos a partir de un portal web, el cual le permite a cualquier ciudadano acceder a los datos y generar un valor agregado a partir de ello, siempre y cuando los datos no se alteren.
c. *Participación universal.* Los datos deben y pueden ser accedidos por cualquier ciudadano, es decir, no debe existir ninguna restricción de acceso a los datos a ciudadanos o grupos por su condición racial, étnica, etc.

Calidad de datos abiertos

A partir de las leyes que apoyan la gestión de datos abiertos en los países, las entidades publican y comparten la información; sin embargo, los formatos y las estructuras de los datos varían drásticamente entre las entidades del Estado, dado que la naturaleza de cada entidad es diferente. La calidad de datos surge como un proceso de control debido a estas múltiples fuentes de datos y la manera como se

comparten. El nivel de calidad de los conjuntos de datos es determinado por el grado de apoyo en diferentes escenarios para la toma de decisiones informadas. Este grado es calculado de acuerdo con la aplicación de métricas matemáticas y criterios sobre los conjuntos de datos por utilizar. A continuación se definen cinco métricas para la medición de la calidad de los datos:

a. *Relevancia*. Los conjuntos de datos compartidos y seleccionados representan importancia para la toma de decisiones informadas. Las entidades públicas deben determinar si la información que se va a compartir ayudará a optimizar procesos en el estado, o bien, si simplemente son datos sin ningún valor para la ciudadanía.
b. *Exactitud*. Los datos deben ser totalmente precisos, independientemente de que sean numéricos o no. Por ejemplo, en un proceso contractual, un ciudadano quiere evaluar la variable *valor del contrato*, pero este no es exacto, ya que la entidad únicamente compartió valores globales y aproximados. Esta situación en particular, seguramente, tendrá imprecisiones en la información que se genera para la toma de decisiones y, en consecuencia, conllevará la pérdida de credibilidad.
c. *Oportunidad*. Los datos pueden y deben estar disponibles en el portal de internet, con el fin de que sean accesibles en el momento en que los ciudadanos los requieran. Por ejemplo, si la persona quiere acceder a un conjunto de datos, pero la entidad le manifiesta que esto no es posible y que, por tanto, debe esperar para que estos sean compartidos, se indica que no hay oportunidad en los datos, ya que estos no pueden ser accedidos.
d. *Comparabilidad*. Indica la facilidad con la que diferentes variables se pueden contrastar, independientemente del conjunto de datos. Hay situaciones en las que se requieren comparar valores de gasto a través de diferentes conjuntos de datos; sin embargo, hay circunstancias en las que las variables, a pesar de que tienen el mismo nombre y objetivo, no cuentan con iguales valores, lo cual indica que no hay comparabilidad entre los conjuntos de datos.
e. *Completitud*. Esta característica evalúa si los valores de cada variable se encuentran definidos. Por ejemplo, en un conjunto de datos donde la variable *ciudad* es crucial para la toma de decisiones, la entidad, no

obstante, no comparte los datos con estos valores; para este ejemplo en particular, el conjunto de datos no está completo.

Datos abiertos en Colombia

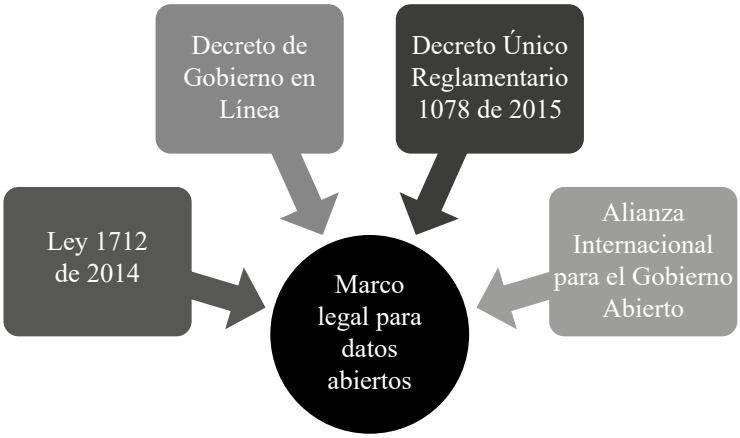
Datos Abiertos es un proyecto que inició en 2011 en el marco de la estrategia de Gobierno en Línea (GEL) y la Alianza Internacional para el Gobierno Abierto (AGA-OGP), cuyo objetivo principal fue compartir y publicar los datos derivados de la gestión administrativa pública. Los principales retos que se abordaron en el proyecto fueron brindar datos de forma oportuna y compartirlos en un formato que facilitara la utilización, reutilización y distribución por parte de los ciudadanos o empresas que quisieran su acceso.

