ISSN 0188-7297

INSTITUTO MEXICANO DEL TRANSPORTE

SCT
SECRETARÍA DE COMUNICACIONES Y TRANSPORTES

Datos masivos geoespaciales aplicados al transporte

Juan Carlos Vázquez Paulino
Miguel Ángel Backhoff Pohls

Publicación Técnica No. 502
Sanfandila, Qro, 2017

SECRETARÍA DE COMUNICACIONES Y TRANSPORTES

INSTITUTO MEXICANO DEL TRANSPORTE

Datos masivos geoespaciales aplicados al transporte

Publicación Técnica No. XXX
Sanfandila, Qro, 2017

Esta investigación fue realizada en la Unidad de Sistemas de Información Geoespacial del Instituto Mexicano del Transporte, por el Lic. Juan Carlos Vázquez Paulino; es el producto final del proyecto de investigación interna VI-01/17 “Datos masivos geoespaciales aplicados al transporte”.

Se agradece la colaboración del M. G. Miguel Ángel Backhoff Pohls, Jefe de la Unidad de Sistemas de Información Geoespacial del IMT, por sus acertados comentarios y apoyo para la realización de esta publicación.

Así mismo, se reconoce la participación, comentarios y aportaciones del Laboratorio Nacional Sistemas de Transporte y Logística (SiT-LOG Lab).

Contenido

Índice de figuras

Sinopsis

En este trabajo, en primer lugar se define lo que se entiende por datos masivos. Se presentan diversas definiciones destacando el hecho de que el tamaño es sólo una dimensión de los datos masivos, otras dimensiones, como la velocidad, la variedad, la veracidad, la viabilidad, la visualización y el valor son igualmente importantes. Se revisan las técnicas de análisis para datos de texto, audio, vídeo y medios sociales.

Se revisan los datos masivos geoespaciales (*geospatial big data*) y se identifican algunas fuentes de datos, también se analiza el modo en que se relacionan con el transporte y su posible aplicación en distintas áreas.

Se identifican las fases principales del proceso de los datos masivos geoespaciales, así como las principales técnicas de representación de éstos y los distintos tipos de algoritmos para procesamiento de datos.

Como parte de los objetivos del proyecto, se realizaron ejercicios con el uso de datos de dos distintas plataformas, en primer lugar Twitter, en donde la aplicación identifica los tweets que presentan geolocalización y se encuentran a cierta distancia de un punto de coordenadas inicial; el segundo ejercicio utiliza datos de Waze, los cuales se descargaron durante determinado lapso y al procesarlos y generar mapas digitales se pueden identificar patrones de ocurrencia de eventos capturados por la comunidad.

Abstract

In this paper, we first define what is meant by big data. There are several definitions highlighting the fact that size is only one dimension of big data, other dimensions such as speed, variety, accuracy, viability, visualization and value are equally important. Analysis techniques for text, audio, video and social media data are reviewed.

Geospatial big data is reviewed and some data sources are identified, as well as the way in which they relate to transport and its possible application in different areas.

It identifies the main phases of the geospatial big data process, as well as the main techniques of data representation and the different types of algorithms for data processing.

As part of the objectives of the project, exercises were carried out using data from two different platforms, firstly Twitter, where the application identifies the tweets that have geolocation and are some distance from an initial coordinate point; the second exercise uses data from Waze, which were downloaded during a certain period of time and when processing and generating digital maps can identify patterns of occurrence of events captured by the community.

Resumen ejecutivo

Esta investigación tuvo como objetivo principal conocer y reconocer los principales aspectos acerca de los datos masivos (*Big Data en inglés*) en donde una parte importante es el componente geoespacial, el cual viene en diversas formas como códigos postales, dirección IP, localización geográfica (coordenadas x,y,z), y recientemente se comienza a usar el análisis del contenido para identificar la ubicación exacta o lo más cercana posible de los datos o eventos relacionados.

Al explorar el estado del arte de los datos masivos geoespaciales y determinar su potencial de utilización en el transporte, mediante la realización de un estudio piloto que maneje datos representativos (espaciales y temporales) de dos plataformas distintas, se podrá analizar el flujo del autotransporte carretero.

Durante el estudio se analizaron los paradigmas existentes relacionados a datos masivos así como sus procesos geo-informáticos; se exploró una muestra representativa de datos masivos para identificar aplicaciones al transporte y se generó una aplicación piloto o caso de uso que muestra la utilización de datos masivos geoespaciales en el transporte. Así mismo se identifican algunas de las áreas de oportunidad para el desarrollo e investigación de la disciplina de datos masivos geoespaciales.

Al conocer el panorama de lo que está sucediendo sobre el terreno le da oportunidad a la analítica geoespacial en el sector de transporte para aprovechar las herramientas de datos y análisis predictivo y de este modo ayudar a las agencias de transporte a mejorar las operaciones, reducir los costos y servir mejor a los viajeros.

1 Introducción

En los últimos años ha habido una explosión en la cantidad de datos que se encuentra disponible. Ya sean los registros de un servidor web, tuits (mensajes enviados a través de twitter[1]), registros de transacciones en línea, datos “ciudadanos", datos de sensores, datos gubernamentales o de alguna otra fuente; el problema no es encontrar datos, es averiguar qué hacer con ellos. Ahora, las instituciones y empresas utilizan sus propios datos, o los datos aportados por sus usuarios. Cada vez es más común la combinación de datos de un gran número y diversidad de fuentes.

La pregunta que enfrenta cada institución hoy en día es cómo usar los datos de manera efectiva (no sólo sus propios datos, sino todos los datos disponibles y relevantes). Utilizarlos efectivamente requiere algo diferente de las estadísticas tradicionales, donde solo se hacen gráficas y análisis sin representación geoespacial. Ahora se hace necesario utilizar análisis de tendencias espaciales, contigüidad, continuidad, distancia y tiempo, concentración de eventos, pronósticos de ubicación y comportamiento, entre otras.

En términos generales, los datos pueden haber "nacidos digitales" o "nacido análogos" (PCAST, 2014). Cuando "nacen digitales" los datos son creados por los usuarios o por un dispositivo informático específicamente para su uso en un entorno de procesamiento de la máquina.

Ejemplos de datos:

- Sistema de Posicionamiento Global (GPS) y otros tipos de datos geo-localizados
- Registros de proceso y de horas
- Metadatos relativos a la identidad del dispositivo, estado y ubicación utilizados por los dispositivos móviles para mantenerse conectados a varias redes (GSM, Wi-Fi, etc.).
- Datos producidos por dispositivos, vehículos y objetos en red.
- Datos de acceso de tarjetas de transporte público y datos asociados con el acceso a portales (tarjetas de identificación, Etiquetas RFID) o paso por arcos detectores (por ejemplo, autopistas, sistemas de carga de congestión).
- Datos de transacciones comerciales (uso de tarjetas de crédito y registros de transacciones, código de barras y lectura de etiquetas RFID).

[1] www.twitter.com

- Correos electrónicos y SMS, metadatos relacionados con llamadas telefónicas.

Los datos "digitales" se diseñan y producen para abordar necesidades específicas. Las consideraciones de eficiencia han significado que sólo los datos específicos requeridos para un proceso fueron generados con el fin de evitar problemas de almacenamiento y capacidad de procesamiento o para inflar los costos. Sin embargo, los costos de procesamiento y almacenamiento significan que la recolección excesiva de datos es fácilmente posible y el costo de realizar tal acción es muy bajo.

Los datos "analógicos" son datos que surgen de una impresión de un fenómeno físico (luz, sonido, movimiento, presencia de un compuesto químico o biológico, impedancia magnética, etc.) sobre un dispositivo de detección, y su posterior conversión en una señal digital. Los sensores pueden incluir cámaras, micrófonos, dispositivos de detección de campos magnéticos, monitores de frecuencia cardíaca, acelerómetros, sensores térmicos, entre otros. Los ejemplos de datos "análogos" (PCAST, 2014), incluyen:

- Secuencias de vídeo desde cámaras de vigilancia, en vehículo, en carretera u otras.
- Contenido de audio de llamadas telefónicas de voz, audio ambiental de cámaras de video o redes de micrófonos.
- Movimiento / inercia (acelerómetros, sensores ultrasónicos)
- Rumbo (brújula), temperatura, infrarrojos
- Radiación, campos electromagnéticos, presión de aire, etc.
- Datos relativos al ritmo cardíaco, la respiración, la marcha y otros parámetros físicos y de salud.
- Reflectancia electromagnética o luminosa (láser) de objetos (por ejemplo, radar de apertura sintética -AR o láser- Basados en sistemas LIDAR).

En la *Figura 1.1 Los datos nunca duermen 3.0*, se observan algunos datos relevantes acerca de lo que sucede cada minuto en cuanto a los datos, su tamaño, métodos de creación y utilidad; por ejemplo, se toman 694 viajes a través de la plataforma de Uber, los usuarios generan más de 4,000,000 de “me gusta” en Facebook, Amazon recibe la visita de 4300 usuarios, los usuarios de Apple descargan 51,000 aplicaciones, los usuarios de Twitter envían 347,222 tweets, en Youtube los usuarios suben 300 horas de nuevos videos, 1,736,111 fotos de Instagram reciben un “me gusta”.

Fuente www.domo.com

Figura 1.1 Los Datos nunca duermen 3.0.

El crowdsourcing - (del inglés crowd –multitud– y outsourcing –recursos externos–) se podría traducir al español como colaboración abierta distribuida o externalización

abierta de tareas y consiste en externalizar tareas que, tradicionalmente, realizaban empleados o contratistas, dejándolas a cargo de un grupo numeroso de personas o una comunidad a través de una convocatoria abierta.

Jeff Howe[2], uno de los primeros autores en emplear el término, estableció que el concepto de "crowdsourcing" depende esencialmente del hecho de que, debido a que es una convocatoria abierta a un grupo indeterminado de personas, reúne a los más aptos para ejercer las tareas, para responder ante problemas complejos y para así contribuir aportando las ideas más frescas y relevantes. Por ejemplo, se podría invitar al público a desarrollar una nueva tecnología, o a llevar a cabo una tarea de diseño (diseño basado en la comunidad o diseño participativo distribuido), o a mejorar e implementar los pasos de un algoritmo o ayudar a capturar, sistematizar y analizar grandes cantidades de datos (ciencia ciudadana).

El término se ha hecho popular entre las empresas, autores y periodistas como forma abreviada de la tendencia a impulsar la colaboración en masa, posibilitada por las nuevas tecnologías como la Web 2.0 o las redes sociales, para así lograr objetivos de negocios o eventualmente propuestas sociales. Sin embargo, el término y sus modelos de negocio han generado controversia y críticas.[3]

Ventajas
- Compilación de una gran variedad de propuestas de alta calidad
- Disminución de costos
- Retroalimentación interna y permanente
- Generación continúa de ideas innovadoras

Desventajas
- Puede generarse mucha basura, por lo que se deben revisar los datos
- Procesos lentos

Para más información revisar el Anexo I. Antecedentes del Big Data (cronología).

[2] http://www.penguinrandomhouse.com/books/83579/crowdsourcing-by-jeff-howe/9780307396211/

[3] http://www.empresasenred.es/empresasenred/blog/%C2%BFqu%C3%A9-es-crowdsourcing-y-ventajas-tiene

2 Marco teórico

2.1 Definición de los datos masivos (Big Data)

El término "grandes datos", “datos masivos” o “big data” apareció por primera vez en las comunidades científicas a mediados de la década de 1990 (*ver anexo I. Antecedentes del Big Data - cronología*), es un concepto de inciertos orígenes, Francis X. Diebold (2012) argumenta que el término probablemente se originó en una conversación en Silicon Graphics Inc. en donde John Mashey, lo utilizó en una presentación titulada “*Big Data and the next wave of infrastress*”[4] en donde mostró un pensamiento muy avanzado y estructurado con respecto a lo que era el Big Data en esa época. A partir de entonces el término comenzó a utilizarse en los trabajos de estadística y econometría y poco a poco se popularizó alrededor de 2008 y comenzó a ser reconocido en 2010. Hoy en día, los datos masivos son dos palabras de moda en internet; se trata de buscar su aplicación, así como beneficios para todas las áreas del conocimiento humando y se usa en todo tipo de conferencias relacionadas al tema.

El creciente flujo de datos, que proviene de los diferentes tipos de sensores, sistemas de mensajería y redes sociales, además de sistemas de medición y observación más tradicionales, ya ha invadido muchos aspectos de la existencia cotidiana. Por un lado, los datos masivos (incluidos los geoespaciales), tienen un gran potencial para beneficiar y ampliar el conocimiento de muchas áreas de estudio como el cambio climático, la vigilancia de enfermedades, la respuesta a desastres, el transporte, el monitoreo de infraestructuras críticas, la logística, la accidentabilidad, etc. Por otra parte, los beneficios de los datos masivos para la sociedad suelen estar limitados por cuestiones necesarias como la privacidad de los datos, la confidencialidad y la seguridad, así como por la dificultad de obtener los conjuntos de datos y el costo asociado a su obtención.

Los datos masivos todavía no son un término claramente definido por lo que desde las diferentes perspectivas tecnológicas, industriales, de investigación o académicas se les da un significado diferente. En general, se consideran como conjuntos de datos estructurados y no estructurados con volúmenes de datos

[4] https://www.usenix.org/legacy/publications/library/proceedings/usenix99/invited_talks/mashey.pdf

Ver Anexo I. Antecedentes del Big Data (cronología)

masivos que no pueden ser fácilmente capturados, almacenados, manipulados, analizados, administrados y presentados por tecnologías tradicionales de hardware, software y bases de datos.

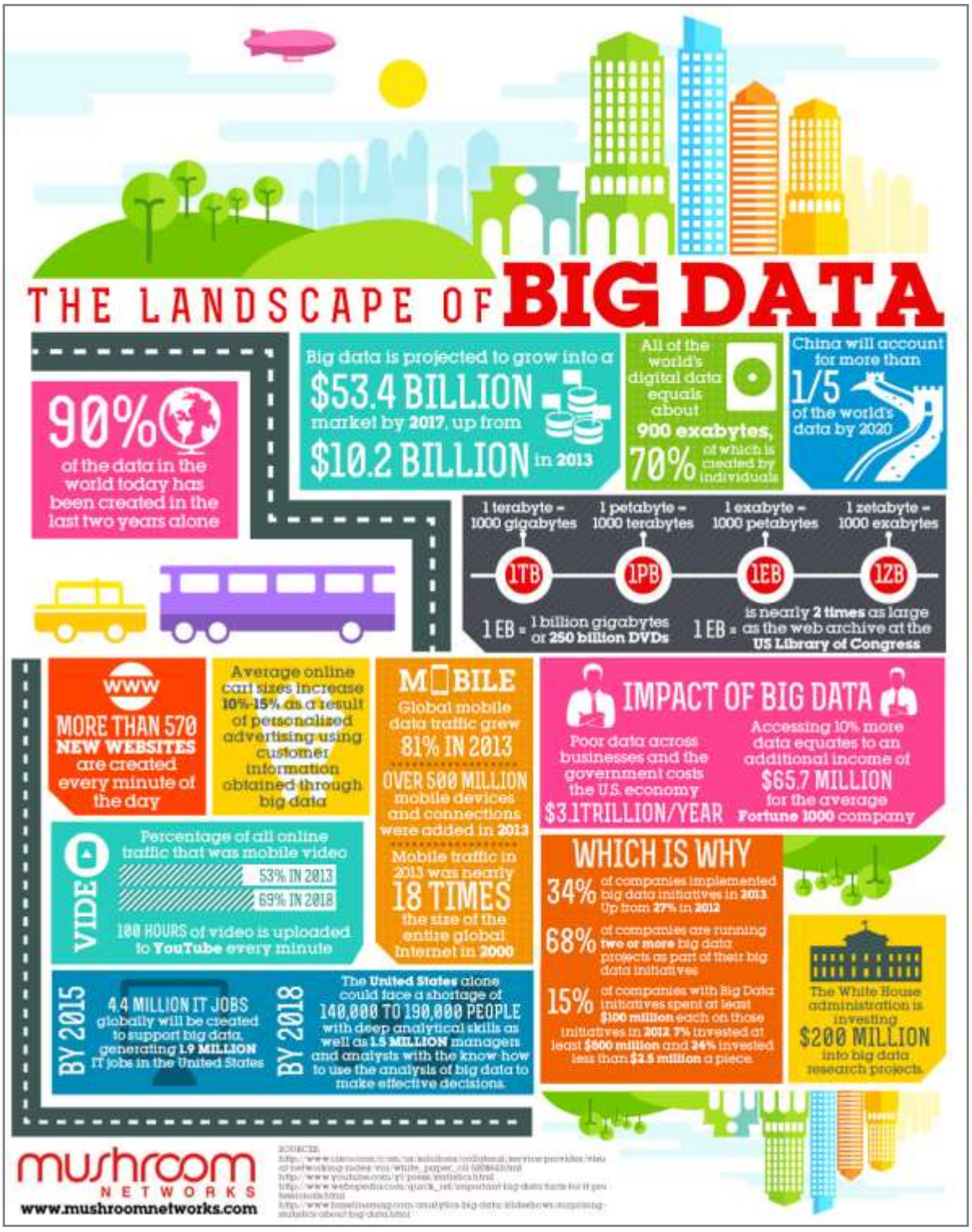
Junto con sus definiciones, los datos masivos se describen a menudo por sus características únicas. Laney (2001) propuso tres dimensiones que caracterizan los desafíos y oportunidades de aprovechar los datos masivos: Volumen, Velocidad y Variedad (3Vs)[5]. A la par de las 3Vs, la característica Veracidad se ha añadido para describir la integridad y calidad de los datos. También se han sugerido Vs adicionales tales como variabilidad, validez, volatilidad, visibilidad, valor o visualización y, debido a que no necesariamente expresan cualidades de magnitud, solo se mencionan en ciertas ocasiones y bajo ciertas condiciones. Aunque estos términos adicionales no ayudan a comprender a qué se refiere la palabra *"Big"* en los datos masivos, se incluyen conceptos importantes relacionados con el gran trabajo de recopilación de datos, procesamiento y presentación. Es evidente que la definición de los datos masivos y sus características es un esfuerzo continuo, sin embargo no tendrá un impacto negativo en el manejo y procesamiento de estos.

Gartner Inc. define el Big data como "el conjunto de información con alto volumen, alta velocidad y alta variedad que demanda innovadoras y efectivas formas de procesar la información para aumentar su valor para la toma de decisiones"[6].

En la *Figura 1.2 The Landscape of Big Data*, se puede observar algunos hechos relevantes acerca del tema, destaca el valor económico de los datos así como el volumen de crecimiento y el impacto que tiene en las empresas o instituciones: por ejemplo, el 90% de los datos actuales en el mundo se han creado en últimos dos años, los datos digitales a nivel mundial equivalen a 900 exabytes *(ver Figura 1.3 Escala del tamaño de los datos),* cada minuto se crean 750 nuevas páginas o *sitios,* más de 500 millones de dispositivos móviles y sus conexiones fueron agregados en 2013, el uso de pocos datos entre empresas y el gobierno le cuestan 3.1 trillones de dólares por año a la economía estadounidense.

[5] https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[6] http://www.gartner.com/it-glossary/big-data



Fuente https://www.mushroomnetworks.com/infographics/the-landscape-of-big-data-infographic

Figura 2.1 Visión del Big Data.

2.1.1 Volumen

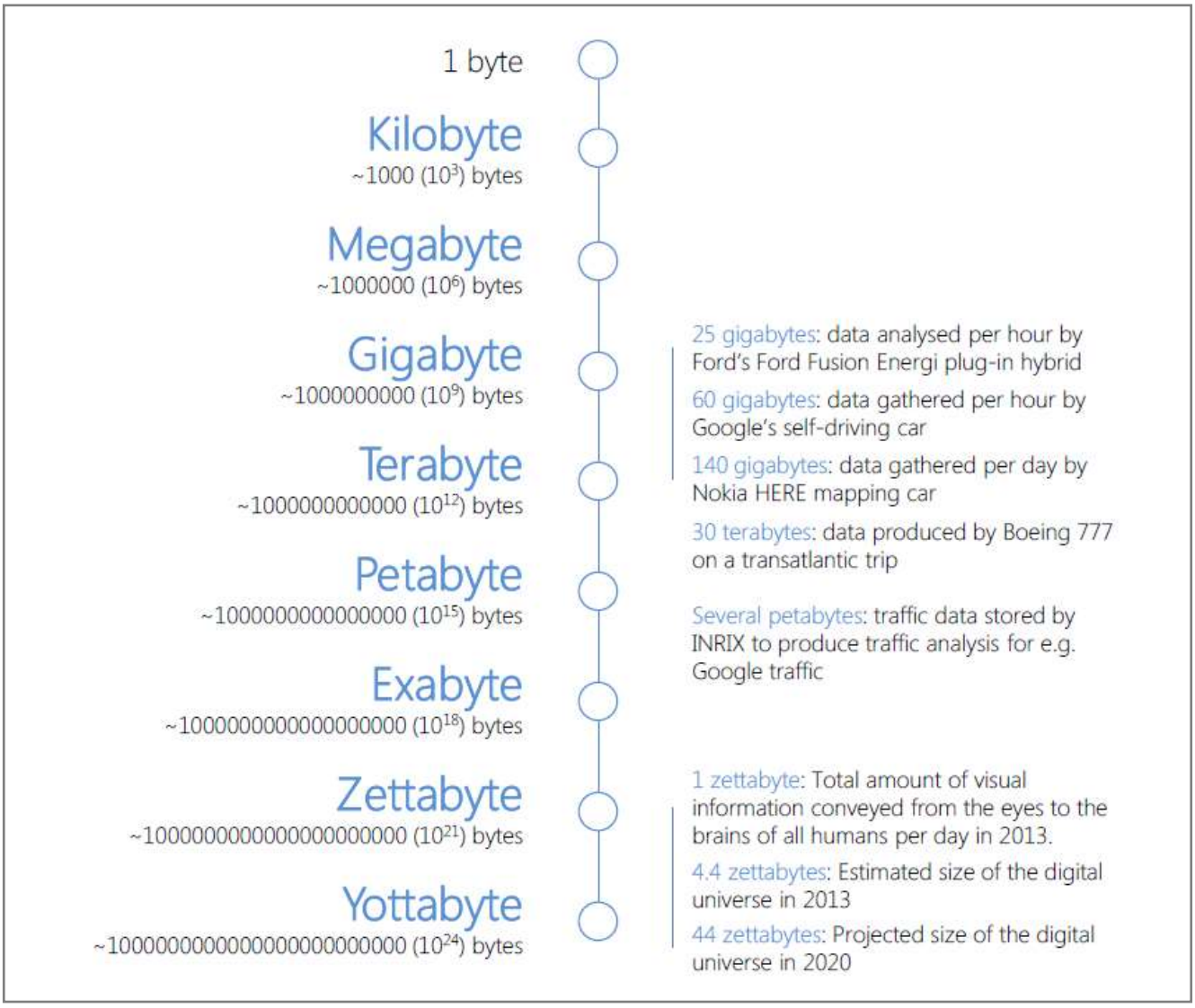
El volumen se refiere a la magnitud de los datos. El tamaño de los datos masivos se reporta en múltiples terabytes o petabytes (*ver Figura 3. Escala del tamaño de los datos)*. Muchos usuarios consideran que los datos masivos debe ser arriba de un terabyte (un terabyte equivale a 1500 discos compactos o 220 dvds. Un petabyte equivale a 1024 terabytes).