Los objetivos del Proyecto de Datos Abiertos abarcaron diferentes puntos de vista: fomentar la utilización de datos abiertos, con el fin de generar servicios innovadores que suministren soluciones a problemas sociales en los ámbitos de la educación, la cultura, la salud, la seguridad, entre otros; involucrar a sectores como la academia, la industria, la sociedad civil y las organizaciones no gubernamentales, con miras a motivar su participación mediante la financiación y el desarrollo de proyectos que generen valor agregado a los datos abiertos; ser un ejemplo en la región por las buenas prácticas en la implementación de la estrategia de datos abiertos.

Los actores identificados que eventualmente podrían generar valor sobre los datos abiertos en Colombia son: academia, entidades públicas y privadas, emprendedores, ciudadanos inteligentes, medios de comunicación y cualquier veedor público. El proyecto definió cuatro etapas para su implementación: a) soporte a las entidades púbicas para la publicación de datos abiertos en formatos estructurados que permitan su utilización, reutilización y distribución en el portal *www.datos.gov.co*; b) otorgamiento de incentivos a los actores que implementen soluciones o servicios basados en datos abiertos; c) mantenimiento y mejora continua del portal *www.datos.gov.co*, con el fin de garantizar el funcionamiento del portal destinado para publicar los datos abiertos; f) apoyar el uso de herramientas para la visualización de datos, las cuales ayudarán a los actores identificados anteriormente a interpretar y analizar datos de manera eficaz.

En el aspecto legal se han implementado leyes, decretos y estrategias que incentivan, motivan y apoyan a las entidades públicas a compartir los datos abiertos a través del portal web *www.datos.gov.co* (figura 3). La Ley 1712 de 2014 establece los procedimientos para garantizar la transparencia y el acceso a la información pública. El Decreto de Gobierno en Línea (Decreto 2573 de 2014) plantea las obligaciones para abrir, divulgar y promover la reutilización de los datos abiertos. El Decreto Único Reglamentario del Sector TIC - 1078 de 2015 suministra lineamientos para adoptar buenas prácticas en referencia a los estándares abiertos, con el fin de contribuir a la eficiencia y transparencia en el Estado colombiano.

Figura 3. Marco legal que apoya los datos abiertos en Colombia



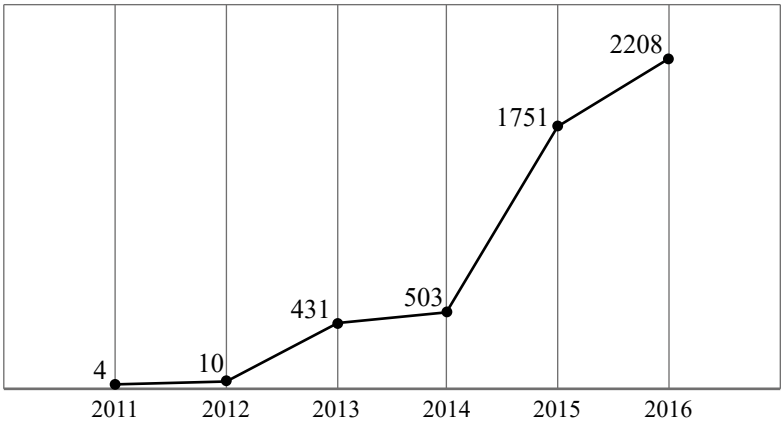
En 2015, el Banco Mundial realizó un diagnóstico al Proyecto de Datos a Abiertos en Colombia. El resultado de esta actividad arrojó recomendaciones para una adecuada implementación de una estrategia y política en esta materia. Entre las recomendaciones más destacadas se encuentran la definición de conjuntos de datos en formatos estándar que puedan ser interpretados por máquinas y usuarios finales, para de esta manera asegurar el intercambio de datos; por otro lado, la definición y el despliegue de un portal web público para compartir datos abiertos; finalmente, mantener los datos bajo una licencia abierta que permita dar cumplimiento a los principios de datos abiertos: uso, reutilización y redistribución.