De acuerdo a Gandomi (2014), dos conjuntos de datos del mismo tamaño podrían requerir distintas tecnologías de proceso si se toma en cuenta su tipo, por jemplo, datos tabulares vs datos de video, por lo que las consideraciones acerca del volumen son imprácticas dependiendo de la fuente de los datos.

El reporte *Big Data and Transport*[7], se refiere una estimación del año 2013 que dimensiona el tamaño total del "universo digital" (entendiendo por contenido digital las fotografías, películas, videos de vigilancia, datos producidos por sensores y dispositivos asociados, contenido de internet, correo electrónico, mensajes cortos de celular, metadatos de llamadas telefónicas y archivos de audio) en 4.4 zettabytes[8] y se proyecta que el tamaño de este universo crecerá hasta 44 zettabytes (*ver Figura 3. Escala del tamaño de los datos)* en 2020. Estas estimaciones representan una cantidad asombrosa de datos y una porción significativa se refiere a eventos y personas (tarjetas de crédito y transacciones de pago, videos de vigilancia, salidas de sensores de vehículos, señales de acceso Wi-Fi, texto colaborativo e imágenes en redes sociales).

[7] Big Data and Transport. International Transport Forum. OECD. 2015

[8] Ibidem.



Fuente: Nokia HERE, Forbes, Idealab, GE, ITF

Figura 2.2 Escala del tamaño de los datos.

2.1.2 Velocidad

La velocidad se refiere a la rapidez con que los datos son generados, almacenados y procesados y de esta manera satisfacer la demanda de datos en tiempo real (Dasgupta, 2017). La proliferación de dispositivos digitales tales como los teléfonos "inteligentes" y otros sensores permiten ahora una rapidez nunca antes vista en la creación de datos y así se acrecienta la necesidad de análisis en tiempo real y planeación basada en evidencias. Los datos generados en estos dispositivos a través de diversas aplicaciones producen grandes cantidades de información que puede ser usada para satisfacer alguna necesidad de los usuarios en tiempo real. Estos datos proveen información acerca de consumidores, como la posición geoespacial, aspectos demográficos y patrones de consumo.

Los sistemas tradicionales de administración y manejo de datos no son capaces de almacenar, procesar y analizar grandes volúmenes de manera instantánea.

2.1.3 Variedad

La variedad se refiere a que los datos pueden provenir de fuentes heterogéneas o diversas. Los avances tecnológicos permiten la generación de datos estructurados, semiestructurados y sin estructura. Los datos estructurados, según Cukier (s/f), representan el 5% de los datos existentes y se refieren a los datos tabulares de hojas de cálculo y bases de datos relacionales. El texto, imágenes, audio y videos son ejemplos de datos sin estructura. Los que no tienen estructura y los semiestructurados no se ajustan a estándares estrictos. El formato XML (*eXtensible Markup Languaje* por sus siglas en inglés)[9] es un ejemplo de datos semiestructurados. Esta gran cantidad de datos, al provenir de distintas fuentes y con diferentes temporalidades se pueden insertar en alguno de los siguientes tipos de estructuras[10]:

- Datos estructurados (*Structured Data*): Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres. Se almacenan en tablas. Un ejemplo son las bases de datos relacionales y las hojas de cálculo.

[9] XML, siglas en inglés de *eXtensible Markup Language*, traducido como "Lenguaje de Marcado Extensible" o "Lenguaje de Marcas Extensible", es un meta-lenguaje que permite definir lenguajes de marcas desarrollado por el World Wide Web Consortium (W3C) utilizado para almacenar datos en forma legible. Proviene del lenguaje SGML y permite definir la gramática de lenguajes específicos (de la misma manera que HTML es a su vez un lenguaje definido por SGML) para estructurar documentos grandes. A diferencia de otros lenguajes, XML da soporte a bases de datos, siendo útil cuando varias aplicaciones deben comunicarse entre sí o integrar información. http://definicion.de/xml/

[10] http://consultec-ti.com/blog/big-data-analiza-toda-informacion-que-no-puede-ser-procesada-con-herramientas-tradicionales/

• Datos no estructurados (*Unstructured Data*): Datos en el formato tal y como fueron recolectados, carecen de un formato específico. No se pueden almacenar dentro de una tabla ya que no se puede separar su información a tipos básicos de datos. Algunos ejemplos son los PDF, documentos multimedia, e-mails o documentos de texto.

• Datos semiestructurados (*Semistructured Data*): Datos que no se limitan a campos determinados, pero que contiene marcadores para separar los diferentes elementos. Es una información poco regular como para ser gestionada de una forma estándar. Estos datos poseen sus propios metadatos semiestructurados que describen los objetos y las relaciones entre ellos y pueden acabar siendo aceptados por convención. Un ejemplo es el HTML, el XML o el JSON.

2.1.4 Veracidad

La veracidad se refiere a la fiabilidad de los datos inherente para ciertas fuentes. En ocasiones los datos tienen una veracidad baja y se debe buscar las fuentes más confiables. Por ejemplo, los sentimientos en las redes sociales es un aspecto difícil de obtener y de comprobar. Según un estudio de IBM, uno de cada tres directores o altos directivos aún no confían en los datos masivos ni en los beneficios que estos les podrían generar (IBM).

2.1.5 Viabilidad

La viabilidad se refiere a la capacidad que tiene el usuario de los datos para manejarlos, administrarlos, procesarlos y transformarlos en información útil para sus propios fines. Estos datos son generados por múltiples fuentes que tienen distintas configuraciones, formatos de salida, velocidades de proceso y se hace necesario conectar, limpiar, transformar y unir todos estos datos en un mismo conjunto.

2.1.6 Visualización

Se refiere a la generación de procedimientos para visualizar los grandes conjuntos de datos. Las visualizaciones permiten identificar patrones, tendencias y relaciones entre puntos y zonas específicas.

2.1.7 Valor

El valor se refiere a que generalmente los datos recibidos tienen poco valor sí se comparan con su volumen, sin embargo, al aplicarles procesos de análisis especializados se genera valor agregado al conjunto de datos.

No existen límites definidos en ninguna de las características mencionadas anteriormente para poder decir cuando un conjunto de datos entra en la categoría de Big Data, ya que estas son dependientes una de otra. Cuando una de las

características cambia, es probable que otra también lo haga y hasta una tercera lo hará también.

2.1.8 Análisis de texto

Gandomi (2014) refiere algunas técnicas para extraer información de datos en formato de texto. Los mensajes de redes sociales, correos electrónicos, blogs, foros en línea, encuestas, documentos corporativos y noticias son ejemplos de los textos que pueden ser utilizados. El análisis de texto involucra análisis estadístico, lingüística computacional, algoritmos con autoaprendizaje e inteligencia artificial. El análisis de texto permite convertir grandes volúmenes de texto en resúmenes entendibles los cuales dan evidencia para la toma de decisiones. Existen distintas variantes del análisis de texto, entre las cuales se identifican la extracción de información, la creación de resúmenes, respuestas a preguntas específicas, análisis de sentimientos.

2.1.9 Análisis de audio

El análisis de audio extrae información de los datos de audio sin estructura. Los centros de llamadas (*call centers*) utilizan el análisis en las llamadas que graban y que representan miles de horas de audio. Estas técnicas son capaces de identificar y evaluar la satisfacción del consumidor, evaluar el desempeño del operador, identificación de voz para cuestiones de seguridad, etc.

El análisis de audio cuenta con dos grandes aproximaciones tecnológicas: Aproximación basada en la transcripción (LVCSR *Large-Vocabulary Continuous Speech Recognition* por sus siglas en inglés) y la aproximación basada en fonética.

2.1.10 Análisis de video

Se conoce también como análisis de contenido de video e involucra una variedad de técnicas para monitorear, analizar y extraer información significativa de cadenas de video. La principal aplicación del análisis de video es la seguridad automatizada y los sistemas de vigilancia, esto sirve para detectar intrusos en áreas no autorizadas, identificar objetos que hayan sido robados y actividades sospechosas. Con más frecuencia se utiliza para identificar patrones de compra de consumidores y a través de algoritmos avanzados se pueden obtener datos demográficos de estos. Un área muy importante se refiere a la organización, indexación, fechado, identificación y almacenamiento de todos los videos que producen las cámaras de manera automática y que no han sido revisados por personal humano. Es necesario generar los metadatos que contiene datos como la fecha, el lugar, la hora, descripción del contenido, restricciones de visualización, la transcripción y el audio si existen.

2.1.11 Análisis de los medios sociales

El análisis de los medios sociales se refiere al análisis de datos estructurados y no estructurados de los canales de medios sociales. Los medios sociales abarcan una variedad de plataformas en línea que permiten a los usuarios crear e intercambiar contenido. Los medios sociales se pueden clasificar en los siguientes tipos de acuerdo con Barbier (2011):

- Redes sociales (por ejemplo, Facebook y LinkedIn)
- Blogs (por ejemplo, Blogger y WordPress)
- Microblogs (por ejemplo, Twitter y Tumblr)
- Noticias sociales (por ejemplo, Digg y Reddit)
- *Socialbookmarking* (por ejemplo, Wikipedia y Wikihow)
- Sitios de preguntas y respuestas (por ejemplo, Yahoo! Answers y Ask)
- Sitios de revisión (por ejemplo, Yelp , TripAdvisor)

Aunque la investigación sobre redes sociales se remonta a principios de 1920, la analítica de medios sociales es un campo nuevo que ha surgido después del advenimiento de la Web 2.0 a principios del año 2000. La característica clave de la analítica moderna de los medios sociales es su naturaleza centrada en los datos. La investigación en el análisis de los medios de comunicación social se extiende a través de varias disciplinas, incluyendo la psicología, la sociología, la antropología, la informática, las matemáticas, la física y la economía. El *marketing* ha sido la principal aplicación de la analítica de los medios sociales en los últimos años, esto puede atribuirse a la generalizada y creciente adopción de los medios de comunicación social por parte de los consumidores de todo el mundo. El contenido generado por el usuario (por ejemplo, sentimientos, imágenes, videos y etiquetas) y las relaciones e interacciones entre las entidades de red (por ejemplo, personas, organizaciones y productos) son las dos fuentes de información en las redes sociales. Basado en esta categorización, el análisis de los medios sociales se puede clasificar en dos grupos:

• Análisis basado en contenido. El análisis basado en contenido se centra en los datos publicados por los usuarios en las plataformas de medios sociales, como la retroalimentación de los clientes, las revisiones de productos, las imágenes y los videos. Tal contenido en las redes sociales es a menudo voluminoso, no estructurado, ruidoso y dinámico. Los análisis de texto, audio y video, como se indicó anteriormente, se pueden aplicar para obtener información de estos datos.

• Análisis basado en la estructura. También conocido como análisis de redes sociales, este tipo de analítica se ocupa de sintetizar los atributos estructurales de una red social y extraer inteligencia de las relaciones entre las entidades participantes. La estructura de una red social se modela a través de un conjunto de nodos y aristas, representando a participantes y relaciones, respectivamente. El modelo se puede visualizar como un gráfico compuesto por los nodos y los bordes. Se identifican dos tipos de gráficos de red, a saber, gráficos sociales y gráficos de actividad. En los gráficos sociales, una arista entre un par de nodos sólo significa la existencia de un enlace (por ejemplo, amistad) entre las entidades correspondientes. Tales gráficos se pueden extraer para identificar comunidades o determinar centros (es decir, los usuarios con un número relativamente grande de enlaces sociales directos e indirectos). En las redes de inactividad los bordes representan interacciones reales entre cualquier par de nodos. Las interacciones implican intercambios de información (por ejemplo, gustos y comentarios). Los gráficos de actividad son preferibles a los gráficos sociales, porque una relación activa es más relevante para el análisis que una simple conexión que puede tener meses sin ser usada, (Barbier 2011).

2.2 Datos masivos geoespaciales (*geospatial big data*)

Cada vez más, los conjuntos de datos con localización son de un tamaño, variedad y tasa de actualización que excede la capacidad de las tecnologías de computación enfocada a los procesos geoespaciales, estos datos se llaman *GeoSpatial Big Data* (GSBD por sus siglas en inglés) o *Datos Masivos Geoespaciales* (DMG), incluyen trayectorias de teléfonos celulares y dispositivos GPS, mediciones de motores de vehículos, mapas de carreteras detallados. La gran diversidad de fuentes de datos masivos geoespaciales aumenta sustancialmente la diversidad de métodos de solución. Los nuevos algoritmos pueden surgir a medida que se encuentren disponibles los conjuntos de datos masivos geoespaciales y de esta manera se crea la necesidad de una arquitectura flexible para integrar rápidamente nuevos conjuntos de datos y algoritmos asociados.

El componente geoespacial de los datos masivos geoespaciales puede venir en diversas formas tales como códigos postales, direcciones IP, localización geográfica (coordenadas x,y,z), nombres de calles o carreteras, recientemente se comienza a usar el análisis del contenido para identificar la ubicación exacta o lo más cercana posible de los datos o eventos relacionados.

De acuerdo con Muñiz (2014), "el 80% de los datos tienen una componente geográfica", esto quiere decir que gran parte de los datos en el mundo pueden ser

georreferenciados y también indica la importancia del manejo geoespacial de los datos masivos. Los datos geoespaciales describen objetos y cosas con relación al espacio geográfico, con coordenadas de ubicación en un sistema de referencia espacial. Estos datos tradicionalmente se recopilan mediante la topografía terrestre, la fotogrametría, la teledetección y más recientemente, mediante el escaneo láser, la cartografía móvil, los contenidos geo-etiquetados, la información geográfica participativa o colaborativa, los sistemas mundiales de navegación por satélite y sensores geoposicionados. Los datos geoespaciales pueden presentar al menos una de las 3Vs pero las otras Vs mencionadas anteriormente también son relevantes.

El volumen creciente y el formato variable de los datos masivos geoespaciales recolectados plantean retos adicionales en cuanto al almacenamiento, gestión, procesamiento, análisis, visualización y verificación de la calidad de los datos. Shekhar (2012) afirma que "el tamaño, la variedad y la tasa de actualización de los conjuntos de datos exceden la capacidad de las tecnologías de computación espacial y de bases de datos espaciales comúnmente utilizadas para aprender, gestionar y procesar los datos con un esfuerzo razonable". La comprobación de la calidad de los datos masivos geoespaciales y de los productos entregados a los usuarios finales se observa como uno de los grandes desafíos.

Los datos geoespaciales, que son abstracciones y observaciones de una realidad continua, son por naturaleza inciertos, idealmente estampados en el tiempo y a menudo incompletos. En consecuencia, los datos grandes geoespaciales, con sus características definitorias de ser grande (voluminoso), heterogéneo (variedad), procesado en tiempo real (velocidad), inconsistente (variabilidad) y, por tanto, también de calidad variable (veracidad), deben sufrir incertidumbre, desfase y fragmentación. Sin embargo, si bien ciertos efectos sobre la calidad de los datos se enfatizan para los grandes datos geoespaciales, los fenómenos que se describen siguen siendo los mismos.

2.2.1 Fuentes de datos masivos geoespaciales

Desde Google Maps hasta los dispositivos GPS (*Global Positioning System*), la sociedad se ha beneficiado enormemente de los servicios de posicionamiento y localización. Cada vez más, sin embargo, el tamaño, la variedad y la tasa de actualización de los conjuntos de datos exceden la capacidad de las tecnologías de computación geoespacial y de bases de datos geoespaciales comúnmente utilizadas para aprender, administrar y procesar los datos con un esfuerzo razonable.

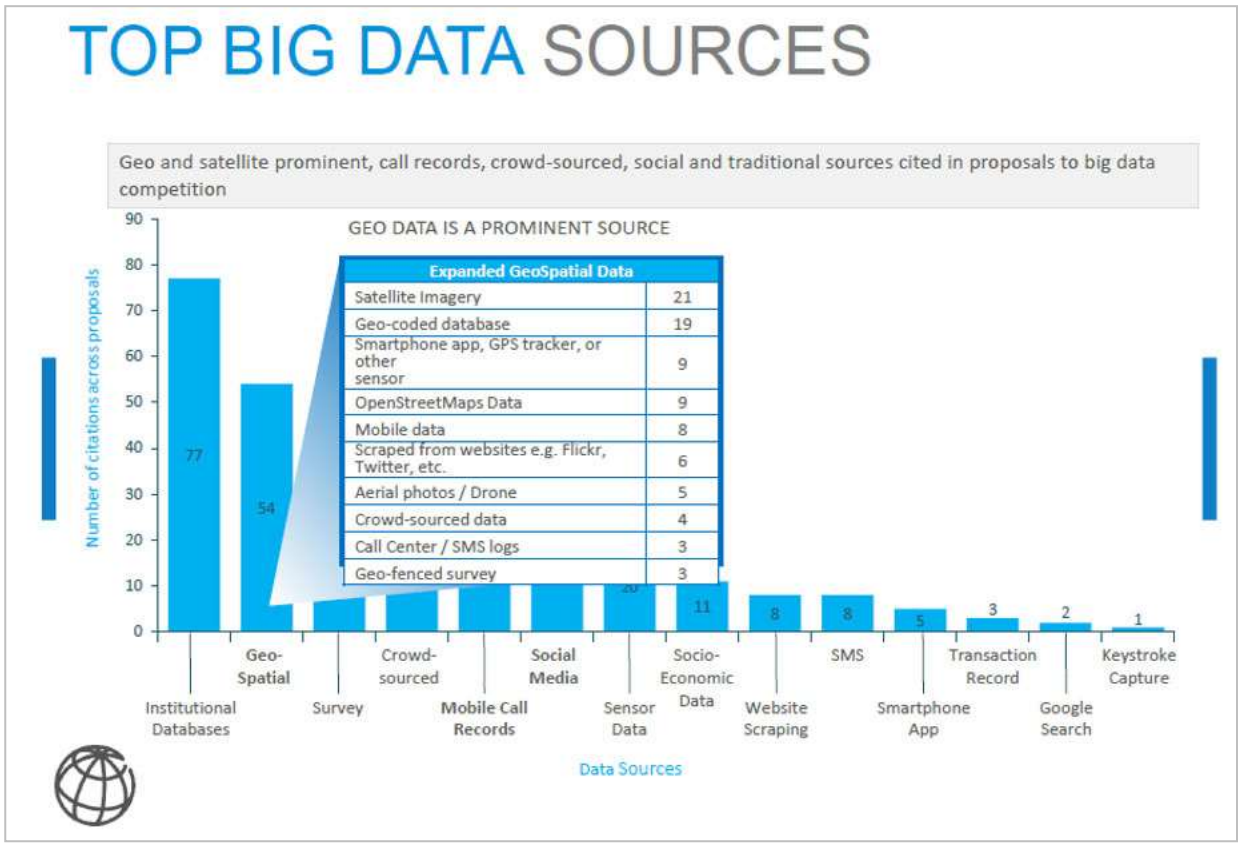
Por ejemplo, los conjuntos de datos que proporcionan velocidades cada minuto para cada segmento de carretera, datos de rastreo GPS de teléfonos celulares, mediciones del consumo de combustible de un motor de acuerdo al tipo de terreno y altitud, emisiones de gases, fotos georrefenciadas tomadas con una cámara de

un dron, barridos LiDAR, avisos de congestionamientos a través de plataformas colaborativas como Waze, sucesos y “sentimientos” geolocalizados a través de Twitter, Facebook, colecciones de fotos como Instagram, rastreo de mensajes de texto y llamadas, todos demuestran la gran cantidad de fuentes de datos masivos geoespaciales con posibilidad de ser almacenados, procesados y analizados.

Cuando se combinan, estos datos revelan patrones hasta ahora insospechados o no observados en la vida cotidiana que pueden beneficiar tanto a los individuos como a la sociedad. Existe también el riesgo de que los patrones o tendencias puedan abrir nuevas vías para el mal uso de los datos y la manipulación potencial de los individuos y su comportamiento. Es así que los datos masivos pueden ser vistos como una oportunidad y un desafío al mismo tiempo.

De acuerdo con la *Figura 1.4. Fuentes de los datos masivos geoespaciales*, algunas de éstas pueden ser:

- Imágenes de satélite.
- Datos existentes (BD, texto, video, fotos).
- Dispositivos móviles.
- GPS (*Global Positioning System*).
- Sensores (llantas, motores, RFID, tanques, válvulas).
- Internet de las cosas.
- Plataformas específicas (*waze, twitter, openstreetmaps, google maps*).
- Radar.
- LiDAR.
- Foto y video (ambos con geoposición) que proviene de sensores colocados sobre drones.
- Crowsourcing.



Fuente http://opendatacon.org/aiming-right-human-scale-measure-sustainable-development-goals/

Figura 2.3 Fuentes de los datos masivos geoespaciales.

2.3 Los datos masivos geoespaciales y el transporte

De acuerdo con la 1ª Ley de Tobler[11] con respecto a la geografía y que establece "*todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes*", entonces, el transporte (en su expresión más básica) se trata simplemente de conectar ubicaciones y luego relacionarlas con flujos. Estas ubicaciones pueden ser cercanas, estar bien conectadas y mostrar altos niveles de acceso, como en áreas urbanas. Los flujos entre estos lugares pueden referirse a personas o bienes y pueden referirse al número de vehículos, tipo, frecuencia, capacidad, etc.
Los dispositivos y redes móviles en los hogares, escuelas, espacios recreativos y lugares de trabajo están conectados a Internet, proporcionando accesibilidad y

[11] Waldo Tobler. 1930. Geógrafo.

seguimiento, incluso los viajes están conectados, de hecho, la industria del transporte es un líder en la creación de Internet de las Cosas (*IoT Internet of Things*, por sus siglas en inglés), al generar grandes volúmenes de datos cada día a través de sensores en los sistemas de conteo de pasajeros, localización y conteo de vehículos, sistemas de pago en carreteras, videos de vigilancia, recolección de características de las superficies de rodamiento de carreteras. Con el transcurso del tiempo, sin embargo, esos terabytes de datos recopilados han añadido complejidad a las operaciones de la institución y consumen cantidades inmensas de almacenamiento en los centros de datos.