De acuerdo con el Programa Gobierno en Línea (2016), los principios definidos bajo el marco del Proyecto de Datos Abiertos en Colombia son:

a. *Primarios.* Los datos deben mantenerse en su forma y detalle original, es decir, no deben ser agregados ni transformados, con el fin de garantizar la exactitud.
b. *Accesibles.* Los datos deben estar disponibles para los ciudadanos o las empresas que requieran usarlos, reutilizarlos y distribuirlos, independientemente de su propósito particular.
c. *Completos.* Los datos deben estar en el mayor detalle posible, sin datos nulos, en clave de garantizar suficiencia y consistencia en el proceso de interpretación y análisis de información.
d. *Procesables por máquinas.* Los formatos para publicar y compartir información deben ser estándar, con el fin de que máquinas y usuarios puedan intercambiar datos.
e. *No propietarios.* Los datos publicados en el portal web no son exclusivos de ninguna persona o entidad pública.
f. *Licenciados de forma abierta.* Los datos abiertos que se publican deben tener asociados términos y condiciones de utilización por parte de quienes van a consumirlos. Adicionalmente, los datos abiertos deben tener licenciamiento abierto.
g. *No discriminados*. Los datos abiertos publicados en el portal web pueden ser consumidos o accedidos por cualquier usuario; no debe exigir ningún tipo de autenticación o registro para hacer uso de los datos.
h. *Oportunos y actualizados.* Las entidades públicas deben mantener una frecuencia de actualización periódica, con el fin de garantizar que los datos se encuentren al día.

A continuación, en la figura 4, se presenta la evolución del número de conjuntos de datos entre 2011 y 2016.

Figura 4. Número de conjuntos de datos publicados en el portal *www.datos.gov.co*, periodo 2011-2016



Fuente: www.datos.gov.co (2016).

Datos abiertos en el mundo

Los países pioneros en la adopción de los principios y políticas de *open data* fueron Estados Unidos, Dinamarca, Noruega, Francia, Holanda y Gran Bretaña, que visualizaron las oportunidades de negocio para la misma sociedad, pues mediante la reutilización de datos podrían generarse nuevas innovaciones y cambios de paradigmas en los ciudadanos.

En América Latina, los países que ya han adoptado políticas y principios de *open data* son Colombia, Chile y México. Aunque existen otras naciones que están iniciando este proceso, en estos tres países ya existe una normatividad que permite y obliga a las entidades del Estado a compartir la información. En concordancia, las entidades gubernamentales están obligadas a publicar y compartir los datos con el fin de conseguir eficiencia y transparencia en sus procesos, mediante la participación de los ciudadanos y las empresas en el análisis y la interpretación de datos.

En general, los países que adoptan *open data* en sus políticas de gobierno coinciden en que esto les permitirá establecer una conversación constante entre el gobierno y la ciudadanía; tomar decisiones de administración pública basadas en las necesidades y preferencias de la comunidad; facilitar y promover la colaboración de la ciudadanía y las instituciones que comunican sus decisiones de

forma abierta; garantizar el acceso a la información. Este derecho es legitimado y defendido por el principio de transparencia (Mahecha, López y Velandia, 2017).