Los sistemas de transporte que hacen uso de la información de las cámaras y microcontroladores para optimizar el flujo de transito público, vigilar el medio ambiente y ejecutar acciones de seguridad son conocidos como sistemas de transporte inteligente (ITS). En general, la información para ITS se extrae a partir de dos tipos de redes de sensores personalizados, sensores fijos y sensores móviles, (Calabrese 2013).

Existen aplicaciones para el transporte inteligente, mediante la visualización y análisis del uso en tiempo real de las redes de transporte y seguridad mediante el procesamiento de datos en tiempo real sobre el funcionamiento del vehículo y su entorno para evitar o minimizar conflictos potencialmente peligrosos.

La exactitud de localización derivada de estos métodos varía en precisión, confiabilidad y puntualidad. Algunos de los datos pueden ser imprecisos pero producidos en tiempo real (con un retraso de milisegundos a segundos), algunos pueden ser muy precisos pero sólo disponibles ex-post a través de los registros de la máquina. Gran parte de los datos de localización producidos por dispositivos personales o sistemas embebidos en vehículos, sin embargo, son precisos y entregados casi en tiempo real. Es probable que los avances en las arquitecturas de los sensores aumenten la cantidad de datos con ubicación, al igual que la capacidad de procesar los datos capturados se hace menos costosa.

Los datos masivos geoespaciales se vislumbran como muy prometedores para mejorar la planificación y la gestión de la actividad del transporte aumentando radicalmente la disponibilidad de datos relacionados con la movilidad casi en tiempo real, para esto se requiere contar con los datos de la infraestructura, tales como carreteras, puentes, vías, aeropuertos, puertos, estaciones de autobuses. Asimismo, el acceso a datos más detallados y en relación con el funcionamiento de los vehículos y del medio ambiente en el que operan, mejorará la seguridad del transporte. Estos tres campos - *operaciones, planificación y seguridad* - son áreas en las que las autoridades deben evaluar críticamente dónde y cómo los datos disponibles y las perspectivas relacionadas, pueden mejorar la política de transporte.

Las autoridades de transporte disponen hoy de herramientas de información y de planificación avanzada y cada una tiene un propósito específico; sin embargo, mediante la fusión de estas fuentes de datos se puede crear una imagen mucho

más completa de la dinámica y los factores que contribuyen a la eficacia de la red y, en consecuencia, la satisfacción (en todos sus aspectos) de los usuarios finales.

- El gran volumen de los datos masivos geoespaciales se debe a la gran cantidad de agentes (vehículos, personas y bienes) que están en camino en cualquier momento. Además, para poder predecir la demanda o el tráfico del transporte, no sólo se requieren datos en tiempo real sino también datos históricos.

- La gran variedad de los datos masivos geoespaciales en el transporte (como GNSS[12], sensores inerciales, compás, sensores de rueda, radar, escaneado láser, reconocimiento de matrículas, ciclos de inducción, peaje electrónico o boletaje, sensores de estacionamiento) que se combinan en grandes análisis de datos

- La variabilidad de los datos masivos geoespaciales en el transporte proviene tanto de la variedad entre los diferentes canales de datos como de la falta de fiabilidad de cualquiera de estas fuentes. Por ejemplo, la información geográfica voluntaria puede ser inconsistente si los "ciudadanos como sensores" (Goodchild) o "ciudadanos como bases de datos" (Richter & Winter) no están de acuerdo con sus observaciones, o están *discapacitados* por la falta de canales de comunicación (por ejemplo, un sensor se está moviendo fuera del alcance de la cobertura WiFi / teléfono celular). En otro ejemplo, el posicionamiento por satélite tiene algunos descriptores de calidad bien conocidos, pero entre las calles de una ciudad estas medidas varían con la ubicación del sensor, el canal de transmisión o el tiempo de transmisión.

- La veracidad de los datos masivos geoespaciales en el transporte indica que los datos son de distintas calidades. Esto se extiende también a tasas de muestreo irregulares (espaciales y temporales), errores de entrada, redundancia, corrupción, falta de sincronización o una variedad de propósitos de recolección (criterios de diseño de las bases de datos, semántica).

- Dado que la gran recopilación de datos requiere conectividad, también se requieren medidas para tratar con contribuciones maliciosas, ataques y robo, así como la privacidad. Esto último es particularmente cierto para los datos geoespaciales recolectados para los movimientos de seguimiento en tiempo real de personas y vehículos.

- Típicamente, el valor de los datos masivos geoespaciales (análisis) se ve en la información para el apoyo a la toma de decisiones. Los métodos analíticos tales como la minería de datos y el aprendizaje automático permiten solamente el

[12] GNSS(Global Navigation Satellite System), es el acrónimo que se refiere al conjunto de tecnologías de sistemas de navegación por satélite que proveen de posicionamiento geoespacial con cobertura global de manera autónoma.

razonamiento inductivo en datos grandes, es decir, la detección de correlaciones globales o predicciones basadas en estas correlaciones. En el transporte, un ejemplo es el descubrimiento de patrones en la ocurrencia de "embotellamientos" así como de accidentes de tráfico.

Algunas de las áreas de aplicación, que se pueden relacionar al transporte son:

- Tendencias sociales (*twitter, facebook*) y colaboración participativa: mapas, levantamiento de información, encuestas de origen-destino
- Medición de datos espacio-temporales para identificación de patrones: logística, percepción remota, dirección del viento, precipitación, oleaje
- Seguimiento de datos GPS: movilidad, accesibilidad, cálculo de rutas alternas
- Perfiles históricos y de predicción: contaminación, velocidades, causa de embotellamientos, rendimiento de maquinaria
- Identificación de relaciones topológicas: nubes de puntos masivas, optimización y planeación del transporte

2.4 Proceso de los datos masivos geoespaciales

Los datos masivos - en el transporte - han surgido de la convergencia de los costos decrecientes de recolección, almacenamiento y procesamiento, así como de la difusión de los datos. La disminución de los costos de los sensores ha llevado a una proliferación de plataformas de detección que transforman grandes extensiones del mundo analógico en señales procesadas digitalmente. La reducción de los costos de almacenamiento de datos ha permitido retener los datos que se habían descartado anteriormente. Como lo señaló el historiador de ciencias Dyson (S/F), "Big Data es lo que sucedió cuando el costo de almacenar la información fue menor que el costo de tomar la decisión de tirarlo".

Para que las organizaciones hagan un uso eficiente de los datos requieren procesos que les permitan extraer y generar información correcta a partir de los datos masivos. Labrinidis & Jagadish (2015), han identificado cinco fases principales agrupadas en dos grandes procesos *(Figura 5. Procesos de los datos masivos).*

Traducción y adaptación a partir de http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf

Figura 2.4 Procesos de los datos masivos.

Administración de datos

De acuerdo con la *Figura 1.5 Procesos de los datos masivos, e*l proceso de administración de datos se refiere a subprocesos y tecnologías de soporte que adquieren y almacenan los datos y los preparan para recuperarlos para su análisis.

Adquisición y almacenamiento

En los últimos años, junto con la disponibilidad de nuevos sensores, han surgido nuevas formas de recolección de datos geoespaciales que han dado lugar a fuentes de datos completamente nuevas y tipos de datos de naturaleza geográfica. Los datos adquiridos por el público, llamados *información geográfica voluntaria* (*Volunteered Geographic Information* VGI, por sus siglas en inglés) y los datos de las redes de sensor con geoposición han llevado a una mayor disponibilidad de información espacial. Mientras, hasta hace poco tiempo, los conjuntos de datos generados por una autoridad eran la mayoría, ahora estos nuevos tipos de datos

vienen a complementar, ampliar y enriquecer el conjunto de datos geográficos en términos de variación temática y por el hecho de que están más centrados en el usuario. Esto último es especialmente cierto para los datos recopilados a través de medios sociales, (Li 2016).

La recopilación de datos geoespaciales está evolucionando de ausencia de datos a un panorama con gran disponibilidad de estos. Mientras que hace unos años la captura de datos geoespaciales se basaba en dispositivos técnicamente exigentes, precisos, caros y complicados, ahora la adquisición de datos geoespaciales es un proceso implementado en dispositivos móviles de uso cotidiano para mucha gente. Estos dispositivos son capaces de adquirir información geoespacial a un nivel sin precedentes, con respecto a la resolución geométrica, temporal y temática; son pequeños, fáciles de manejar y capaces de adquirir datos incluso sin que el usuario se percate.

En general, se pueden distinguir las siguientes configuraciones de sensores:

1) Objetos equipados con sensores que se mueven a través del espacio y captan sus propias trayectorias y en algunos casos, los entornos locales, por ejemplo, automóviles

2) Sensores estáticos que observan constantemente el ambiente y graban alguna característica, por ejemplo, estaciones meteorológicas, contadores de vehículos

En la última década, con el surgimiento del concepto de "ciudad inteligente", se ha instalado sensores que proporcionan gran cantidad de datos en tiempo real a través de distintos sistemas como los lectores de tarjetas inteligentes, dispositivos de rastreo de vehículos, circuitos cerrados de televisión y vigilancia, sistemas de peaje y otros sensores. Con el aumento de la comunicación a través de las redes sociales también se generan grandes cantidades de datos; por ejemplo, los mensajes cortos de Twitter, que pueden estar *geolocalizados* y se podrían utilizar para ayudar en la gestión de desastres y socorro en situaciones de emergencia. Asimismo existen enormes repositorios de imágenes de satélite que se utilizan para detección remota.

Las personas ayudan a capturar *información geográfica voluntaria* relacionada con el tráfico o la movilidad. La adquisición de datos se debe a un proceso consciente de un usuario, que explícitamente selecciona objetos, observa sus características y aporta esta información, generalmente a través de una plataforma, por ejemplo *Waze y OpenStreetMap*. En el caso de *Waze*, la captura de los puntos de congestionamiento y los distintos tipos de eventos son sólo subproductos que generan los usuarios voluntariamente y por otro lado, el principal producto es todo el conjunto de lecturas GPS que se registran sin que el usuario intervenga y se envían del teléfono celular a los servidores de la plataforma, en donde se procesan y al final se generan mapas dinámicos que muestran las calles, avenidas o carreteras a partir de la congestión vial que presenta cada uno de ellos.

En cuanto al almacenamiento se requieren equipos que cumplan con las condiciones requeridas para efectuar tal función, estos equipos pueden ser locales

o también se puede utilizar el almacenamiento en la nube aunque para las dos vertientes se hace necesario implementar mecanismos de organización que permita almacenar cada conjunto de datos en una forma lógica y correcta de tal manera que cuando se requiera utilizarlos sea fácil identificarlos, descargarlos y procesarlos con los métodos establecidos para ello. Es indispensable la creación de los metadatos de cada uno de los conjuntos de datos, en donde se describen las características principales de estos.

También es factible la utilización de Unidades de Estado Sólido (SSD) que son dispositivos que sirven para almacenar datos en una computadora.

Los discos duros en estado sólido no disponen de partes mecánicas móviles, y ofrecen una velocidad de lectura y escritura superior a los discos duros tradicionales. En las unidades SSD no hay ningún plato girando a velocidad constante, ni brazos moviéndose sobre él, sino que emplean unidades de memoria *flash*, similares a las de los *pendrives* o memorias USB, pero con una capacidad de almacenamiento mayor. Las unidades de memoria utilizadas son no volátiles y siguen manteniendo los datos después de una pérdida de potencia.

Extracción, limpieza y anotación

Más allá de las cuestiones de disponibilidad y costos de recolección, un factor importante a considerar al seleccionar una fuente de datos es su aptitud para el análisis. Las técnicas de extracción se refieren a todas las formas en que la información se extrae de un conjunto de datos determinado. Una vez validados los campos que dan origen a un identificador único y a la geoposición (por ejemplo, el nombre, el origen y el destino, la longitud y la latitud), se puede realizar una serie de operaciones para limpiar, transformar y modelar los datos en busca de conclusiones significativas.

Se han desarrollado o adaptado una gama de técnicas y herramientas para agrupar, manipular, limpiar y extraer datos masivos. Estos se basan en la experiencia de distintas áreas del conocimiento como la estadística, informática, matemáticas aplicadas y economía.

En el contexto de la planificación del transporte, se suele extraer las propiedades topológicas, geométricas o geográficas codificadas en un conjunto de datos y las técnicas pueden agruparse en las siguientes categorías, sin limitarse exclusivamente a ellas, (Manyika, 2011*)*:

- Minería de datos[13]: técnicas para extraer patrones de grandes conjuntos de datos, tales como las relaciones entre nodos discretos en una red de transporte.

- Optimización: técnicas para reorganizar sistemas y procesos complejos para mejorar su desempeño de acuerdo con uno o más parámetros, tales como tiempo de viaje o eficiencia de combustible.

Un factor importante a considerar al seleccionar una fuente de datos es el alcance y la calidad del conjunto de datos resultante. Los datos extraídos de una sola fuente generalmente se consideran limpios y precisos. La realidad es que los datos suelen ser "desordenados", ya que son heterogéneos, "sucios" (esto incluye datos incorrectos, mal etiquetados, faltantes o potencialmente falsos) y en su formato nativo, son incompatibles con otras fuentes de datos. Parte del desafío reside en el hecho de que algunos datos pueden estar altamente estructurados (por ejemplo, datos de latitud y longitud GPS y datos de transacciones comerciales) que facilitan un análisis rápido mientras que otros datos pueden incluir conjuntos de datos altamente desestructurados (correos electrónicos, contenido de medios sociales, audio) y por lo tanto será más difícil y tomará más tiempo analizarlos. Los avances en las técnicas de procesamiento y análisis de datos permiten mezclar datos estructurados y no estructurados con el fin de obtener nuevos conocimientos, pero esto requiere datos "limpios".

La limpieza de datos y su preparación para el uso analítico es una tarea no trivial que puede implicar un costo significativo. Los datos estructurados deben ser analizados y los faltantes o datos potencialmente incorrectos contabilizados. Los datos no estructurados deben ser correctamente interpretados, categorizados y etiquetados consistentemente.

En algunos casos, la "*edición manual de datos*" sigue siendo un componente necesario del flujo de recolección y análisis de datos. Este trabajo requiere una gran inversión en tiempo y recursos que representan entre el 50% y el 80% del tiempo de los "*científicos de datos*" según algunas estimaciones de Lohr (2014). La preparación de los datos puede facilitarse en gran medida mediante el uso de algoritmos compensatorios apropiados, pero la elección del algoritmo puede

[13] El *datamining* (minería de datos), es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.
Básicamente, el *datamining* surge para intentar ayudar a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales.
De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento.
http://www.sinnexus.com/business_intelligence/datamining.aspx

conducir a errores de imputación o predicción si interpreta incorrectamente los valores faltantes o minimiza los valores atípicos que podrían ser importantes.

Integración, agregación, representación y fusión

La integración de datos es el proceso que permite combinar datos heterogéneos de muchas fuentes diferentes en la forma y estructura de una única aplicación. Este proceso de integración de datos facilita que sus diferentes tipos, tales como matrices de datos, documentos y tablas, sean fusionados por usuarios, organizaciones y aplicaciones para un uso personal, de procesos de negocio o de investigación.

La integración de datos soporta el procesamiento analítico de grandes conjuntos de datos alineando, combinando y presentando cada conjunto de datos y fuentes remotas y externas, para cumplir con los objetivos del integrador.

La integración de datos se implementa generalmente en un almacén de datos (*data-warehouse*) mediante software especializado que aloja grandes repositorios de datos de recursos internos y externos. Los datos se extraen, se mezclan y se presentan de forma unificada. Por ejemplo, el conjunto completo de datos de un usuario puede incluir datos extraídos y combinados de marketing, ventas y operaciones, que se combinan para formar un informe completo[14].

Un proyecto de integración de datos generalmente implica los siguientes pasos:

- Acceso a los datos desde todas las fuentes y localizaciones, tanto si se trata de locales, en la nube o de una combinación de ambos.

- En la integración de datos existe una técnica en donde los registros de una fuente de datos mapean registros en otra. Se trata de un tipo de preparación de datos esencial para que las analíticas y otras aplicaciones sean capaces de utilizar los datos con éxito.

- Para resolver un problema se necesita conocimiento específico sobre él y también se requiere representar este conocimiento de alguna manera útil y eficaz. La forma de representar dicho conocimiento dependerá del problema a analizar y se deberá encontrar la mejor manera de representarlo.

Entre las técnicas de representación de datos se encuentran:

[14] http://www.powerdata.es/integracion-de-datos

- Representación mediante tuplas[15]: (objeto, atributo, valor)
- Representación mediante redes semánticas[16]
- Representación mediante tablas
- Representación mediante *frames*[17]
- Representación mediante reglas: IF condición THEN
- Representación mediante lógica de predicados
- Representación mediante modelos lineales
- Representación mediante árboles

En el caso de agregación de datos, los conjuntos de datos se emparejan y se combinan con base en atributos y variables compartidos, pero se conserva el conjunto de cada conjunto de datos independiente. Este método es adecuado para incrementar el descubrimiento de conocimiento a través del análisis de datos contextuales.

La fusión de datos es un paso especialmente importante en el uso de entradas de múltiples plataformas de sensores. Por ejemplo, los algoritmos de fusión de datos ayudan a procesar entradas de sensores de movimiento, acelerómetros, magnetómetros, datos de señal celular, cámaras, escáneres láser y chips GPS. Todas estas fuentes de datos contribuyen a crear, por ejemplo, una representación precisa de la ubicación de un auto en una calle, donde dicha fusión de datos es necesaria para el desarrollo de vehículos de conducción autónomos. Aquí, la representación final del vehículo (la ubicación exacta, el tamaño, la dirección y velocidad) es todo lo que se retiene, eliminando así la necesidad de almacenar cada flujo de datos individuales de cada sensor. El aspecto desafiante de la fusión de datos está en la extracción de características sobresalientes de múltiples conjuntos de datos generados para diferentes usos.

En general, las técnicas de fusión de datos buscan integrar precisión y semántica. Los datos recolectados de fuentes múltiples suelen contener incoherencias en

[15] Una tupla es una secuencia de valores agrupados. Sirve para agrupar, como si fueran un único valor, varios valores que, por su naturaleza, deben ir juntos. No puede ser modificada una vez que ha sido creada. http://progra.usm.cl/apunte/materia/tuplas.html

[16] Se denomina red semántica al esquema que permite representar, a través de un gráfico, cómo se interrelacionan las palabras. http://definicion.de/red-semantica/

[17] Los sistemas de *frames* razonan acerca de clases de objetos usando representaciones prototípicas, pero que pueden modificarse para capturar las complejidades del mundo real. https://ccc.inaoep.mx/~esucar/Clases-MetIA/MetIA-07.pdf

términos de resolución. Por ejemplo, hay GPS que registran una posición cada segundo mientras que algunas estaciones base lo hacen a cada 15 segundos. La semántica se refiere al sujeto que está siendo representado. En un caso, los datos rastrean un vehículo y la otra vertiente se refiere al usuario del auto y con quién se comunica.

Para esto, se sugiere la siguiente lista de atributos que debe cumplir un paquete de datos:

- Localización (sistema de coordenadas, latitud y longitud) del dispositivo transmisor o acción reportada.
- Tiempo de transmisión de datos o acción reportada.
- Categoría de datos (datos basados en la ubicación del teléfono móvil, datos GPS, boletín de noticias).
- Formato de datos (valor único, matriz, vector, texto, imagen, etc.)
- Representación de datos (por ejemplo, unidad de medida).
- Semántica de los datos (por ejemplo, vehículo de seguimiento o teléfono móvil).

Análisis

De acuerdo con la *Figura 1.5 Procesos de los datos masivos,* el proceso de análisis se refiere a las técnicas usadas para analizar y generar valor agregado e identificar patrones derivados de los datos masivos.

Modelado y análisis

Construir y ejecutar modelos ayuda a probar hipótesis sobre el impacto y la importancia de diferentes variables en los sistemas del mundo real. Al simplificar la simulación de fenómenos del mundo real, los modelos ayudan a caracterizar, comprender, cuantificar y visualizar relaciones que son difíciles de comprender en sistemas complejos. Los modelos de construcción requieren datos sobre las condiciones de referencia y la comprensión de la naturaleza de las relaciones, correlativas o causales, entre múltiples fenómenos. La llegada de los datos masivos ha aumentado drásticamente la escala, el alcance y la accesibilidad de los ejercicios de modelado, aunque debe tenerse en cuenta que la capacidad de estos para seguir con precisión el mundo real está vinculada no sólo, o principalmente, a la cantidad y la calidad de los datos de referencia. La construcción del modelo y la garantía de que las preguntas correctas se formulan siguen siendo esenciales para proporcionar resultados de alto valor. Los modelos bien construidos basados en datos escasos pueden ser tan o más eficaces que los modelos mal diseñados que trabajan en conjuntos masivos de datos en tiempo real. Con esta advertencia en mente, las fuentes y las técnicas de datos masivos han permitido construir nuevos modelos que proporcionen nuevas preguntas y nuevas perspectivas.

Estructuración espacial de datos

Todos los modelos de datos espaciales, incluyendo el modelo de datos vectoriales tipo espagueti[18], el modelo de datos de red, el modelo de datos topológicos y el modelo regular (ráster) e irregular (Voronoi[19], árbol kd[20], árbol de partición binario[21]), se pueden utilizar para manejar datos masivos geoespaciales. Sin embargo, existen algunos modelos que son más adecuados para manejar conjuntos de datos muy grandes y otros que son menos adecuados para datos masivos geoespaciales. Como la red y los modelos de datos topológicos necesitan almacenar la conectividad y la adyacencia, no son adecuados para manejar grandes flujos de datos geoespaciales a menos que un algoritmo muy eficiente realice la indexación de datos así como la actualización de la conectividad y/o la topología en tiempo real. En tales casos, la única manera posible de satisfacer las necesidades de proceso en tiempo real es utilizar un método de indexación de datos espaciales que pueda mantener sus prestaciones con un gran flujo de datos.

Los actuales métodos de indexación de datos espaciales no pueden manejar grandes flujos de datos geoespaciales porque su eficiencia se reduce a medida que los nuevos flujos de datos espaciales superan la capacidad de extensión del índice de datos espaciales. La geoestadística es adecuada para manejar grandes datos. Ofrece oportunidades para resumir los datos y expresar medidas de variación e incertidumbre. La gran preocupación, sin embargo, es que muchos de los procesos y procedimientos se desarrollan para los conjuntos de datos más pequeños. En particular, gran parte del análisis estadístico espacial se realiza en conjuntos de datos que se recogen en una escala puntual (como datos de campo, datos meteorológicos o datos administrativos) y de un contenido relativamente pequeño, o se centra en los conjuntos de datos de imagen relativamente grandes que tienen un carácter muy específico. La estadística espacial depende de la noción de dependencia espacial (y espaciotemporal) y tal dependencia depende a su vez, de la noción de distancia entre puntos. Las estructuras de datos actuales como tales suelen ser capaces de manejar los grandes datos también, pero lo más probable es

[18] Almacena información de los puntos como pares de coordenadas, las líneas como una sucesión de pares de coordenadas y los polígonos cadena de pares de coordenadas con repetición del primer par de coordenadas que indica que es un elemento cerrado.

[19] El diagrama de Voronoi de un conjunto de puntos en el plano es la división de dicho plano en regiones, de tal forma, que a cada punto le asigna una región del plano formada por los puntos que son más cercanos a él que a ninguno de los otros objetos. http://www.abc.es/ciencia/abci-diagrama-voronoi-forma-matematica-dividir-mundo-201704241101_noticia.html

[20] En ciencias de la computación, un Árbol kd (abreviatura de árbol k-dimensional) es una estructura de datos de particionado del espacio que organiza los puntos en un Espacio euclídeo de k dimensiones. http://es.unionpedia.org/%C3%81rbol_kd

[21] Es una estructura de datos usados para organizar objetos dentro de un espacio. Dentro del campo de gráfica de computadores, tiene aplicaciones en la remoción de áreas ocultas y en el trazado de rayos. http://www.symbolcraft.com/products/bsptrees/spanish/

que se desarrollen procedimientos específicos que sean capaces de resolver problemas relativamente novedosos (como combinar datos en el dominio espacio-tiempo) o que tienen que abordar preguntas y problemas específicos, es decir, seleccionar datos de un conjunto de datos grande para una aplicación de modelo particular.

Algunos autores como Lee (S/F) y Shekhar (2012), ya han señalado la necesidad de una programación paralela y distribuida para manejar los grandes conjuntos de datos en el contexto general o incluso en el contexto geoespacial, así como la aplicación de los métodos de minería de datos, a pesar de que difieren de los métodos tradicionales de análisis de bases de datos que no presuponen un modelo que describe las relaciones en los datos ni requieren consultas específicas sobre las que basar el análisis, más bien estos enfoques permiten que los datos hablen por sí mismos, basándose en algoritmos para descubrir patrones que no son evidentes en los conjuntos de datos únicos y unidos.

Los algoritmos de minería de datos realizan diferentes tipos de operaciones[22]:

- Clasificación, donde los objetos o eventos se clasifican según categorías conocidas (por ejemplo, las compañías de seguros emplean algoritmos de clasificación para asignar categorías de riesgo de accidente a los conductores que comparten ciertas características).

- Agrupamiento (*Clustering*), donde se buscan patrones de similitud en los datos brutos (aplicable a *Waze*).

- Regresión o predicción numérica, donde las cantidades numéricas son predichas según el análisis de regresión.

- Asociación, donde se identifican las relaciones entre elementos de conjuntos de datos únicos o unidos.

- Detección de anomalías, donde se identifican los valores atípicos o pausas de patrones en los conjuntos de datos.

- Resumen, tabulación y presentación de características destacadas dentro de conjuntos de datos.

- Los enfoques de minería de datos pueden basarse en ejemplos de relaciones que proporcionan los operadores humanos y se utilizan para guiar el proceso o

[22] Un algoritmo en minería de datos (o aprendizaje automático) es un conjunto de heurísticas y cálculos que permiten crear un modelo a partir de datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa los resultados de este análisis en un gran número de iteraciones para determinar los parámetros óptimos para crear el modelo de minería de datos. https://msdn.microsoft.com/es-es/library/ms175595.aspx

mediante una operación no supervisada donde los patrones se descubren algorítmicamente.

Interpretación

Desde hace tiempo, los seres humanos han confiado en la representación gráfica o pictórica de los datos para que los registros de información sean accesibles, comprensibles y, lo que es más importante, atractivos para la mente humana. En consecuencia, la reciente explosión de datos también ha visto un aumento en las herramientas que se centran en la visualización efectiva de los datos. Muchas herramientas sobresalen en representar la información usando métodos tradicionales como tablas, histogramas, gráficos circulares y gráficos de barras. Sin embargo, los gráficos de barras presentados en documentos extensos o presentaciones de diapositivas no suelen adaptarse a una audiencia más allá de los profesionales. A medida que los conjuntos de datos se hacen más grandes y los esfuerzos para explotar estos datos buscan llegar a más personas, el lenguaje de visualización de datos debe adaptarse y mejorar.

Los ciudadanos viven en un mundo cada vez más visual, observan pantallas de diferentes tamaños con resoluciones que se incrementan cada que cambian de dispositivo. Las visualizaciones sirven no sólo para la entrega de información sino también para generar interés e impactar; son presentaciones de información enmarcadas en la convergencia del arte, los medios digitales y la tecnología de la información.

Visualización de datos

Son las técnicas utilizadas para generar imágenes, diagramas o animaciones para comunicar los resultados del análisis de datos, tales como mapas de tráfico. Las técnicas de visualización se utilizan durante y después del análisis de datos para dar sentido a la información.

Por lo tanto, las visualizaciones son ampliamente reconocidas como parte del proceso de análisis (es decir, no sólo la comunicación), en las que se puede explorar los datos y construir hipótesis durante este proceso (Chen & Zhang, 2014). Es importante señalar que las visualizaciones se han conceptualizado comúnmente como herramientas de comunicación. Si bien esto es cierto y las visualizaciones son muy importantes en la comunicación de hipótesis, resultados e ideas, en el caso de los datos masivos, su papel en la exploración juega un papel muy importante. Los datos masivos tienen muchas cosas para mostrar y a menudo se muestran pantallas muy "ocupadas" o sobresaturadas de información, por lo que se hace necesario un pre-análisis visual de las aplicaciones de datos masivos.

Un SIG es una caja de herramientas muy completa y de amplia utilización en la ciencia de datos, esto por su capacidad para procesar datos espaciales y no espaciales (atributos), incluso cuando no están perfectamente estructurados, a través de medios computacionales y visuales. La ciencia de la información

geográfica y sus aplicaciones asociadas, como la teledetección y la geoinformática, han estado tratando con grandes conjuntos de datos durante un tiempo relativamente largo, incluso antes de que el término datos masivos tomara impulso en la ciencia, la cultura popular y los negocios (Cöltekin & Reichenbacher, 2011). A escala urbana, regional y nacional, las plataformas de suministro de información capaces de combinar capas de información de manera comprensible pueden aumentar la eficiencia general y la sostenibilidad de la planificación y regulación de la infraestructura del transporte en sus diversos ámbitos y facilitar las labores de las instituciones encargadas de su administración.

Para promover el acceso universal y la compartibilidad, las visualizaciones deben poder ser exportadas y distribuidas en varios formatos, desde imágenes a vídeos o páginas web. Las grandes pantallas interactivas son más propensas a atraer a los usuarios, sobre todo, los principiantes. La compartibilidad de la plataforma de visualización es importante ya que permite la integración inteligente con otras plataformas como los portales web públicos o privados. También permite agregar nuevos módulos de terceros en un entorno de código abierto.

La difusión de los productos de análisis de datos, así como la distribución de algunas formas de datos, se realiza con papel o documentos digitales. Sin embargo, la utilidad de estos canales está disminuyendo ya que estos medios son generalmente estáticos. El acceso en línea a datos tabulares o geográficos que son directamente utilizables por diversas plataformas de software puede ser valioso para algunas formas de análisis de datos. Cada vez más, la difusión de datos se producirá a través de Interfaces de Programación de Aplicaciones[23] o formatos de datos que permiten la integración de estos datos directamente en diferentes aplicaciones móviles o de otros tipos. El acceso a datos dirigido a aplicaciones en teléfonos móviles basado en API ha contribuido a muchos nuevos servicios que facilitan enormemente la navegación, los viajes, la logística y otros servicios relacionados con el transporte para individuos y empresas. La difusión de datos a través de las API desempeñará un papel cada vez más importante en el acceso de los ciudadanos y el uso de datos relacionados con la movilidad, pero los beneficios finales de este canal de difusión de datos dependerán de los términos de uso asociados con esos datos.

Ahora bien, estos términos de uso de los datos abarcan desde términos completamente "abiertos" sin restricciones de uso y redistribución, hasta términos altamente restringidos que permiten el acceso comercial a los datos sólo en un conjunto limitado de condiciones. Los defensores de los datos abiertos proponen que la mayor cantidad posible de datos sea de uso abierto. Esto incluye (casi) todos

[23] Una API es un conjunto de funciones y procedimientos que cumplen una o muchas funciones con el fin de ser utilizadas por otro software. Las siglas API vienen del inglés *Application Programming Interface*. Una API permite implementar las funciones y procedimientos que engloba determinado proyecto sin la necesidad de programarlas de nuevo. En términos de programación, es una capa de abstracción. https://hipertextual.com/archivo/2014/05/que-es-api/

los datos recopilados por el gobierno, ya que sostienen que los ciudadanos ya los pagan con sus impuestos y por lo tanto se debería tener acceso libre a ellos. Por supuesto, hay problemas con el modelo de datos abiertos relacionados con la seguridad y la privacidad. Sin embargo, muchas autoridades están tratando de abrir el acceso a gran parte de sus datos, especialmente en el transporte. Por otro lado, las limitaciones que se imponen al acceso a los datos recogidos por empresas comerciales son comprensibles y necesarias, ya que las empresas que recopilan datos y prestan servicios basados en el uso y análisis del mismo, deben generar valor y ganancias económicas a sus propietarios. Para ello, se debe generalizar cuidadosamente, hacer hincapié en lo importante, eliminar lo poco importante, agrupar la información tanto de forma temática como perceptual y prestar atención a la jerarquía visual cuando se diseñan las pantallas.

3 Plataforma Twitter

Twitter es un servicio de microblogging[24], con sede en San Francisco, California, fue creado originalmente por Jack Dorsey en 2006, tiene amplia popularidad a nivel mundial y se estima que tiene más de 500 millones de usuarios, genera 65 millones de tweets al día y maneja más de 800,000 peticiones de búsqueda diarias. Ha sido denominado como el "SMS de Internet".

La plataforma permite enviar mensajes de texto plano de corta longitud, con un máximo de 140 caracteres, a estos mensajes se les llama *tweets* y se muestran en la página principal del usuario. Los usuarios pueden suscribirse a los *tweets* de otros usuarios, a esto se le llama *seguir* y a los usuarios se les llama *seguidores* o *followers*. Por default, los mensajes son públicos, pudiendo difundirse privadamente y mostrándolos únicamente a unos seguidores determinados. Los usuarios pueden *twitear* (acto de envíar tweets) desde la página web del servicio, con aplicaciones oficiales externas (para teléfonos inteligentes), o mediante el servicio de mensajes cortos (*SMS*) disponible en algunos países. Si bien el servicio es gratis, acceder a él vía SMS implica pagar tarifas que son fijadas por el proveedor de telefonía móvil. Actualmente, Twitter factura más de 2.500 millones de dólares anuales y tiene un valor en bolsa superior a los 10.000 millones de dólares.

La versión definitiva se lanzó el 15 de julio de 2006 y en palabras de su fundador, era "una corta ráfaga de información intrascendente", el "pio de un pájaro", que en inglés es *tweet*. Jack Dorsey es el actual presidente del Consejo de Administración de Twitter, Inc.

Según un estudio realizado por Semiocast[25] en 2012, analizando 383 millones de cuentas creadas antes de dicho año, los países con mayor número de usuarios en Twitter son los Estados Unidos (141 millones), Brasil (33,3 millones), Japón (29,9 millones), Reino Unido (23 millones), Indonesia (19 millones), India (12 millones),

[24] Microblogging es una forma de comunicación o sistema de publicación que consiste en el envío de mensajes cortos de texto (longitud máxima de 140 caracteres) a través de herramientas creadas para esta función. Su finalidad es explicar qué se está haciendo en un determinado momento, compartir información con otros usuarios u ofrecer enlaces hacia otras páginas web. http://microblogging18.blogspot.mx/2012/01/definicion.html

[25] http://semiocast.com/en/publications/2012_01_31_Brazil_becomes_2nd_country_on_Twitter_superseds_Japan

México (10,5 millones), Filipinas (8 millones), España (7,9 millones) y Canadá (7,5 millones)[26]

Las características más importantes de twitter son[27]:

- Asimétrica: twitter es una red social de relaciones optativas (seguir/ ser seguido), en la que no se requiere el consentimiento mutuo entre los usuarios.

- Breve: es un formato de escritura limitado a 140 caracteres por mensaje.

- Descentralizada: posee una arquitectura variable multipunto – multipunto, definida por las decisiones de cada usuario.

- Global: twitter es un servicio disponible en varios idiomas y en todo el planeta, incluida la Estación Espacial Internacional.

- Hipertextual: es un entorno de lectura-escritura en el que cada mensaje contiene enlaces por defecto en el que el uso del símbolo de la @ y del # genera enlaces de manera automática.

- Intuitiva: es un concepto de aplicación y una interfaz web orientados a usuarios no expertos, basados en la simplicidad y usabilidad.

- Multiplataforma: twitter es una aplicación con la que se puede interactuar desde clientes de mensajería de correo, de SMS, navegadores web y sus extensiones, ordenadores de sobremesa, portátiles, netbooks, tablets, móviles y redes sociales.

- Social: porque es un conjunto de comunidades y relaciones definidas por cada usuario.

- Viral: es una plataforma que, por su carácter global, social y sincrónico, facilita la rápida circulación y multiplicación de los mensajes

La interfaz web de Twitter está escrita en el lenguaje Ruby on Rails[28], los mensajes se mantienen en un servidor que funciona con software programado en Scala[29] y además dispone de una API abierta para todo tipo de desarrolladores, lo cual

[26] Se muestran los datos para el año 2012, debido a que no existen datos estadísticos recientes del uso y envío de tweets, así como del número de usuarios en distintos países.

[27] https://andrewzuniganajar.wordpress.com/2013/09/20/10-caracteristicas-comunicativas-de-twiiter/

[28] Rails es un conjunto de aplicaciones web que incluye todo lo necesario para crear aplicaciones web respaldadas por bases de datos de acuerdo con el modelo Model-View-Controller (MVC). http://rubyonrails.org/

[29] https://www.scala-lang.org/

supone una gran ventaja para todos aquellos que quieran integrar Twitter como un servicio tanto en otras aplicaciones web como en aplicaciones de escritorio o móviles. Según Biz Stone, más del 50 por ciento del tráfico se da a través de la API de Twitter.

Como red social, Twitter gira en torno al principio de los seguidores. Cuando se elige seguir a otro usuario de Twitter, los tuits de ese usuario aparecen en orden cronológico inverso, en la página principal de Twitter.

Los usuarios pueden agrupar mensajes sobre un mismo tema mediante el uso de etiquetas, palabras o frases iniciadas mediante el uso de un "#" (numeral) conocidas como *hashtag*, su nombre original en inglés y el que se utiliza normalmente en Twitter. De forma similar, la "@" (arroba) seguida de un nombre de usuario se usa para mencionar o contestar a otros usuarios. Para volver a postear un mensaje de otro usuario, y compartirlo con los propios seguidores, la función de retuit se marca con un "RT" en el mensaje.

A finales de 2009 se añadió la opción de listas, haciendo posible seguir (así como mencionar y contestar) listas de usuarios en vez de usuarios individuales.

Los mensajes fueron fijados a 140 caracteres máximo para la compatibilidad con los mensajes SMS, introduciendo la notación de la taquigrafía y el argot de Internet comúnmente usado en los SMS. El límite de 140 caracteres también ha llevado a la proliferación de servicios de reducción de URLs, como bit.ly, goo.gl, y tr.im, y web de alojamiento de material, como Twitpic, memozu.com y NotePub para subir material multimedia y textos superiores a 140 caracteres. El 11 de junio de 2015, Twitter anunciaba que esta restricción de caracteres se eliminaría en los Mensajes privados a partir de julio del mismo año, quedando el límite establecido en 10.000 caracteres

Como se mencionó, existe la posibilidad de que los tweets que se envían contengan información acerca de la posición desde donde fueron enviados, esto significa que pueden tener las coordenadas geográficas del lugar en donde se encontraba el dueño y el dispositivo de envío.

De acuerdo con estimaciones de Wade (S/F), un porcentaje muy bajo (5% - 7%) de tweets presentan dichas coordenadas y aunque el volumen diario de tweets es sumamente alto, el subconjunto es extremadamente bajo.

3.1.1 API de Twitter

Una API (siglas de *Application Programming Interface*) es un conjunto de reglas (código) y especificaciones que las aplicaciones pueden seguir para comunicarse entre ellas: sirviendo de interfaz entre programas diferentes de la misma manera en que la interfaz de usuario facilita la interacción humano-software.

Las API pueden servir para comunicarse con el sistema operativo (WinAPI), con bases de datos (DBMS) o con protocolos de comunicaciones (Jabber/XMPP). En los últimos años, se han sumado múltiples redes sociales (Twitter, Facebook, Youtube, Flickr, LinkedIn, etc) y otras plataformas online (Google Maps, WordPress).

La API permite hacer uso de funciones ya existentes en otro software (o de la infraestructura ya existente en otras plataformas) para no estar reinventando o reescribiendo codigo constantemente, que se sabe que está probado y que funciona correctamente. En el caso de herramientas propietarias (es decir, que no sean de código abierto), son un modo de hacer saber a los programadores de otras aplicaciones cómo incorporar una funcionalidad concreta sin por ello tener que proporcionar información acerca de cómo se realiza internamente el proceso[30].

El uso de las API como una '*subcontratación*' de funciones, en donde se imponen condiciones al subcontratante: algunos sitios como Twitter o eBay limitan el número de llamadas que un determinado software o web pueden hacer a su API en un determinado período de tiempo (por minuto, hora o día) antes de tener que pagar una licencia.

3.1.2 Software R y librerías asociadas

R es un lenguaje y un entorno para la informática estadística y los gráficos[31]. Es similar al lenguaje **S** y el entorno que fue desarrollado en Bell Laboratories (anteriormente AT&T, ahora Lucent Technologies) por John Chambers. Hay algunas diferencias importantes, pero gran parte del código escrito para **S** se ejecuta sin alteración bajo **R**.

R proporciona una amplia variedad de modelos estadísticos (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupación) y técnicas gráficas. Una de las fortalezas de **R** es la facilidad con la que se pueden producir marcos para publicación bien diseñadas, incluyendo símbolos matemáticos y fórmulas donde sea necesario. **R** está disponible como Software Libre bajo los términos de la Licencia Pública General GNU de la Free Software Foundation en forma de código fuente. Compila y ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluyendo FreeBSD y Linux), Windows y MacOS.

[30] http://www.ticbeat.com/tecnologias/que-es-una-api-para-que-sirve/

[31] https://www.r-project.org/about.html

3.1.2.1 El entorno R

R es un conjunto integrado de extensiones de software para la manipulación de datos, cálculo y visualización gráfica. Incluye una extensión para el tratamiento y almacenamiento de datos, una serie de operadores para cálculos sobre matrices, una gran colección de herramientas intermedias para el análisis de datos, extensiones gráficas para el análisis de datos y visualización en pantalla o en papel y un lenguaje de programación bien desarrollado, simple y eficaz que incluye condicionales, bucles, funciones recursivas definidas por el usuario y extensiones de entrada y salida.

3.1.2.2 R Studio

RStudio es un entorno de desarrollo integrado o entorno de desarrollo interactivo, (*Integrated Development Environment* – IDE), es una aplicación informática que proporciona servicios integrales para facilitarle al desarrollador o programador el desarrollo de software. Las características de ***R Studio*** son[32]:

- Acceso local a ***RStudio***
- Resaltado de sintaxis, finalización de código y sangría inteligente
- Ejecución de código R directamente desde el editor fuente
- Acceso rápido las definiciones de funciones
- Administración de múltiples directorios de trabajo mediante proyectos
- Ayuda y documentación
- Depurador interactivo para diagnosticar y corregir errores rápidamente
- Extensas herramientas de desarrollo de paquetes

3.1.2.3 Aplicación y utilización de los datos

Tomando como base el código fuente original desarrollado por *amsantac.co*[33] y que se encuentra disponible en su página web, se desarrolló una aplicación interactiva que permite identificar tweets con geoposición a partir de una etiqueta (hashtag #), de una posición inicial y un radio de extensión territorial.