A pesar de que algunos países se encuentran más maduros que otros en los procesos inherentes a datos abiertos, aún se deben seguir atacando múltiples desafíos en función de asegurar el éxito en la implementación de políticas en esta materia. A continuación, de acuerdo con Charalabidis, Alexopoulos y Loukis (2016), se presenta la lista de desafíos que aún faltan por resolver en el contexto internacional:

a. *Instrumentos de medición*. A través de las estrategias y los proyectos de datos abiertos se deben incorporar modelos de medición que permitan determinar indicadores tales como el grado de efectividad de la implementación de datos abiertos, el porcentaje de participación de los ciudadanos y empresas en la gestión de datos abiertos, el grado de innovación de proyectos a partir de la implementación de aplicaciones de *software* en la sociedad, etc.
b. *Anonimizar datos abiertos*. Las entidades públicas deben velar y reservar conjuntos de datos que forman parte de la seguridad nacional y de tratados internacionales que solo les competen a ciertos actores en un país; sin embargo, una posible manera de publicar y compartir estos datos es a través de la anonimización, a fin de que nuevos servicios y aplicaciones se puedan desarrollar a partir de estos datos, salvaguardando los datos originales.
c. *Limpieza de datos.* Los ciudadanos y las empresas interesados en desarrollar nuevos servicios o aplicaciones a partir de datos abiertos deben considerar métodos para la limpieza de datos, con el fin de extraer, transformar y cargar datos a otros repositorios, y de esta manera aplicar métodos de minería y visualización de datos. Por lo tanto, es crucial definir métodos estandarizados para realizar limpieza de datos abiertos.
d. *Visualización de datos.* Los repositorios públicos de datos abiertos en el ámbito mundial están diseñados para almacenar, procesar y mostrar *big data*; ahora el objetivo es identificar, seleccionar y aplicar métodos adecuados para la fácil visualización, interpretación y análisis de datos por parte de cualquier ciudadano. De esta forma, las ciudades podrán

llamarse *ciudades inteligentes*, ya que la tecnología y las comunicaciones sirven como soporte para la toma decisiones informadas.

e. *Datos enlazados (linked data)*. El siguiente paso después de publicar y compartir la información es ofrecerles a los usuarios herramientas de *software* que permitan crear y mostrar datos relacionados entre diferentes conjuntos de datos, es decir, datos con una naturaleza diferente. Por ejemplo, crear relaciones entre un conjunto de datos financieros con un conjunto de datos de salud.

f. *Proceso de publicación*. Si bien los procesos para publicar se encuentran definidos en documentos y guías para ejecutar correctamente las actividades asociadas, aún no se tiene automatizado un flujo que permita hacer más expedito el proceso de publicación; por consiguiente, la participación activa de flujos de procesos a través de herramientas de *software* simplificaría el proceso actual para la publicación de datos.

g. *APIs y servicios*. La creación de estándares para el consumo de datos entre sistemas de información debe estandarizarse, por cuanto cada país es libre de implementar y utilizar librerías para exponer los datos; por eso, se complejiza la manera de extraer los datos de los portales que suministran datos abiertos.

Recientes estudios han demostrado que los datos abiertos han ayudado al desarrollado de servicios y aplicaciones en las siguientes áreas: innovación, analítica de datos, toma de decisiones, anticorrupción, ciudades inteligentes, entre otras. Ello trae como consecuencia la optimización de procesos y la generación de valor agregado en variables tales como economía, social y eficiencia de gobierno. Attard *et al.* (2015) presentan algunas aplicaciones que evidencian lo mencionado anteriormente:

- *Innovación*. La ganancia económica y el valor social no se encuentran directamente relacionados con esta área. La innovación está asociada a la manera en que un paradigma cambia disruptivamente a partir de un invento que implica el cambio en la manera de ejecutarse un proceso, normalmente con repercusión en la calidad de vida de las personas. *DontEat.at* es una aplicación que ayuda a clientes de restaurantes a validar la calidad del lugar que van a visitar (Global Open Data Index, s. f.).

- *Analítica de datos*. A través de herramientas de *software* que se enfocan en la aplicación de métodos para realizar análisis de *big data,* se han desarrollado nuevos procesos que involucran la optimización de actividades asociadas a medioambiente, transporte y educación. Un ejemplo es la aplicación *Street Bump*, que indica el estado de las calles en Estados Unidos.
- *Toma de decisiones y anticorrupción*. El alcalde de Seúl, Corea del Sur, tiene la posibilidad de integrar los datos abiertos de diferentes sectores que afectan a los ciudadanos: transporte, salud, precios de productos, entre otros. Toda la información se muestra a través de indicadores que le permiten a él tomar decisiones importantes. La información que se muestra en el *software* que el alcalde utiliza también está disponible para los ciudadanos, con lo cual se aplica el concepto de transparencia y como resultado se construye la confianza de los ciudadanos hacia el gobierno.
- *Ciudades inteligentes*. El indicador social enfocado a la seguridad también se ha visto beneficiado por la aplicación *Crime Finder*, que revela los delitos denunciados por las personas, y ayuda a la policía y a los ciudadanos a estar alerta, de acuerdo con la ubicación que se capture a través del celular (AppAdvice, s. f.).

Caso de estudio

El prototipo de *software* que se desarrolló para mejorar la calidad de vida de los ciudadanos, considerando los conceptos de *smart cities, open data* y *big data*, se encuentra publicado en internet con el fin de iniciar una democratización de la información generada por las entidades públicas del territorio colombiano. En específico, el caso de estudio se basa en el análisis de los datos suministrados por las entidades públicas al repositorio de datos *www.datos.gov.co*. El rango de datos que se tomó comprende las variables de los contratos celebrados en 2017, independientemente de su naturaleza. En la siguiente sección de análisis de los resultados se abordará el nivel de calidad que las entidades públicas comparten a la ciudadanía; la calidad se calculará a partir de las variables *trazabilidad*, *completitud* y *cumplimiento*.

A continuación se presentan los módulos que contiene la herramienta, para que de esta manera la organización o el ciudadano que requieran utilizarla puedan hacerlo con el conocimiento mínimo, de una manera eficaz y efectiva.

Módulos de la herramienta

Módulo 'Ingreso de estructura'

Este módulo tiene como objetivo almacenar los datos que se importan desde *www.datos.gov.co.* El módulo requiere dos variables para iniciar el flujo de importación de datos del repositorio en mención. La primera variable es llamada *ID del conjunto de datos,* que indica el bloque de datos que se va a descargar; por ejemplo, el conjunto de datos de los contratos celebrados en Bogotá. La segunda variable define el repositorio de datos del cual se quiere obtener los datos; para el caso de estudio únicamente se abordó el repositorio de *www.datos.gov.co*; no obstante, el prototipo puede descargar datos de cualquier otro repositorio (figura 5).

Figura 5. Módulo para ingresar estructuras de datos

El valor de los dos campos mencionados son suministrados por el repositorio de datos abiertos *www.datos.gov.co*; sin embargo, el prototipo también es capaz de soportar la importación de datos de otros países, como Inglaterra y Estados Unidos. Por lo tanto, el prototipo valida la estructura de datos para cualquier repositorio que haga uso de los estándares mínimos para el intercambio de datos abiertos.

La figura 6 presenta la validación que realiza el *software* para garantizar que los datos importados se almacenan correctamente. De esta manera, se asegura que las estructuras de datos por importar funcionen correctamente de acuerdo con lo definido por el portal de datos en mención. Adicionalmente, el *software* realiza la validación de que la estructura que se va a crear no exista previamente, con el fin de evitar estructuras duplicadas (figura 7).