[32] https://www.rstudio.com/products/rstudio/

[33] http://amsantac.co/blog/es/2016/05/28/twitter-r-es.html

Para esto, en primer lugar se creó una cuenta en twitter y una vez que se ha generado la cuenta, en apps.twitter.com[34] se debe crear una nueva aplicación para de este modo tener acceso a la API de Twitter.
En la API de Twitter se deben generar los siguientes valores o claves:

- Consumer key
- Consumer secret
- Access Token
- Access Token Secret

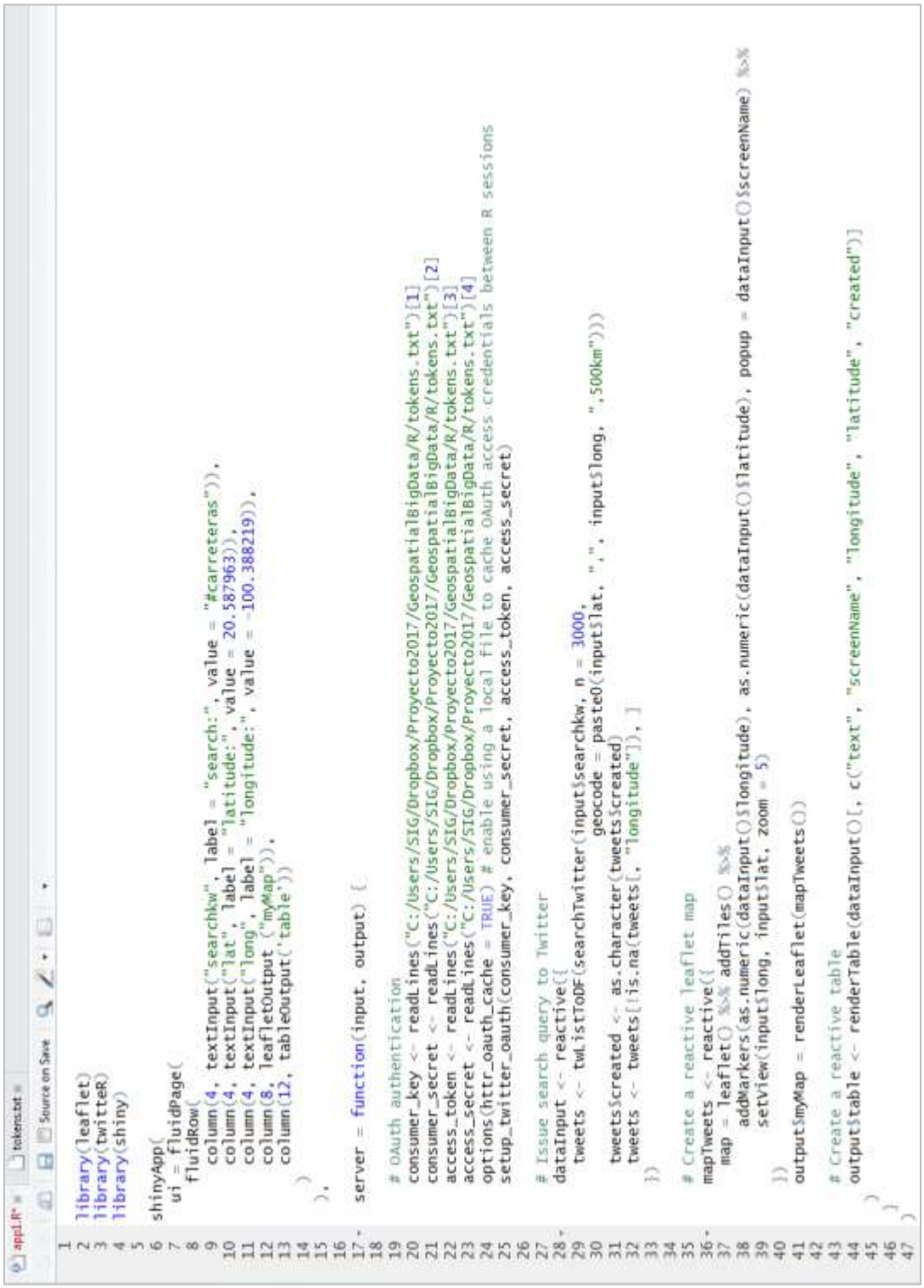
Se instaló **R** versión x64 3.3.3 y así como los paquetes complementarios twitteR[35], shiny[36] y leaflet[37]; estos paquetes contienen instrucciones específicas para ejecutar determinadas operaciones del código.
Después, en una sesión de trabajo de **R**, visualizada a través de R Studio se implementa, prueba y depura el siguiente código, ver *Figura 3.1 Código de búsqueda y despliegue de tweets en un mapa*

[34] https://apps.twitter.com/

[35] https://cran.r-project.org/web/packages/twitteR/README.html

[36] https://cloud.r-project.org/web/packages/shiny/index.html

[37] https://cloud.r-project.org/web/packages/leaflet/index.html

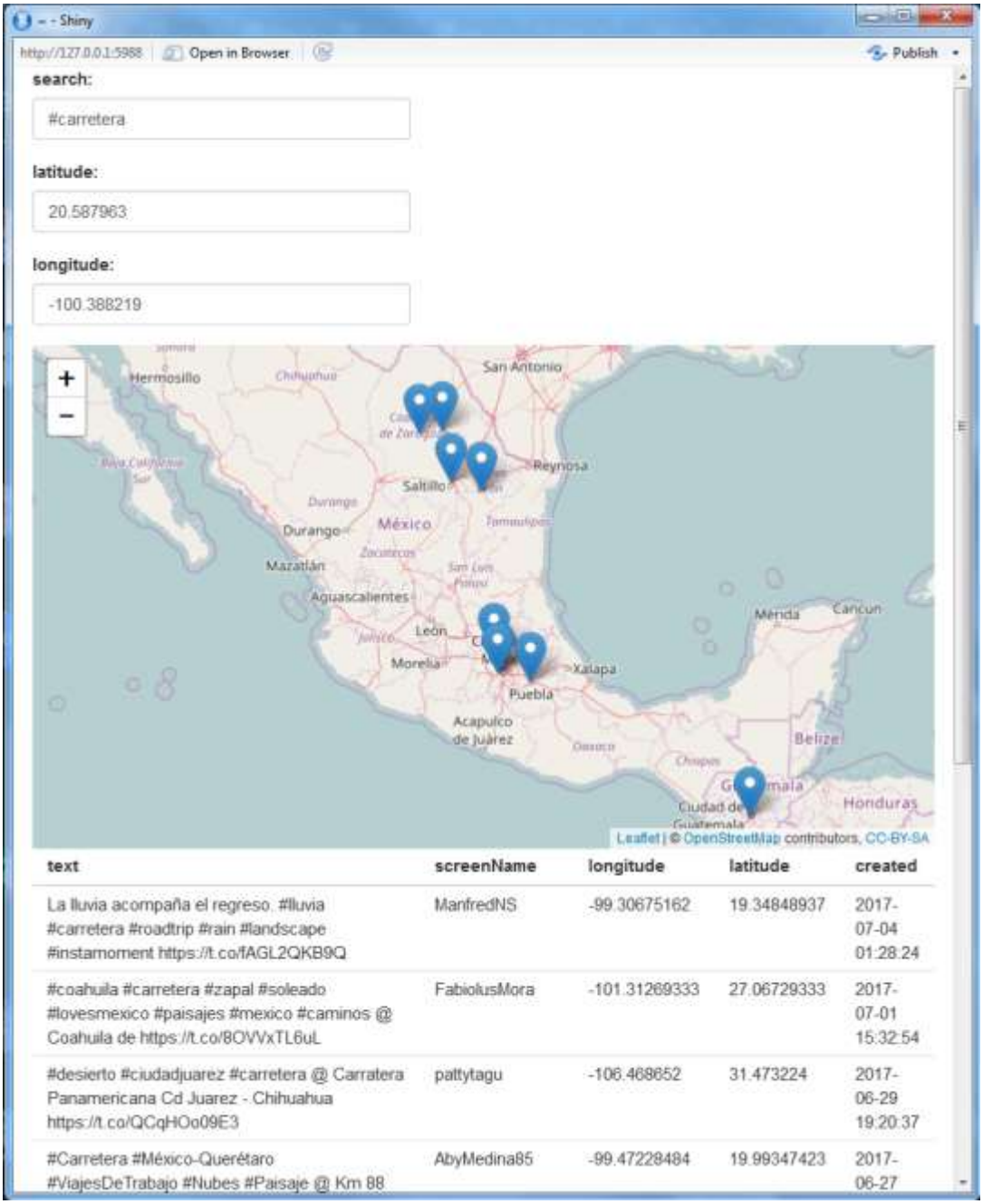


```
library(leaflet)
library(twitteR)
library(shiny)

shinyApp(
  ui = fluidPage(
    fluidRow(
      column(4, textInput("searchkw", label = "search:", value = "#carreteras")),
      column(4, textInput("lat", label = "latitude:", value = 20.587963)),
      column(4, textInput("long", label = "longitude:", value = -100.388219)),
      column(8, leafletOutput("myMap")),
      column(12, tableOutput('table'))
    )
  ),

  server = function(input, output) {

    # OAuth authentication
    consumer_key <- readLines("C:/Users/SIG/Dropbox/Proyecto2017/GeospatialBigData/R/tokens.txt")[1]
    consumer_secret <- readLines("C:/Users/SIG/Dropbox/Proyecto2017/GeospatialBigData/R/tokens.txt")[2]
    access_token <- readLines("C:/Users/SIG/Dropbox/Proyecto2017/GeospatialBigData/R/tokens.txt")[3]
    access_secret <- readLines("C:/Users/SIG/Dropbox/Proyecto2017/GeospatialBigData/R/tokens.txt")[4]
    options(httr_oauth_cache = TRUE) # enable using a local file to cache OAuth access credentials between R sessions
    setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

    # Issue search query to Twitter
    dataInput <- reactive({
      tweets <- twListToDF(searchTwitter(input$searchkw, n = 3000,
                                         geocode = paste0(input$lat, ",", input$long, ",500km")))
      tweets$created <- as.character(tweets$created)
      tweets <- tweets[!is.na(tweets[, "longitude"]), ]
    })

    # Create a reactive leaflet map
    mapTweets <- reactive({
      map = leaflet() %>% addTiles() %>%
        addMarkers(as.numeric(dataInput()$longitude), as.numeric(dataInput()$latitude), popup = dataInput()$screenName) %>%
        setView(input$long, input$lat, zoom = 5)
    })
    output$myMap = renderLeaflet(mapTweets())

    # Create a reactive table
    output$table <- renderTable(dataInput()[, c("text", "screenName", "longitude", "latitude", "created")]
    )
  }
)
```

Fuente: Elaboración propia tomando como base el código fuente original desarrollado por amsantac.co

Figura 3.1 Codigo de búsqueda y despliegue de tweets en un mapa

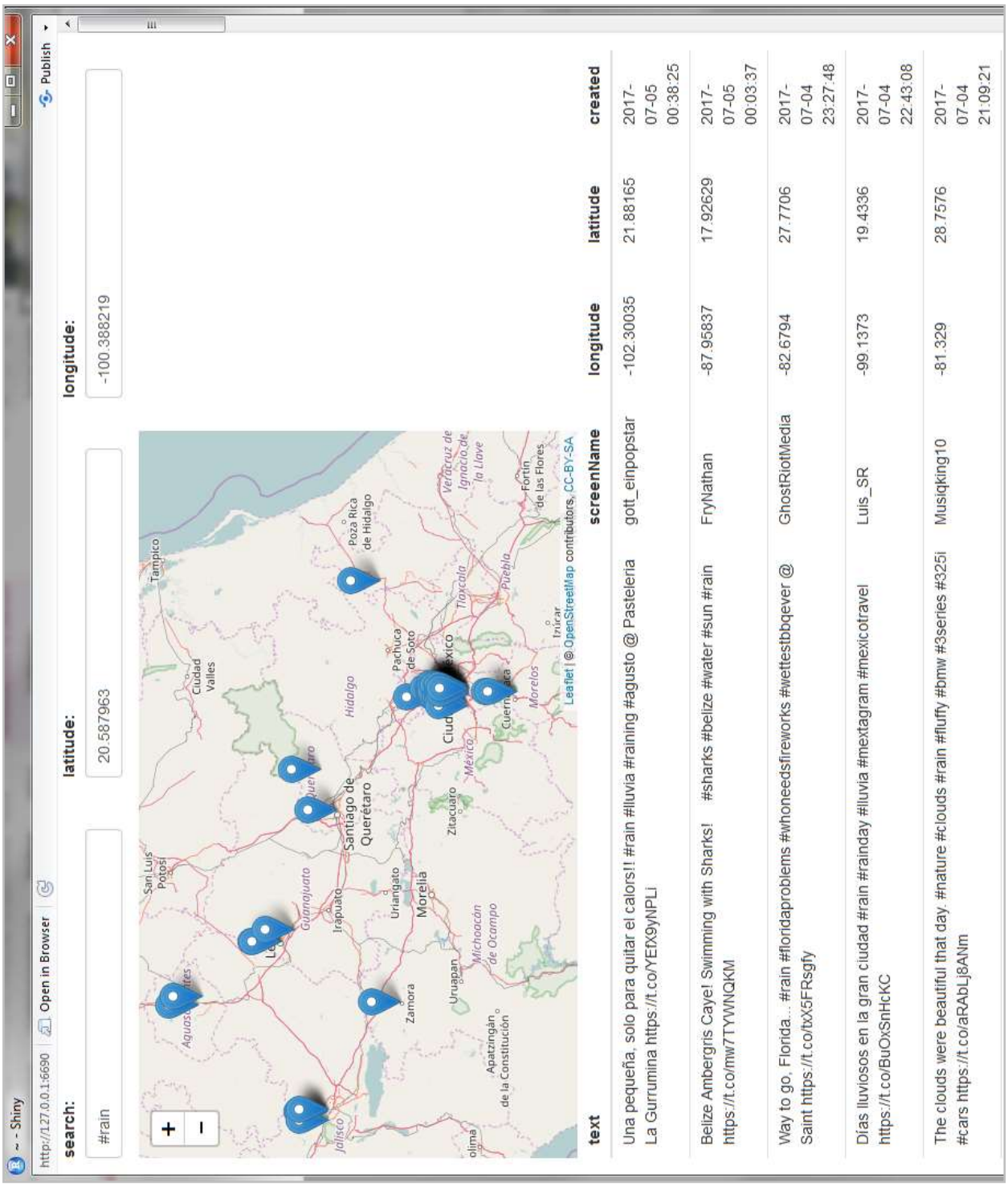
Algunos de los resultados se muestran en las Figuras 3.2 y 3.3; se debe notar que la cantidad de tweets que cuentan con georreferenciación es muy poca (entre 3-5% del total). Se espera que esto aumente dramáticamente en los próximos años debido a la proliferación de dispositivos inteligentes y aplicaciones móviles.



text	screenName	longitude	latitude	created
La lluvia acompaña el regreso. #lluvia #carretera #roadtrip #rain #landscape #instamoment https://t.co/fAGL2QKB9Q	ManfredNS	-99.30675162	19.34848937	2017-07-04 01:28:24
#coahuila #carretera #zapal #soleado #lovesmexico #paisajes #mexico #caminos @ Coahuila de https://t.co/8OVVxTL6uL	FabiolusMora	-101.31269333	27.06729333	2017-07-01 15:32:54
#desierto #ciudadjuarez #carretera @ Carratera Panamericana Cd Juarez - Chihuahua https://t.co/QCqHOo09E3	pattytagu	-106.468652	31.473224	2017-06-29 19:20:37
#Carretera #México-Querétaro #ViajesDeTrabajo #Nubes #Paisaje @ Km 88	AbyMedina85	-99.47228484	19.99347423	2017-06-27

Fuente: Elaboración propia

Figura 3.2 Resultado de la ejecución del codigo de busqueda y despliegue de tweets en un mapa, en este caso #carretera



text	screenName	longitude	latitude	created
Una pequeña, solo para quitar el calors!! #rain #lluvia #raining #agusto @ Pasteleria La Gurrumina https://t.co/YEfX9yNPLi	gott_einpopstar	-102.30035	21.88165	2017-07-05 00:38:25
Belize Ambergris Caye! Swimming with Sharks! https://t.co/mw7TYWNQKM	#sharks #belize #water #sun #rain	FryNathan	-87.95837	17.92629	2017-07-05 00:03:37
Way to go, Florida... #rain #floridaproblems #whoneedsfireworks #wettestbbqever @ Saint https://t.co/txX5FRsgfy	GhostRiotMedia	-82.6794	27.7706	2017-07-04 23:27:48
Días lluviosos en la gran ciudad #rain #rainday #lluvia #mextagram #mexicotravel https://t.co/BuOxSnHcKC	Luis_SR	-99.1373	19.4336	2017-07-04 22:43:08
The clouds were beautiful that day. #nature #clouds #rain #fluffy #bmw #3series #325i #cars https://t.co/aRAbLj8ANm	Musiqking10	-81.329	28.7576	2017-07-04 21:09:21

Fuente: Elaboración propia

Figura 3.3 Resultado de la ejecución del codigo de búsqueda y despliegue de tweets en un mapa, en este caso #rain

4 Plataforma Waze

4.1 Cómo funciona

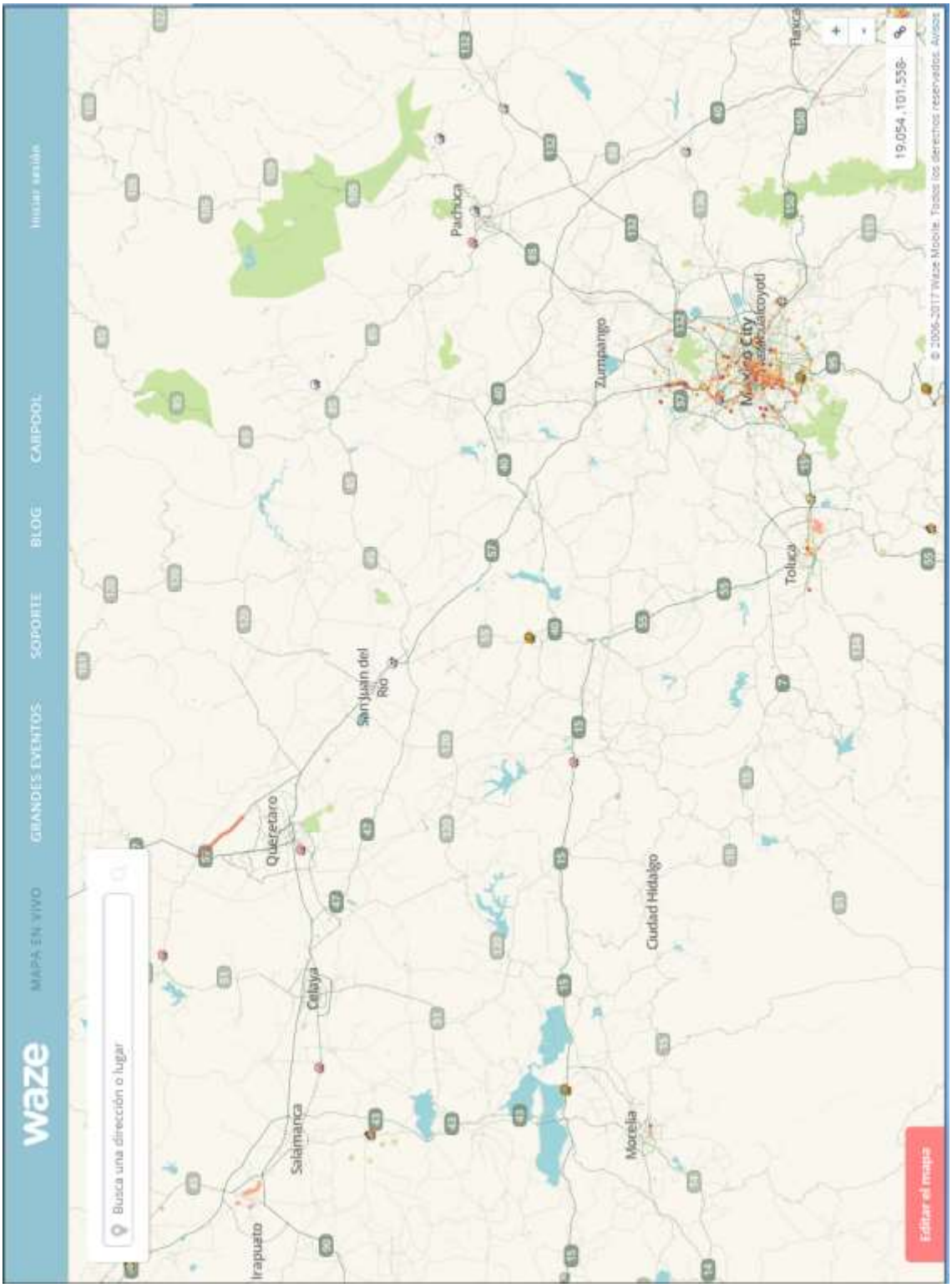
El *crowdsourcing* (*Introducción, pag. 12*), es la base fundamental de la aplicación *Waze*, la cual fue creada en 2008 en Israel, utiliza navegación GPS, proporciona información del tránsito, además de varias funciones sociales. Los Wazers (usuarios de waze) se pueden informar unos a otros sobre el tránsito, controles policiales, obras, radares, accidentes y eventos meteorólogicos. Al ser información generada por los propios usuarios, cuantos más usen la plataforma, sera mejor y habrá mas datos. En 2013 Waze fue adquirida por parte de Google por 1.3 miles de millones de dólares, fundamentalmente para tener acceso a la información social de la plataforma y para recolectar información en tiempo real en una forma más barata[38]. Se estima que en todo el mundo la aplicación cuenta con 70 millones de usuarios y para la Ciudad de México se cree que hay 1.5 millones de usuarios[39].

Existen dos maneras de participar en la plataforma, la primera es cuando los usuarios conducen con la aplicación abierta para contribuir pasivamente con información del tránsito; también pueden tomar un rol más activo al compartir alertas de incidentes en el tránsito como accidentes, controles policiales o cualquier peligro que encuentren en la vía. Esto ayuda a otros usuarios que están en la zona porque reciben un aviso de lo que sucede más adelante en su ruta. Existe una comunidad de editores de mapas que aseguran que los datos esten lo más actualizados que sea posible[40], ver Figura 4.1 y Figura 4.2

[38] https://www.entrepreneur.com/article/266030

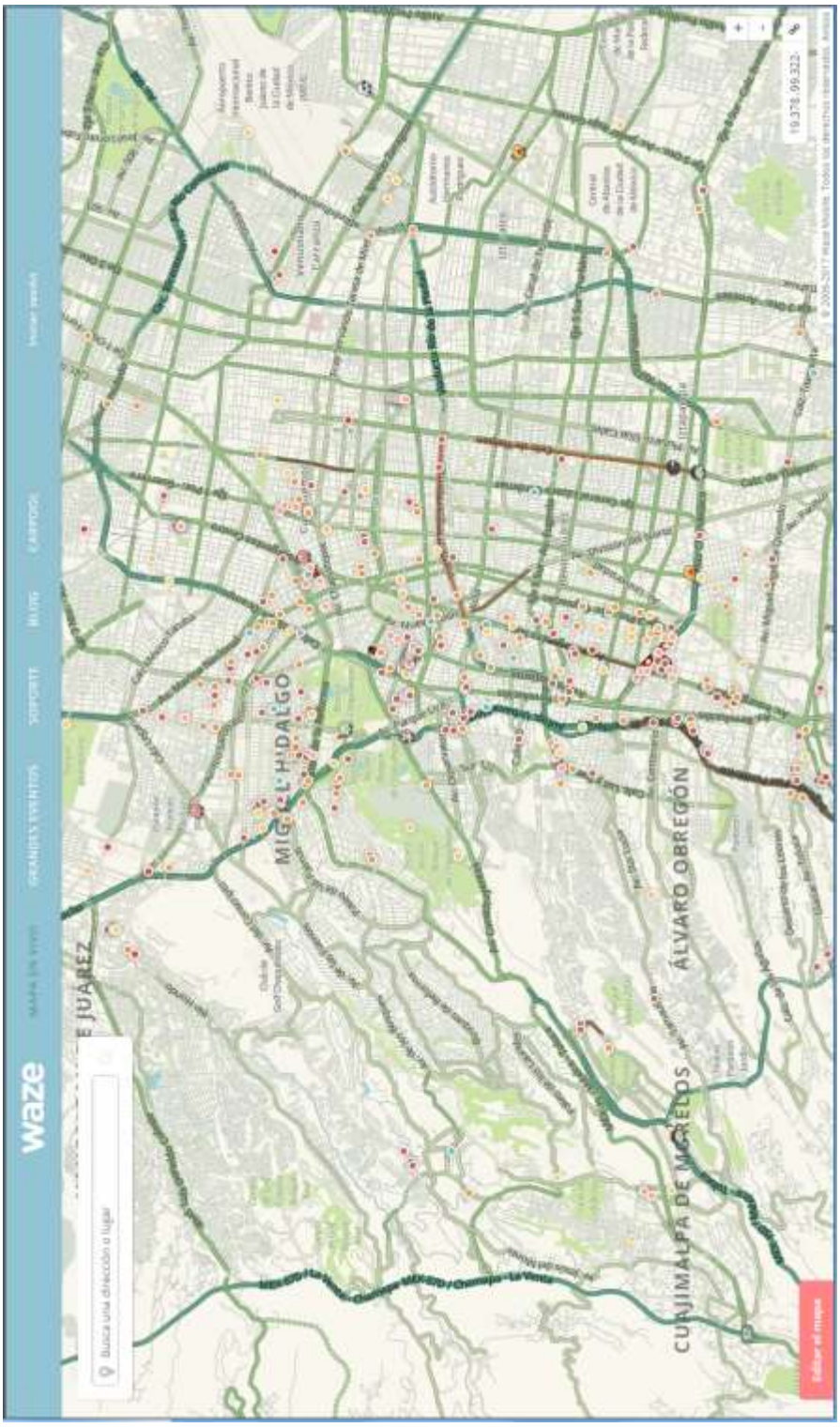
[39] Estos son números estimados ya que la plataforma, por políticas internas, no publica el número de usuarios totales a nivel mundial y tampoco por país.