Figura 6. Validación exitosa del ingreso de datos

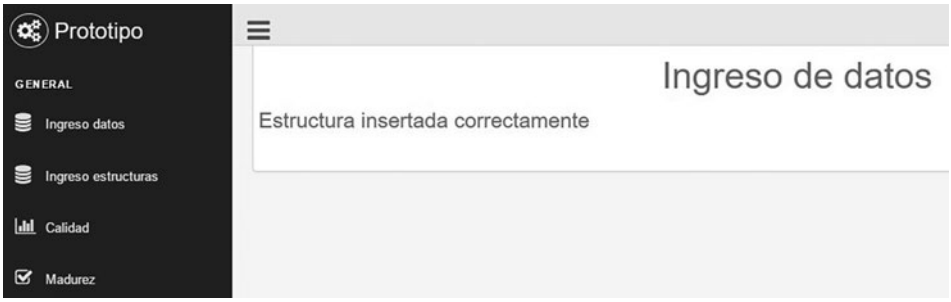


Figura 7. Validación del tipo de estructura utilizada

Módulo 'Ingreso de datos'

Este módulo tiene como función definir los conjuntos de datos por descargar, los cuales son identificados a partir de ID's definidos por el portal *www.datos.gov.co* (figura 8). Otra característica del *software* consiste en que el usuario final puede seleccionar varios conjuntos de datos. Un ejemplo de datos por importar es *r4jv-v36v*. Adicionalmente, se define el portal de datos abiertos por utilizar, lo cual se realiza a través del campo llamado "Ingrese el dominio", ya que el *software* tiene la capacidad de importar datos abiertos desde otros países (por ejemplo, el portal de datos abiertos de Estados Unidos). Una vez se envían, se define el conjunto de datos

y el repositorio sobre el cual se van a importar, el *software* corre un algoritmo que mide las siguientes métricas de calidad: completitud, trazabilidad y conformidad.

Figura 8. Módulo de ingreso de datos

La figura 9 muestra si el conjunto de datos se insertó correctamente. Esta operación se hace mediante la validación y el aseguramiento de que los datos se almacenen en la base de datos, con el fin de correr el proceso de calidad de datos. Por otra parte, y con el fin de garantizar que los datos no se dupliquen, el prototipo de *software* valida que los datos no se encuentren previamente guardados en la base de datos; esta es una manera de mantener la calidad de los datos (figura 10). Ahora bien, considerando las estructuras definidas por el portal *www.datos.gov.co* para el intercambio de datos abiertos, el prototipo de *software* se encarga de validar que la estructura de los datos se encuentre correctamente definida. Por ejemplo, si la estructura define que todos los datos tengan una columna ID, y la estructura por importar no la contiene, el *software* no ejecutará el proceso de importación del conjunto de datos (figura 11).

Figura 9. Validación exitosa de almacenamiento de datos

Figura 10. Mensaje de error que valida datos previamente guardados

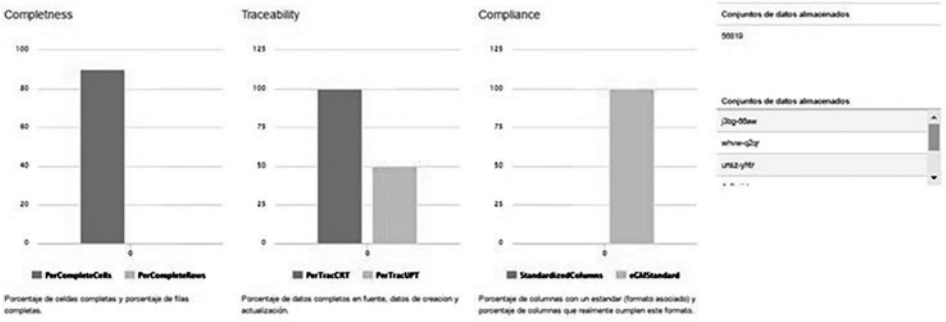
Figura 11. Mensaje de error por estructura no válida

Módulo de reportes

El módulo de reportes fue construido con el fin de mostrar los resultados de la evaluación que el modelo propuesto ejecutó. Los resultados se muestran en tres categorías: completitud, trazabilidad y cumplimiento. Cada una de estas se definieron y explicaron en la sección de calidad de datos abiertos. La figura 12 es un resumen de los registros suministrados por *www.datos.gov.co* en lo referente a los contratos legalizados del 2017.

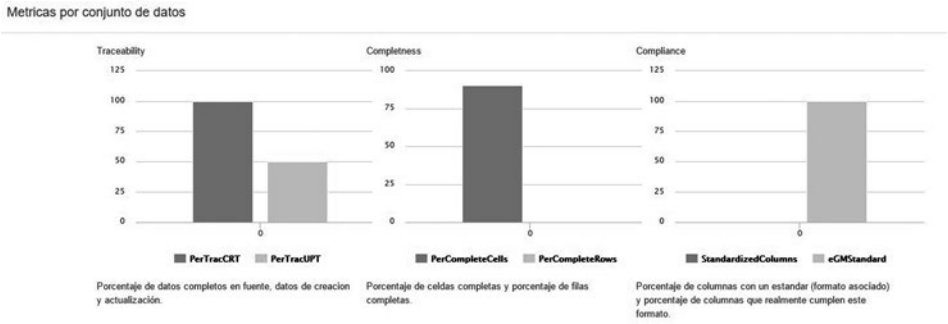
El módulo de reportes consta de dos partes: la presentación de resultados acumulada y la presentación de resultados para un conjunto de datos en particular. Los resultados acumulados son presentados a partir del análisis de todos los conjuntos de datos cargados al prototipo de *software*, independientemente de su naturaleza; es decir, el modelo de análisis podrá evaluar al mismo tiempo un conjunto de datos relacionado con educación y otro con contratación.

Figura 12. Módulo de resultados de calidad de datos acumulado



La figura 13 presenta un análisis de la calidad de datos para un conjunto de datos específico. Al igual que la anterior figura, existen las mismas categorías y se aplica el mismo análisis de datos; la única diferencia consiste en que el usuario final debe seleccionar el conjunto de datos por analizar.

Figura 13. Módulo de reporte de calidad de datos filtrado por conjunto de datos



Análisis de resultados de la herramienta

Conjunto de datos acumulados

Con respecto a la métrica de *completitud*, el modelo planteado evalúa que celdas, filas y columnas se encuentren con algún dato, independientemente de su valor. La figura 13 evidencia que la muestra de conjuntos de datos seleccionada tiene un 90% de completitud en las celdas, mientras que existe un 5% de completitud en sus filas; es decir, las entidades públicas deben seguir trabando en asegurar la amplitud y profundidad de los datos que compartan con la ciudadanía.

La métrica de *trazabilidad* evalúa si los metadatos (fuente de datos, frecuencia de actualización, periodo de datos actualizados, fecha de corte, entre otros) se encuentran definidos para cualquier conjunto de datos establecido durante el proceso de creación y actualización de dichos datos. Para el conjunto de datos evaluados, se puede concluir que el 100% de los conjuntos de datos contiene metadatos cuando se trata de la creación en *www.datos.gov.co*; sin embargo, cuando las entidades públicas actualizan datos, apenas un 50% contienen metadatos, dejando incompleta la información relacionada con los conjuntos de datos.

Con respecto al *cumplimiento,* el estándar define metadados asociados a datos abiertos, como fuente, fecha de creación, categoría y título. Aunque existen metadatos opcionales, como descripción, publicación, cobertura, entre otros, estos no son considerados dentro de la evaluación. De acuerdo con la figura 13, el 99% del conjunto de datos analizados por la herramienta tiene metadatos definidos.

Conjunto de datos específico

De manera similar, el conjunto de datos (código de datos abiertos: *j3bg-66aw*) analizado corresponde a los contratos celebrados entre la Contraloría General de Antioquia y diferentes proveedores; para mayor detalle, la figura 14 muestra la ficha técnica del conjunto de datos. Puntualmente, 90% de los datos suministrados por la entidad se encuentran completos; el 100% de los datos creados contienen metadatos y el 48% de metadatos fueron definidos durante el proceso de actualización; por último, el 100% de los datos cumplen con el estándar de formato asociado a los datos.