[40] Tomado de ¿Cómo funciona Waze? https://support.google.com/waze/answer/6078702?hl=es



Fuente: waze.com

Figura 4.1 Pantalla de mapa en vivo de Waze.



Fuente: waze.com

Figura 4.2 Pantalla de mapa en vivo de Waze, acercamiento a la Ciudad de México.

4.2 Descripción del programa *Citizen Connected Partner*

El Programa de Ciudadanos Conectados (*Citizen Connected Partner - CCP*) es una asociación continua entre *Waze* y varias agencias gubernamentales internacionales para compartir datos de cierre de caminos e incidentes que se distribuyen públicamente. Los participantes en este programa (a Octubre 2016), divididos por regiones son[41]:

North America

1. Alabama – Department of Transportation
2. California – Caltrans
3. California – City of Cupertino City Hall
4. California – City of Los Angeles
5. California – City of Sacramento
6. California – City of San Francisco
7. California – Los Angeles Metropolitan Transportation Authority
8. California – Paramedics Plus (Genesis Pulse)
9. California – Town of Los Gatos
10. Canada – Ville de Montreal
11. Colorado – City of Colorado Springs
12. Colorado – Douglas County
13. District of Columbia – D.C. Department of Transportation
14. Florida – City of Miami Beach
15. Florida – City of Tampa
16. Florida – Florida Department of Transportation
17. Florida – Paramedics Plus (Genesis Pulse)
18. Florida – Miami-Dade County
19. Florida – Sunstar EMS (Genesis Pulse)
20. Georgia – Bartow County
21. Georgia – City of Atlanta

[41] https://wiki.waze.com/wiki/Connected_Citizens_Program

22. Georgia – City of Johns Creek
23. Georgia – City of West Jackson
24. Georgia – Georgia Department of Transportation
25. Georgia – Georgia Emergency Management & Homeland Security Agency
26. Illinois – City of Evanston
27. Illinois – City of Naperville
28. Indiana – City of Bloomington
29. Indiana – Paramedics Plus (Genesis Pulse)
30. Indiana – Three Rivers Ambulance Authority (Genesis Pulse)
31. Iowa – Iowa Department of Transportation
32. Kentucky – City of Louisville
33. Kentucky – Kentucky Transportation Cabinet
34. Louisiana – City-Parish of Baton Rouge
35. Louisiana – Louisiana Department of Transportation and Development
36. Maine – Maine Department of Transportation
37. Maryland – University of Maryland
38. Maryland – St. Mary's Emergency Services and Technology
39. Massachusetts – City of Boston
40. Massachusetts – City of Cambridge (Kleinfelder East)
41. Massachusetts – Capital Strategic Solutions
42. Massachusetts – Massachusetts Department of Transportation
43. Missouri – CoxHealth (Genesis Pulse)
44. Missouri – Mercy EMS (Genesis Pulse)
45. Missouri – Taney County Ambulance Directory (Genesis Pulse)
46. National – SeeClickFix (nonprofit partner)
47. National – United States Department of Transportation
48. Nebraska – Nebraska Department of Roads
49. Nevada - Regional Transportation Commission of Southern Nevada (RTCSVN)
50. New Hampshire – New Hampshire Department of Transportation
51. New Jersey – City of Jersey City

52. New Jersey – Jersey City EMS (Genesis Pulse)
53. North Carolina – City of Charlotte
54. North Carolina – City of Greensboro
55. North Carolina – City of Raleigh
56. Ohio – Town of Dublin
57. Oregon – Oregon Department of Transportation
58. Pennsylvania – Pennsylvania Department of Transportation
59. Pennsylvania – Pennsylvania Turnpike Authority
60. Pennsylvania – Wilkes-Barre Township Police Department
61. Rhode Island – City of Providence
62. Rhode Island – Rhode Island Turnpike & Bridge Authority
63. South Dakota – Paramedics Plus (Genesis Pulse)
64. Tennessee – Tennessee Department of Transportation
65. Texas – CareFlite (Genesis Pulse)
66. Texas – Champion EMS (Genesis Pulse)
67. Texas – ETMC EMS (Genesis Pulse)
68. Texas - City of Fort Worth
69. Texas – LifeNet EMS (Genesis Pulse)
70. Utah – Utah Department of Transportation
71. Vermont – Vermont Department of Transportation
72. Virginia – City of Arlington
73. Virginia – Portsmouth Police Department

Latin America

1. Brazil – City of Petropolis
2. Brazil – City of Vitoria
3. Brazil – Juiz de Fora Secretary of Transport and Transit, Secretaria de Transporte e Transito
4. Brazil – Rio de Janeiro Center for Traffic Operations (COR)
5. Colombia – Bogotá Instituto de Desarrollo Urbano (IDU)
6. Colombia – Medellín Alcaldía de Medellín
7. Costa Rica – Ministry of Transport

8. México – City of Puebla
9. México – Delegación Miguel Hidalgo (México City)
10. México – La Sultana de Norte (Monterrey)
11. Perú – Municipalidad de Miraflores

Europe

1. Belgium – City of Ghent
2. Estonia – Tarktee (Smart Roads)
3. France – Department of Var
4. France – Northern France, Tollway Authority
5. Hungary – BKK Center for Budapest Transport
6. Latvia – City of Riga
7. Latvia – Latvia State Roads
8. Lithuania – Lithuania Road Administration
9. Netherlands – National Data Warehouse for Traffic Information
10. Portugal – Brisa/Via Verde (Portugal Tollway Authority)
11. Rome – Rome Center for Mobility
12. Spain – City of Barcelona
13. Spain – Government of Catalonia
14. United Kingdom – Transport for London

Middle East

1. Israel - City of Tel-Aviv
2. Israel - Holon Municipality

Asia-Pacific

1. Indonesia - City of Jakarta
2. Australia - Transportation Management Centre of New South Wales

Los objetivos del programa son reducir la congestión, aumentar la eficiencia de la respuesta a incidentes y tomar decisiones basadas en datos de lo que está sucediendo en la infraestructura vial, ya sean calles o carreteras.

El Programa de Ciudadanos Conectados de *Waze* es un intercambio de datos bidireccional que permite transferir información en ambas direcciones para ayudar a la toma de decisiones y de este modo lograr un impacto concreto en la comunidad en cuanto a identificación de rutas alternativas al tráfico, localización de accidentes y eventos que afectan el transporte en las zonas urbanas y carreteras que las conectan.

Lanzado en octubre de 2014 con 10 socios de distintas ciudades, el programa se ha expandido a más de 100 socios, incluyendo agencias gubernamentales de ciudades, estados y países, así como organizaciones sin fines de lucro.

En este modo, *Waze* proporciona en tiempo real, información sobre cierres de calles o carreteras, tráfico, rangos de velocidad y diversos eventos directamente desde el origen, en este caso, los propios conductores. El CCP al proporcionar más datos, da a los *Wazers* una mayor capacidad para evitar los cierres de carretera y los lugares con tráfico.

Los miembros de CCP reciben la información de los incidentes en tiempo real más rápido que otros métodos de transferencia de informes. *Waze* identifica de una manera precisa y verifica dónde ocurren los incidentes, creando tiempos de respuesta más rápidos. Los socios pueden utilizar estándares de datos diseñados por *Waze* para el cierre y la notificación de incidentes para reducir la fragmentación de datos y, de este modo, promover la agregación de datos de transporte así como de distintas instancias de gobierno.

En el programa CCP se encuentran disponibles las herramientas que permiten realizar algunas de las siguientes actividades:

- Especificación de incidentes.
- Herramienta de visualización de tráfico.
- Mensajes para aviso de tráfico inusual.
- Mapa en vivo de *Waze.*
- Herramienta de cierre de carretera.
- Foro de socios en línea.
- Herramienta para eventos de crisis o situaciones de emergencia, por ejemplo, tormentas o lluvias torrenciales.

Waze pone sus datos a disposición de los socios CCP a través de un *feed*[42] en formato JSON o XML que se actualiza cada 2 minutos.

4.3 Formas de distribución de datos

4.3.1 JSON (JavaScript Object Notation)

Es un formato para el intercambio de datos, básicamente JSON describe los datos con una sintaxis dedicada que se usa para identificar y gestionar los datos. JSON nació como una alternativa a XML. Una de las mayores ventajas que tiene el uso de JSON es que puede ser leído por cualquier lenguaje de programación. Por lo tanto, puede ser usado para el intercambio de información entre distintas tecnologías.

JSON (*JavaScript Object Notation* - Notación de Objetos de JavaScript) es un formato ligero de intercambio de datos. Está basado en un subconjunto del Lenguaje de Programación JavaScript, Standard ECMA-262 3rd Edition de diciembre 1999[43]. JSON es un formato de texto que es completamente independiente del lenguaje pero utiliza convenciones que son ampliamente conocidos por los programadores de la familia de lenguajes C, incluyendo C, C++, C#, Java, JavaScript, Perl, Python, y muchos otros. Estas propiedades hacen que JSON sea un lenguaje ideal para el intercambio de datos.

JSON está constituído por dos estructuras:
Una colección de pares de nombre/valor. En varios lenguajes esto es conocido como un objeto, registro, estructura, diccionario, tabla hash, lista de claves o un arreglo asociativo.

Una lista ordenada de valores. En la mayoría de los lenguajes, esto se implementa como arreglos, vectores, listas o sequencias.

Estas son estructuras universales; virtualmente todos los lenguajes de programación las soportan de una forma u otra. Es razonable que un formato de intercambio de datos que es independiente del lenguaje de programación se base en estas estructuras.

Un archivo JSON, se representa de esta forma:

[42] Una fuente web o canal web (en inglés *web feed*) es un medio de redifusión de contenido web. http://www.wordreference.com/es/translation.asp?tranword=feed

[43] http://www.json.org/json-es.html

Un objeto o conjunto desordenado de pares nombre/valor. Un objeto comienza con { (llave de apertura) y termine con } (llave de cierre). Cada nombre es seguido por : (dos puntos) y los pares nombre/valor están separados por , (coma). [44]

4.3.1.1 Muestra de datos JSON generados por Waze

Ejemplo:


```
{"country":"MX","roadType":2,"magvar":44,"subtype":"JAM_STAND_STILL_TRAFFIC","reportRating":0,"confidence":0,"reliability":5,"location":{"x":-99.22162,"y":19.32204},"type":"JAM","uuid":"191035d1-e70b-3cbc-9847-76d85652cee1","pubMillis":1499197477095},{"country":"MX","city":"Magdalena Contreras","reportRating":0,"confidence":0,"reliability":5,"type":"JAM","uuid":"414f68bb-8c82-3ee7-8f81-d182b6a406b6","roadType":2,"magvar":37,"subtype":"JAM_HEAVY_TRAFFIC","street":"Av. México","location":{"x":-99.222258,"y":19.321257},"pubMillis":1499197658949},{"country":"MX","city":"Magdalena Contreras","reportRating":1,"confidence":0,"reliability":5,"type":"JAM","uuid":"880586d1-fc2b-391f-b518-fee1a8f67233","roadType":2,"magvar":42,"subtype":"JAM_STAND_STILL_TRAFFIC","street":"Av. Contreras","location":{"x":-99.21805,"y":19.324541},"pubMillis":1499197717544},{"country":"MX","city":"Magdalena Contreras","reportRating":2,"confidence":0,"reliability":5,"type":"JAM","uuid":"123bc295-e781-310b-8396-b79414aca889","roadType":1,"magvar":149,"subtype":"JAM_STAND_STILL_TRAFFIC","street":"Veracruz","location":{"x":-99.227671,"y":19.319699},"pubMillis":1499197720769}
```


[44] Ibidem.

Fuente: waze.com

Figura 4.3 Captura de pantalla del feed en formato Json.

4.3.2 XML (*eXtensible Markup Language*)

XML, (siglas en inglés de *eXtensible Markup Language)*, traducido como "Lenguaje de Marcado Extensible" o "Lenguaje de Marcas Extensible", es un meta-lenguaje que permite definir lenguajes de marcas desarrollado por el World Wide Web Consortium (W3C) utilizado para almacenar datos en forma legible. Proviene del lenguaje SGML y permite definir la gramática de lenguajes específicos (de la misma manera que HTML es a su vez un lenguaje definido por SGML) para estructurar documentos grandes[45]. A diferencia de otros lenguajes, XML da soporte a bases de datos, siendo útil cuando varias aplicaciones deben comunicarse entre sí o integrar información.

XML es un estándar para el intercambio de información estructurada entre diferentes plataformas. Se puede usar en bases de datos, editores de texto, hojas de cálculo. XML es una tecnología sencilla que tiene a su alrededor otras que la complementan y la hacen mucho más grande, con unas posibilidades mucho mayores. Tiene un papel muy importante en la actualidad ya que permite la compatibilidad entre sistemas para compartir la información de una manera segura, fiable y fácil.

Un archivo XML se representa de la siguiente forma:

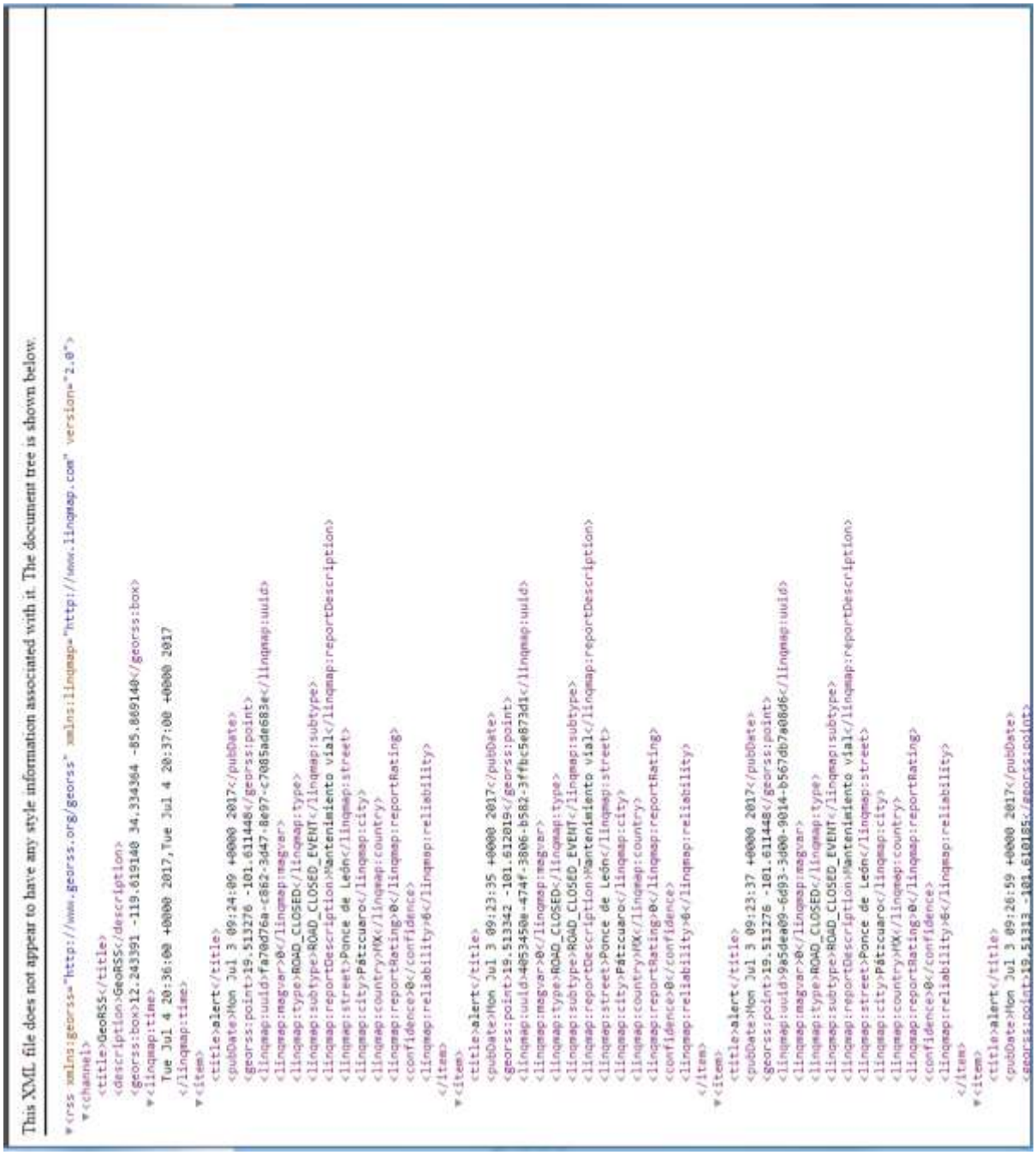
- Elementos: Pieza lógica del marcado, se representa con una cadena de texto (dato) encerrada entre etiquetas. Pueden existir elementos vacíos (
). Los elementos pueden contener atributos.
- Instrucciones: órdenes especiales para ser utilizadas por la aplicación que procesa <?xml-stylesheet type="text/css" href="estilo.css">
- Las instrucciones XML. Comienzan por <? Y terminan por ?>.
- Comentarios: Información que no forma parte del documento. Comienzan por <!-- y terminan por -->.
- Declaraciones de tipo: Especifican información acerca del documento. <!DOCTYPE persona SYSTEM "persona.dtd">
- Secciones CDATA: Se trata de un conjunto de caracteres que no deben ser interpretados por el procesador. <![CDATA[Aquí se puede meter cualquier carácter, como <, &, >, ... Sin que sean interpretados como marcación]][46]>

[45] https://www.w3.org/XML/

[46] Ibidem.

4.3.2.1 Muestra de datos XML generados por Waze

```
<rss xmlns:georss="http://www.georss.org/georss"
xmlns:linqmap="http://www.linqmap.com" version="2.0">
<channel>
<title>GeoRSS</title>
<description>GeoRSS</description>
<georss:box>-85.000000 -179.000000 85.000000 179.000000</georss:box>
<linqmap:time>
Fri May 26 17:58:00 +0000 2017,Fri May 26 17:59:00 +0000 2017
</linqmap:time>
<item>
<title>alert</title>
<pubDate>Tue Jul 4 20:35:54 +0000 2017</pubDate>
<georss:point>19.393525 -99.135527</georss:point>
<linqmap:uuid>c400896b-3692-3510-823e-9ed0b07f0a7c</linqmap:uuid>
<linqmap:magvar>97</linqmap:magvar>
<linqmap:type>WEATHERHAZARD</linqmap:type>
<linqmap:subtype>HAZARD_ON_ROAD_CAR_STOPPED</linqmap:subtype>
<linqmap:street>Eje 4 Sur - Napoleón (carril derecho)</linqmap:street>
<linqmap:city>Benito Juárez</linqmap:city>
<linqmap:country>MX</linqmap:country>
<linqmap:roadType>6</linqmap:roadType>
<linqmap:reportRating>0</linqmap:reportRating>
<confidence>0</confidence>
<linqmap:reliability>5</linqmap:reliability>
</item>
<item>
<title>alert</title>
<pubDate>Tue Jul 4 20:36:06 +0000 2017</pubDate>
<linqmap:uuid>227e18c9-bd22-3408-9a4b-69845c94d0de</linqmap:uuid>
<linqmap:magvar>126</linqmap:magvar>
<linqmap:type>JAM</linqmap:type>
<linqmap:subtype>JAM_STAND_STILL_TRAFFIC</linqmap:subtype>
<linqmap:street>Eje 1 Nte. - Fuerza Aérea Mexicana</linqmap:street>
<linqmap:city>Venustiano Carranza</linqmap:city>
<linqmap:country>MX</linqmap:country>
<linqmap:roadType>6</linqmap:roadType>
<linqmap:reportRating>3</linqmap:reportRating>
<confidence>0</confidence>
<linqmap:reliability>5</linqmap:reliability>
</item>
<item>
<title>alert</title>
<pubDate>Tue Jul 4 20:37:18 +0000 2017</pubDate>
<georss:point>19.418267 -99.079619</georss:point>
<linqmap:uuid>d4512759-ce60-3819-b12a-10244f3ddfbe</linqmap:uuid>
<linqmap:magvar>126</linqmap:magvar>
<linqmap:type>JAM</linqmap:type>
<linqmap:subtype>JAM_STAND_STILL_TRAFFIC</linqmap:subtype>
<linqmap:street>Eje 1 Nte. - Fuerza Aérea Mexicana</linqmap:street>
<linqmap:city>Venustiano Carranza</linqmap:city>
<linqmap:country>MX</linqmap:country>
<linqmap:roadType>6</linqmap:roadType>
<linqmap:reportRating>3</linqmap:reportRating>
```

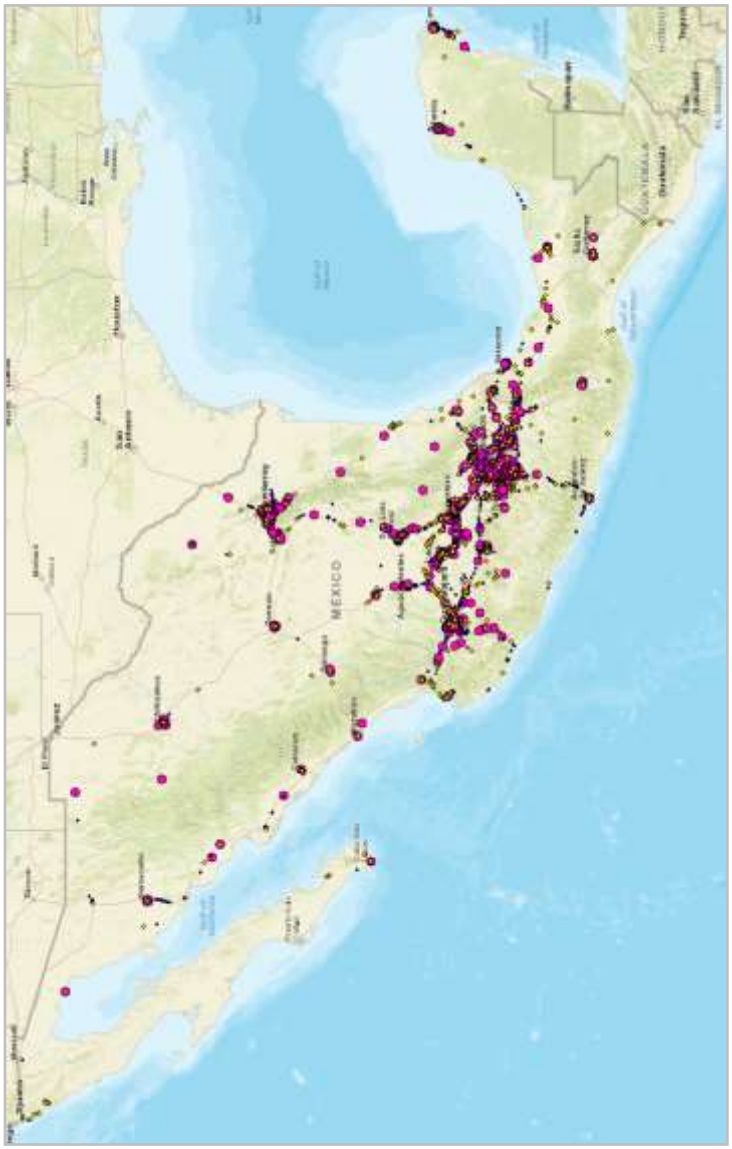


Fuente: waze.com

Figura 4.4 Captura de pantalla del feed en formato XML.

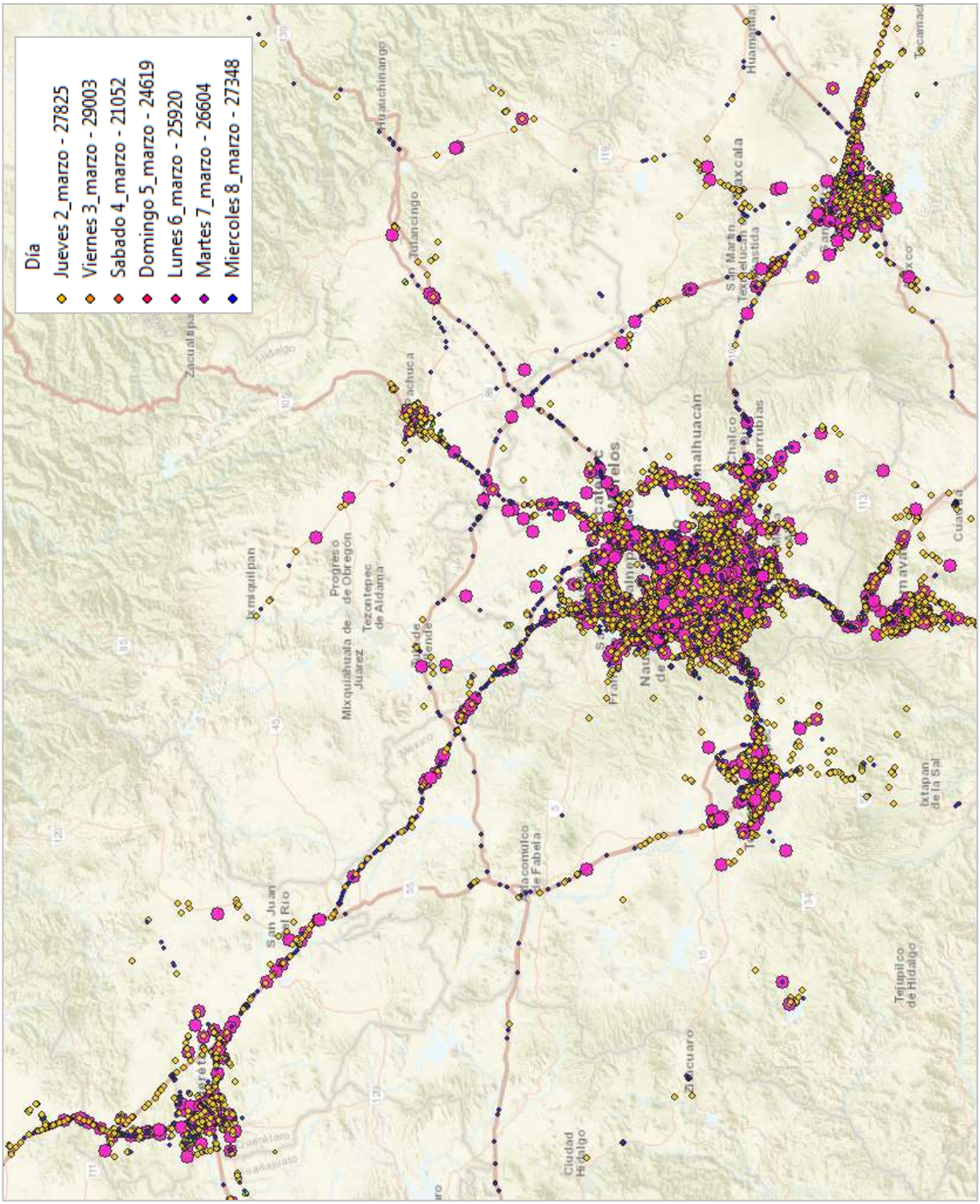
4.4 Proceso y transformación de datos

Dado que la cantidad de datos que se generan es muy grande, se actualiza de manera continua y es difícil de manipular, para el propósito del presente proyecto se seleccionó un periodo de tiempo para procesar solamente los datos de ese intervalo. El periodo seleccionado fue del 2 de marzo de 2017 al 8 de marzo del mismo año. Este conjunto de datos se procesó y convirtió en información geoespacial compatible con el software ArcGIS, ver Figura 4.5.



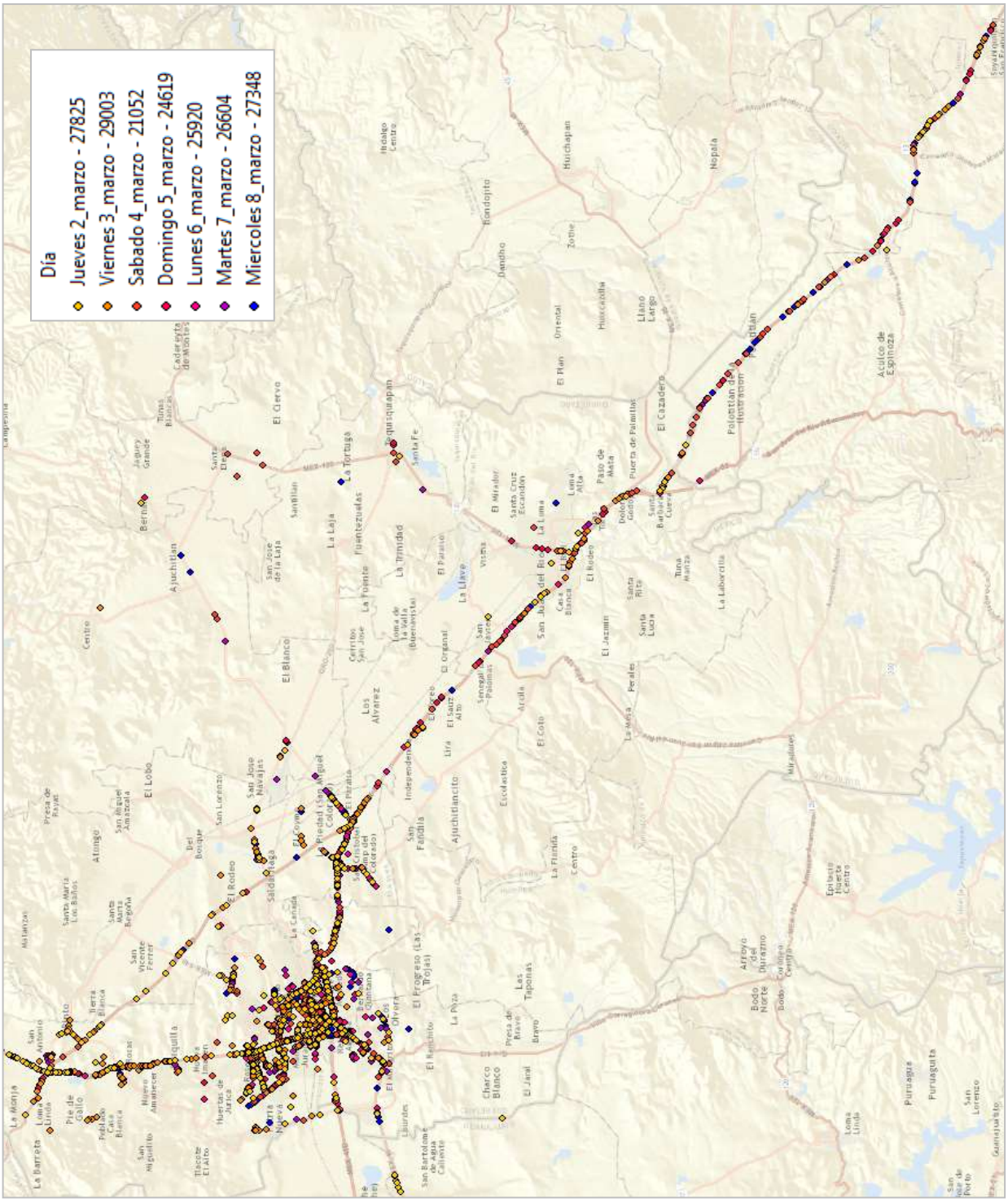
Fuente: Elaboración propia a partir de datos de Waze.

Figura 4.5 Eventos de la semana 2 de marzo al 8 de marzo 2017. Aplica para todo el país. Los circulos definen distintos tipos de eventos, ver Figura 4.6



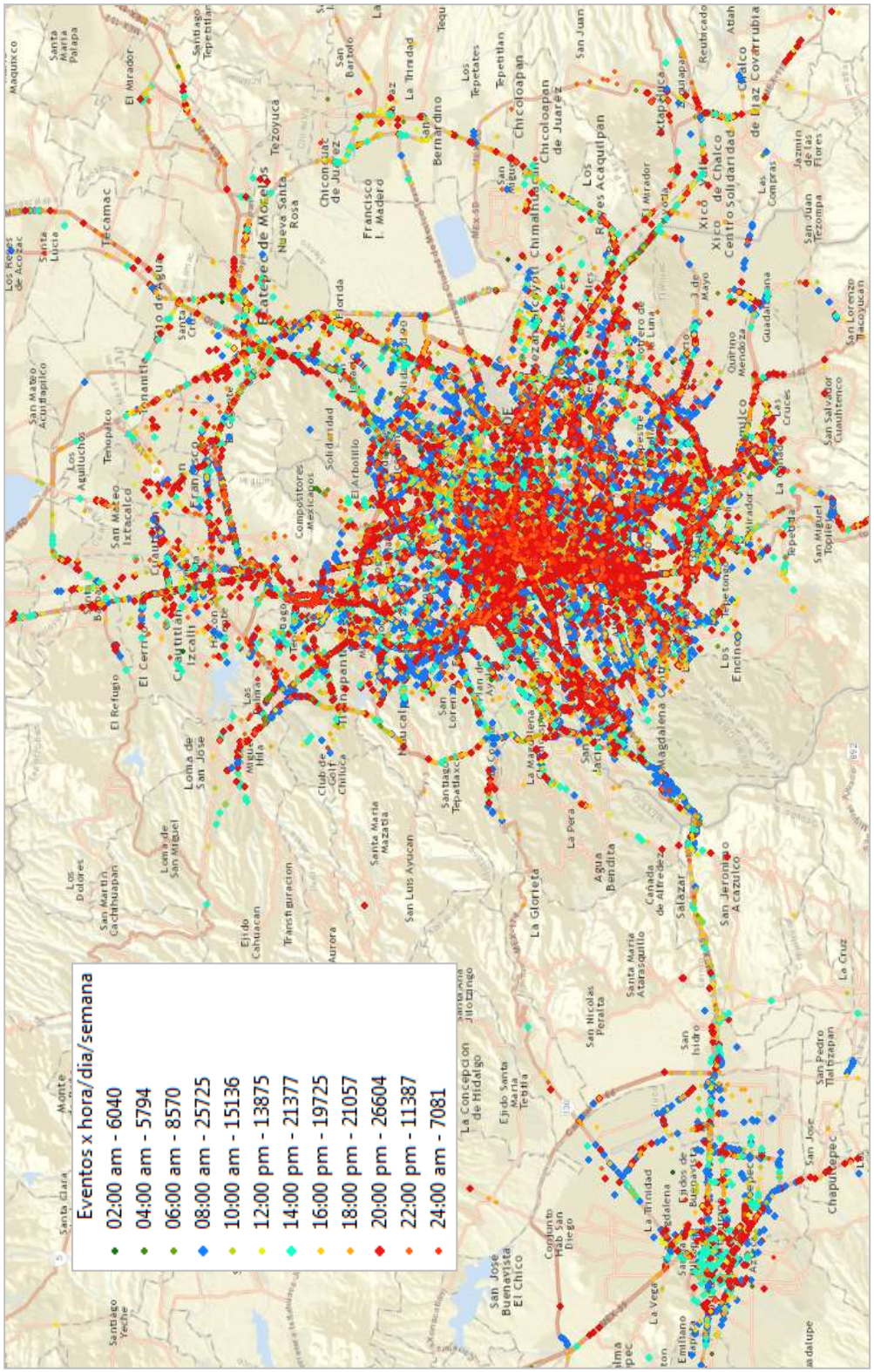
Fuente: Elaboración propia a partir de datos de Waze.

Figura 4.6. Eventos de la semana 2 de marzo al 8 de marzo 2017. Acercamiento a la zona de Querétaro - Ciudad de México - Puebla.



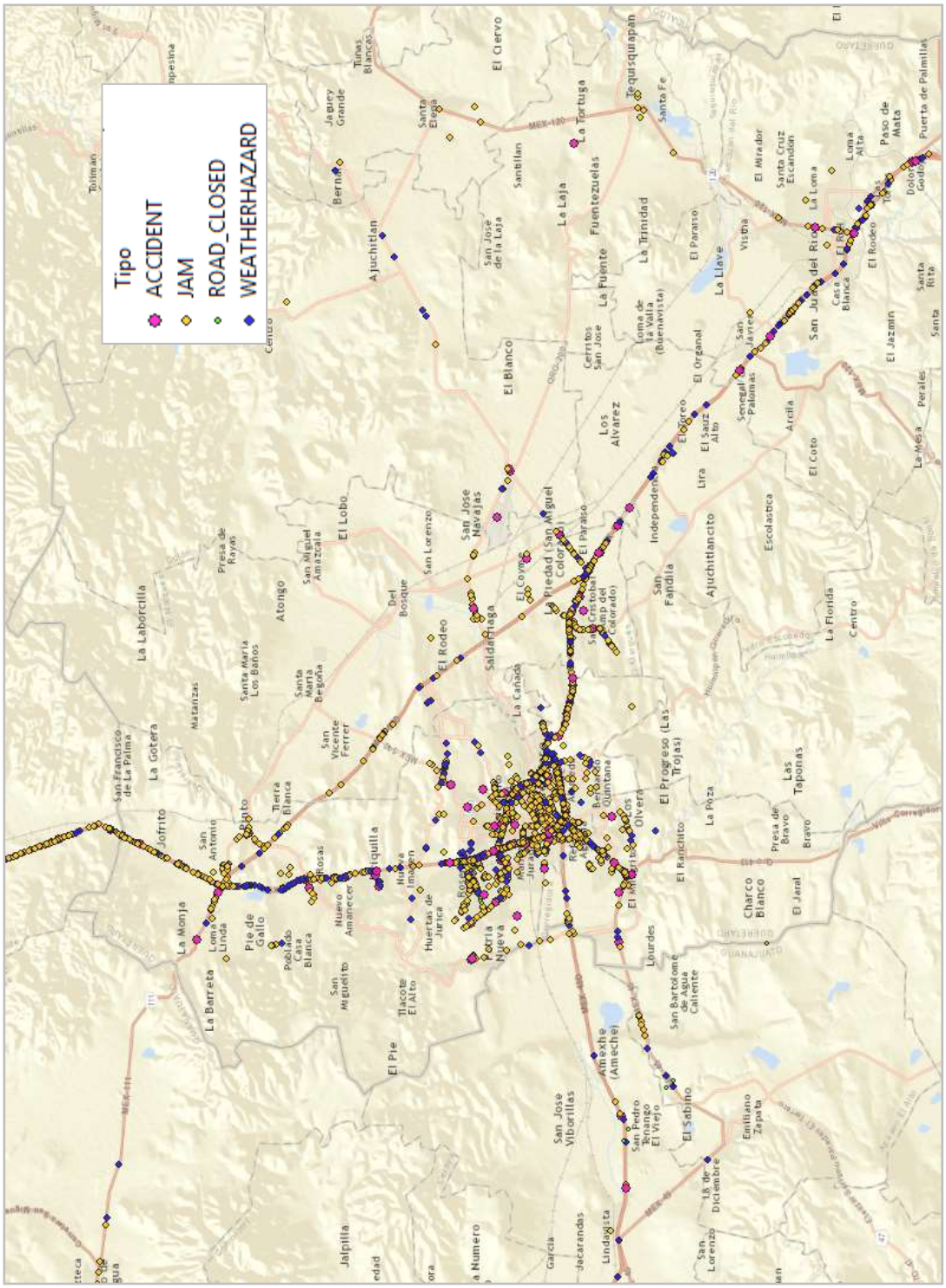
Fuente: Elaboración propia a partir de datos de Waze.

Figura 4.7. Eventos de la semana 2 de marzo al 8 de marzo 2017. Simbolizados por día. Zona de Querétaro - Ciudad de México - Puebla.



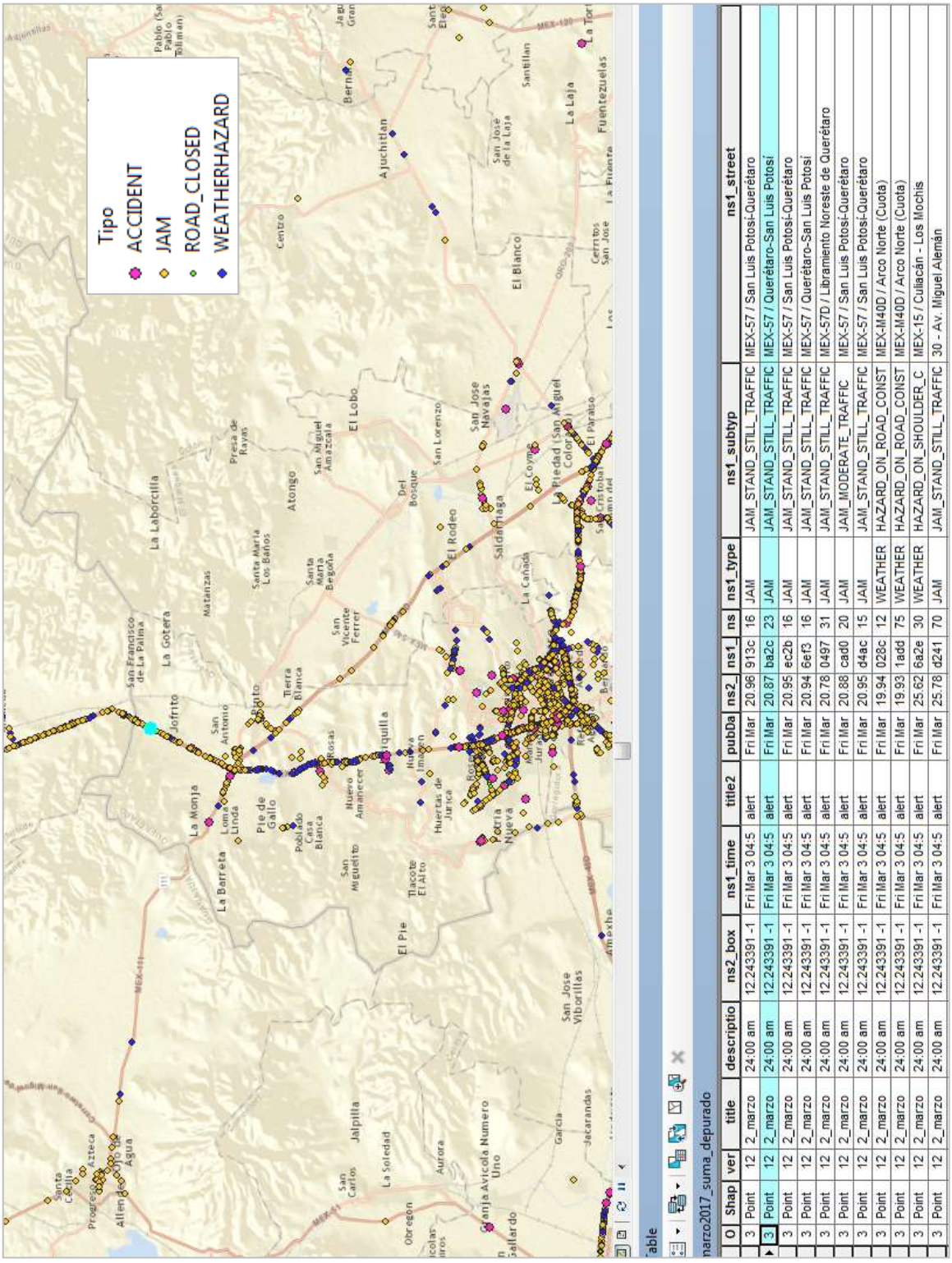
Fuente: Elaboración propia a partir de datos de Waze.

Figura 4.8. Eventos de la semana 2 de marzo al 8 de marzo 2017. Simbolizados por hora. Zona de Ciudad de México - Toluca.



Fuente: Elaboración propia a partir de datos de Waze.

Figura 4.9. Eventos de la semana 2 de marzo al 8 de marzo 2017. Simbolizados por tipo de evento. Zona de Querétaro.



marzo2017_suma_depurado

O	Shap	ver	title	descriptio	ns2_box	ns1_time	title2	pubDa	ns2_	ns1_	ns	ns1_type	ns1_subtyp	ns1_street
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	20.96	913c	16	JAM	JAM_STAND_STILL_TRAFFIC	MEX-57 / San Luis Potosí-Querétaro
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	20.87	ba2c	23	JAM	JAM_STAND_STILL_TRAFFIC	MEX-57 / Querétaro-San Luis Potosí
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	20.95	ec2b	16	JAM	JAM_STAND_STILL_TRAFFIC	MEX-57 / San Luis Potosí-Querétaro
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	20.94	6ef3	16	JAM	JAM_STAND_STILL_TRAFFIC	MEX-57 / Querétaro-San Luis Potosí
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	20.78	0497	31	JAM	JAM_STAND_STILL_TRAFFIC	MEX-57D / Libramiento Noreste de Querétaro
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	20.88	cad0	20	JAM	JAM_MODERATE_TRAFFIC	MEX-57 / San Luis Potosí-Querétaro
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	20.95	d4ac	15	JAM	JAM_STAND_STILL_TRAFFIC	MEX-57 / San Luis Potosí-Querétaro
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	19.94	028c	12	WEATHER	HAZARD_ON_ROAD_CONST	MEX-M40D / Arco Norte (Cuota)
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	19.93	1add	75	WEATHER	HAZARD_ON_ROAD_CONST	MEX-M40D / Arco Norte (Cuota)
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	25.62	6a2e	30	WEATHER	HAZARD_ON_SHOULDER_C	MEX-15 / Culiacán - Los Mochis
3	Point	12	2_marzo	24:00 am	12.243391 -1	Fri Mar 3 04:5	alert	Fri Mar	25.78	d241	70	JAM	JAM_STAND_STILL_TRAFFIC	30 - Av. Miguel Alemán

Fuente: Elaboración propia a partir de datos de Waze.

Figura 4.10. Eventos de la semana 2 de marzo al 8 de marzo 2017. Simbolizados por tipo de evento. Zona de Querétaro

5 Conclusiones

Este estudio revisa una variedad de conceptos sobre los datos masivos, en particular la vertiente de los datos masivos geoespaciales. Aunque no existe una definición estándar de datos masivos, se puede considerar como el conjunto de datos estructurados y no estructurados con volúmenes de datos que no pueden ser capturados, almacenados, manipulados, analizados, administrados y presentados fácilmente por tecnologías tradicionales de hardware, software y bases de datos con y sin coordenadas geográficas que indiquen posición.

Dadas estas características únicas, los métodos tradicionales de manejo de datos a veces son insuficientes y se han identificado las siguientes áreas de oportunidad en el desarrollo e investigación de la disciplina de datos masivos espaciales:

- El desarrollo de métodos de indexación espacial y algoritmos para manejar datos masivos, *streaming*[47] y topología para análisis en tiempo real.
- Los nuevos paradigmas de visualización, especialmente desarrollados para datos masivos, tienden a ser ricos en información. En relación con la gestión de pantallas de visualización complejas, la investigación tecnológica de detalle es muy importante.
- Se requiere el desarrollo de nuevos enfoques para la identificación de errores con el fin de evaluar efectivamente la calidad de los datos.
- Del mismo modo, se requieren métodos de validación de los datos para asegurar la confiabilidad de estos.
- El diseño de métodos de interoperabilidad que permita el intercambio de información entre distintos sistemas o componentes, por ejemplo, los datos generados a través de *Waze* y su aplicación en un SIG (Sistema de Información Geográfica).
- La privacidad y seguridad son igualmente importantes y son preocupaciones clave, sobre todo sí no se abordan adecuadamente. Son una parte esencial de la gestión de datos masivos geoespaciales, pero no se cubren en este estudio ya que el enfoque en este documento es sobre métodos de manejo de datos.

En el futuro, herramientas y modelos específicos trabajarán en conjunto con fuentes de datos producidas por sistemas aún más operativos, rápidos y precisos que los que existen actualmente, tales como datos de tráfico, *feeds* de medios sociales para revelar sentimientos del viajero, demografía de la población, datos geoespaciales, datos meteorológicos y económicos para mejorar la gestión operativa y las actividades de planificación para el transporte.

[47] Transmisión en tiempo real

Los datos de *Waze* procesados en este estudio podrán servir para predecir el impacto en carreteras, autopistas y redes de transporte público causado por cierres o trabajos de mantenimiento en carreteras y recomendar cambios en los cronogramas de tránsito. Se podrá detectar y predecir la probable ocurrencia de incidentes de servicio no planificados como un accidente de tráfico o eventos meteorológicos y con esta información sugerir respuestas óptimas de acción para disminuir los daños y el tiempo de espera.

El futuro del análisis de datos en el transporte tiene muchas aplicaciones y oportunidades. El desafío no es ciertamente la capacidad de generar datos, porque los sistemas y aplicaciones existentes ya están proporcionando más de lo que se está utilizando actualmente. La solución, por lo tanto, está en utilizar medios y métodos significativamente mejorados para recopilar y comprender los datos para que las decisiones que se deban tomar tengan la mejor información posible.

El uso de datos masivos por parte de los responsables de la formulación de políticas debe evaluarse de acuerdo con las preocupaciones específicas y el nivel de conocimiento relacionado con un problema a tratar. Para los desafíos bien definidos en los que se requieren muchas variables y sus interrelaciones, los datos pequeños, cuidadosamente estratificados y representativamente muestreados pueden ser, tal vez, incluso más eficaces que la búsqueda de soluciones con el uso de datos masivos. Sin embargo, para situaciones donde el conocimiento es bajo, el análisis de datos masivos puede ayudar a aclarar preguntas relevantes acerca de algunos aspectos del tema a investigar.

Así, crear una imagen más rica y completa de lo que está sucediendo sobre el terreno es una oportunidad para la analítica geoespacial en el sector de transporte, al aprovechar las herramientas de datos y análisis predictivo para ayudar a las agencias y entidades públicas de transporte a mejorar las operaciones, reducir los costos y servir mejor a los viajeros. Sin embargo, antes de que las agencias de transporte lleguen a esos resultados, hay varios obstáculos que deben eliminarse utilizando las herramientas tecnológicas disponibles y un proceso de datos a medida.

Como se mencionó anteriormente, ya existen conjuntos de datos masivos en las organizaciones de transporte, por lo que estas entidades y dependencias deben decidir cuál de ellas utilizar y en dónde buscar datos complementarios. Este primer obstáculo del proceso de recolección de datos incluye la extracción y recopilación de datos en cualquier formato desde cualquier lugar en el que se encuentren disponibles. Luego deben ser colocados en un repositorio de datos, preferiblemente uno que no esté restringido por las restricciones tradicionales de gestión de datos, antes de ser limpiado y organizado en modelos estandarizados adecuados para el análisis.

Enseguida, se debe superar el siguiente obstáculo del proceso, que es comprender la historia contada por los datos, esto ocurre mediante la fusión de datos específicos que fueron limpiados y organizados en el primer paso, con otros datos modelados y aplicando técnicas de análisis predictivo para llegar a un cuadro más completo. El análisis estadístico, la simulación y la optimización se pueden aplicar para explotar las relaciones identificadas en los datos, para planificar y predecir lo que es probable que suceda en diferentes escenarios. Estas ideas deben ser presentadas en visualizaciones de datos que sean convincentes e informes intuitivos que revelen y comuniquen la información que importa.

Los datos masivos geoespaciales presentan desafíos y oportunidades, en particular para la gestión integral del transporte. En este documento se esbozan algunos de estos desafíos desde una perspectiva técnica y conceptual, y también se presentan las áreas prioritarias que deben ser abordadas en el futuro. Una vez que la investigación de datos masivos geoespaciales madure, las oportunidades para la gestión global de la sociedad y la toma de decisiones se vuelven enormes.

El despliegue de datos de Twitter en el ejemplo desarrollado en este estudio permite la identificación de tweets que contengan coordenadas, así como el cumplimiento de criterios de búsqueda específicos que se relacionen con el transporte y sus áreas afines, es necesario ahondar la investigación en esta área para recabar más datos, procesarlos e identificar patrones espaciales de ocurrencia en esta plataforma social.

En cuanto al ejemplo de utilización de datos de *Waze*, en donde se observan los incidentes de tráfico, peligros y eventos meteorológicos reportados por sus 50 millones de usuarios, es posible vincular estos eventos con herramientas de visualización más predictivas en los centros de gestión del tráfico. Así, serán posibles las visualizaciones en intervalos de tiempo donde se podrá observar cómo se propaga la congestión desde un punto o zona y, posteriormente implementar medidas de mitigación.

En la Unidad de Sistemas de Información Geoespacial se ha estado trabajando para identificar el potencial de utilización de estas herramientas, así como para obtener fuentes de información actualizada y verificada que puedan utilizarse para las líneas de investigación aplicada y desarrollo que están en marcha. Todo esto con la finalidad de proponer soluciones y metodologías que pueden apoyar a la labor sustantiva de la Secretaría de Comunicaciones y Transportes en cuanto a la operación, seguridad, conservación, planeación, seguimiento y administración en general de la infraestructura carretera del país.

6 Bibliografía

- Barbier Geoffrey; Liu Huan. *Data mining in social media*. Arizona State University. 2011. https://pdfs.semanticscholar.org/8a60/b082aa758c317e9677beed7e7776acde5e4c.pdf

- Calabrese F. *Understanding individual mobility patterns from urban sensing data: A mobile phone trace example*. Department of Real Estate, National University of Singapore, Singapore. 2013.

- Chen, P. Zhang, C. *Data-intensive applications, challengues, techniques and technologies: A survey on Big data*. University of Macau, China. 2014.

- Cöltekin, Arzu. Reichenbacher, Tumasch. *High Quality Geographic Services and Bandwidth Limitations*. GIScience-GIVA, Department of Geography, University of Zurich, Switzerland. 2011

- Cukier K. *Big Data. La revolución de los datos masivos*. Turner.

- Dasgupta Arup. *The Continuum: Big Data, Cloud & Internet of Things*. IBM Blog. EUA. 2017. https://www.ibm.com/blogs/internet-of-things/big-data-cloud-iot/

- Diebold Francis X. *A personal perspective on the origins and development of "big data": the phenomenon, the term and the discipline*. University of Pennsylvania. 2012.

- Dyson, George. *The Big Data Problem Will Also Be A Problem For Science 2.0*. edge.org

- Gandomi A. Haider M. *Beyond the hype: big data concepts, methods and analytics*. Ryerson University. Ontario, Canada. 2014

- Goodchild, M. F. *Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0*. International Journal 2007

- IBM. *The four V´s of Big Data*. 2014

- Labrinidis A. Jagadish H. *Challenges and Opportunities with Big Data*. Universidad de Pittsburgh. Universidad de Michigan. 2015.

- Laney Doug. *3D Data Management: Controlling Data Volume, Velocity and Variety*. MetaGroup. 2001.

- Lee, R. *Big Data & Analytics*. Executive Consulting Services LLC.
- Li, Songnian; Dragicevic, Suzana; Anton, François; Sester, Monika; Winter, Stephan; Coltekin, Arzu; Pettit, Chris; Jiang, Bin; Haworth, James; Stein, Alfred; Cheng, Tao. *Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges.* ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 115, 2016.
- Lohr Steve. *For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights*. The New York Times. 2014. https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html
- Manyika, James; Chui, Michael; Brown, Brad; Bughin, Jacques; Dobbs, Richard; Roxburgh, Charles; Hung Byers, Angela. *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute. 2011
- Muñiz G. Rafael. *Marketing en el siglo XXI*. 5ª Edición. Centro de Estudios Financieros. España. 2014.
- PCAST (President´s Council of Advisors on Scince and Technology). *Big Data and privacy: a technological perspective*. White House. EUA. 2014.
- Richter Kai-Florian, Winter Stephan. *Citizens as database: Conscious ubiquity in data collection*. International Symposium on Spatial and Temporal Databases. 2011. Lecture Notes in Computer Science, vol 6849. Springer, Berlin, Heidelberg
- Shekhar S. *Spatial Big Data*. Department of Computer Scince and Engineering, University of Minessota. 2012. http://www-users.cs.umn.edu/~shekhar/talk/2013/13.5.sbd.asu.pdf
- Wade R. *Data Detour: Analytics Will Move Transportation Forward*. Wired. https://www.wired.com/insights/2014/07/data-detour-analytics-will-move-transportation-forward/

Anexo 1. Antecedentes del Big Data (cronología)

Fuente: http://www.winshuttle.es/big-data-historia-cronologica/

1880
El comienzo de la sobrecarga de información.
El Censo de los Estados Unidos del año 1880 tardó ocho años en tabularse, y se calcula que el censo de 1890 hubiera necesitado más de 10 años para procesarse con los métodos disponibles en la época. Si no se hubieran realizado avances en la metodología, la tabulación no habría finalizado antes de que tuviera que realizarse el censo de 1900.

1881
La máquina tabuladora de Hollerith.
La influencia de los datos del censo derivó en la invención de la máquina tabuladora de Hollerith (tarjetas perforadas), que fue capaz de domar esta ingente cantidad de información y permitir realizar el trabajo aproximadamente en un año. Hizo que Hollerith se convirtiera en emprendedor, y su empresa pasó a formar parte de lo que hoy en día conocemos como IBM.

1932
El boom del crecimiento demográfico.
La sobrecarga de información prosiguió con el aumento desmesurado de la población en los Estados Unidos, la emisión de los números de la seguridad social y el crecimiento general del conocimiento (y la investigación), aspectos que exigían un registro de la información más preciso y organizado.

1940
El efecto en las bibliotecas.
Las bibliotecas, fuente original de la organización y el almacenamiento de datos, tuvieron que adaptar sus métodos de almacenamiento para responder al rápido aumento de la demanda de nuevas publicaciones e investigación.

1941
La explosión de la información.
Los académicos comenzaron a denominar a esta increíble expansión de la información como la «explosión de la información». Tras aparecer por primera vez en el periódico *Lawton Constitution* en el año 1941, el término se desarrolló en un artículo del *New Statesman* en marzo del año 1964, en el que se hacía referencia a la dificultad que suponía gestionar los volúmenes de información disponibles.

1944
El primer aviso del problema del almacenamiento y la recuperación de datos.

La primera señal de aviso sobre el crecimiento del conocimiento como problema inminente a la hora de almacenar y recuperar los datos tuvo lugar en 1944, cuando Fremont Rider, bibliotecario de la Universidad Wesleyana, calculó que las bibliotecas de las universidades de EU, duplicaban su tamaño cada dieciséis años. Fuente: The Coming of post-industrial society. Daniel Bell. Basil Book. 1973.

1948
La teoría de la información de Shannon.
Claude Shannon publicó la Teoría matemática de la comunicación, en la que se estableció un marco de trabajo para determinar los requisitos de datos mínimos para transmitir la información a través de canales afectados por ruido (imperfectos). Fue un trabajo histórico que ha hecho posible gran parte de la infraestructura actual. Sin su teoría, el volumen de los datos sería mucho mayor que el actual. Utilizó como referencia «*Certain Factors Affecting Telegraph Speed*», una obra de Nyquist,que permitió muestrear señales analógicas y representarlas digitalmente, lo que es la base del procesamiento de datos moderno. Fuente: http://hspencer.info/2008/09/la-naturaleza-de-la-informacion/

1956
Memoria virtual.
El concepto de memoria virtual fue desarrollado por el físico alemán Fritz-Rudolf Güntsch, como una idea que trataba el almacenamiento finito como infinito. El almacenamiento, administrado mediante hardware integrado y software para ocultar los detalles al usuario, permitió procesar los datos sin las limitaciones de memoria de hardware que anteriormente provocaban la partición del problema.

1961.
El conocimiento científico se amplía.
La revolución científica era la responsable de la comunicación rápida de ideas nuevas como información científica. Este rápido crecimiento se materializaba en la duplicación cada 15 años de los registros nuevos creados.

1962
Los pioneros en el reconocimiento de voz.
Los científicos han trabajado en el reconocimiento de voz casi desde que empezaron a fabricar ordenadores. En el año 1962, William C. Dersch de IBM desveló la máquina Shoebox en la Feria Mundial. Fue la primera máquina capaz de entender 16 palabras y diez dígitos en inglés hablado mediante el uso de los datos disponibles en ese momento y era capaz de procesarlos correctamente. Sin embargo, hasta transformar esta innovación en el reconocimiento de voz en productos con una utilidad comercial real, aún quedaba mucho camino por delante. Este camino exigiría avances importantes en la potencia de procesamiento y la reducción del costo de la tecnología informática. La existencia de un volumen de datos mayor también ayudaría a entrenar los sistemas de reconocimiento de voz.
http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/

1963
En busca de una solución organizativa.
A principios de la década de 1960, Price observó que la enorme mayoría de investigación científica suponía un esfuerzo abrumador para los humanos. Los resúmenes documentales, creados a finales de la década de 1800 como forma de gestionar los conocimientos, cada vez de mayor volumen, crecían también con la misma progresión (multiplicándose por un factor de diez cada cincuenta años), y ya habían alcanzado una magnitud preocupante. Habían dejado de ser una solución de almacenamiento o de organización de la información.

1966
Entran en escena los sistemas de computación centralizados.
La información no solo se encontraba en pleno auge en el sector científico, también lo estaba en el sector de los negocios. Debido a la influencia que tuvo la información en la década de 1960, la mayoría de organizaciones empezaron a diseñar, desarrollar e implementar sistemas informáticos que les permitían automatizar los sistemas de inventario.

1970
Base de datos relacional.
En el año 1970, Edgar F. Codd, un matemático formado en Oxford que trabajaba en IBM Research Lab, publicó un artículo en el que se explicaba la forma en la que podía accederse a la información almacenada en bases de datos de gran tamaño sin saber cómo estaba estructurada la información o dónde residía dentro de la base de datos. Hasta ese momento, para recuperar la información se necesitaban conocimientos informáticos relativamente sofisticados, e incluso hasta los servicios de especialistas, por lo que se convertía en una tarea ardua que exigía tiempo y recursos económicos. Hoy en día, la mayoría de transacciones de datos rutinarias —acceder a cuentas bancarias, utilizar tarjetas de crédito, comerciar con acciones, realizar reservas de viaje, realizar compras a través de Internet— utilizan estructuras basadas en la teoría de la base de datos relacional. http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/reldb/

1975
El crecimiento de la comunicación bidireccional.
El Censo del Flujo de la Información, realizado por el Ministerio de Correos y Telecomunicaciones de Japón, comenzó a realizar un control del volumen de información que circulaba por el país en 1975. Utilizando como unidad de medición el número de palabras utilizadas a través de todos los medios de comunicación, el estudio pudo comprobar que el suministro de información superaba considerablemente al volumen de información consumida y que la demanda de comunicación unidireccional se había estancado. Ahora, la tendencia es el aumento de la demanda de comunicación bidireccional, más personalizada y que responde a las necesidades de las personas.
Fuente: Information Society Studies. Alistair S. Duff, 2000.

1976

Sistemas de Planificación de necesidades de material (MRP).
A mediados de la década de 1970, los sistemas de Planificación de necesidades de material (MRP) se diseñaron como herramienta que ayudaba a las empresas de fabricación a organizar y planificar su información. A esas alturas, la popularidad de las computadoras personales en las empresas estaba en auge. Esta transformación marcó un cambio de tendencia hacia los procesos de negocio y las funcionalidades de contabilidad, y en este ámbito se fundaron empresas como Oracle, JD Edwards y SAP. Fue Oracle la que presentó y comercializó el Lenguaje de Consulta Estructurado o *Structure Query Language* (SQL). *Factory Physics, To Pull or Not to Pull: What is the Question?*
http://pubsonline.informs.org/doi/pdf/10.1287/msom.1030.0028

Abril 1980
La ley de los datos de Parkinson.
A medida que aumentaba la velocidad con la que se creaba información, las opciones de almacenamiento y organización de datos eran cada vez menores. En su charla «*Where Do We Go From Here*?», I.A. Tjomsland afirmó que «aquellos que trabajan en dispositivos de almacenamiento descubrieron hace mucho tiempo que la primera ley de Parkinson puede parafrasearse para describir nuestro sector: “los datos se expanden para llenar el espacio disponible”. Desde mi punto de vista, las grandes cantidades de datos se guardan porque los usuarios no tienen forma de identificar los datos obsoletos; las penalizaciones derivadas de almacenar datos obsoletos tienen una importancia inferior a las que conlleva eliminar datos potencialmente útiles».

Agosto 1983
El crecimiento de la información y el sector de la comunicación.
Los avances tecnológicos permitieron a todos los sectores beneficiarse de nuevas formas de organizar, almacenar y generar datos. Las empresas estaban empezando a usar los datos para tomar mejores decisiones de negocio. En el artículo *Tracking the Flow of Information*, publicado en la revista *Science*, el autor Ithiel de Sola Pool analizó el crecimiento del volumen de información de 17 importantes medios de comunicación desde el año 1960 hasta 1977. Atribuye el enorme crecimiento de la información a la expansión del sector de las comunicaciones.

1985
Sistemas de Planificación de recursos de fabricación (MRP II).
Tras el auge de los sistemas de MRP, se introdujo la Planificación de recursos de fabricación (MRP II) en la década de 1980, con un énfasis en la optimización de los procesos de fabricación mediante la sincronización de materiales con las necesidades de producción. MRP II incluía áreas tales como la gestión del área de producción y la distribución, la gestión de proyectos, las finanzas, los recursos humanos y la ingeniería. No fue hasta mucho después de adoptar esta tecnología cuando otros sectores (p. ej. agencias gubernamentales y organizaciones del sector

servicios) comenzaron a tener en cuenta, y posteriormente adoptar, la tecnología ERP.

1985
La necesidad de datos precisos.
En el año 1985, Barry Devlin y Paul Murphy definieron una arquitectura para los informes y análisis de negocio en IBM (*Devlin & Murphy, IBM Systems Journal 1988*) que se convirtió en la base del almacenamiento de datos. En el centro neurálgico de dicha arquitectura, y en el almacenamiento de datos en general, se encuentra la necesidad de almacenamiento homogéneo y de alta calidad de datos históricamente completos y exactos.

Julio 1986
Desde las tablillas de barro hasta la memoria de semiconductores.
En su artículo «*Can users really absorb data at today's rates? Tomorrow's*?», Hal Becker mencionaba que «la densidad de recodificación lograda por Gutenberg fue aproximadamente de 500 símbolos (caracteres) por pulgada cúbica; 500 veces la densidad de las tablillas de barro [sumerias del año 4000 antes de cristo]. En el año 2000, la memoria de acceso aleatorio de los semiconductores será capaz de almacenar 1,25×10^11 bytes por pulgada cúbica».

1988
La superficie de los nuevos sistemas de software.
A finales de la década de los 80 y principios de los 90, fuimos testigos del aumento de los sistemas de Planificación de recursos empresariales (ERP), ya que pasaron a ser más sofisticados y ofrecían la posibilidad de coordinarse e integrarse entre todos los departamentos de las empresas. Las bases tecnológicas de los sistemas de MRP, MRP II y ERP comenzaron a integrar áreas de empresas entre las que se incluían la producción, la distribución, la contabilidad, las finanzas, los recursos humanos, la gestión de proyectos, la gestión del inventario, el mantenimiento y el transporte, y ofrecer así accesibilidad, visibilidad y homogeneidad en la totalidad de la empresa. Fuente: Enterprise Resource Planning: Global Opportunities & Challenges

1989
Inteligencia empresarial.
En 1989, Howard Dresner amplió el popular término genérico «*Business Intelligence (BI)*» o Inteligencia empresarial, inicialmente acuñado por Hans Peter Luhn en el año 1958. Dresner lo definió como los «conceptos y métodos que mejoran la toma de decisiones de negocio mediante el uso de sistemas de apoyo basados en datos reales». Poco tiempo después, y como respuesta a la necesidad de una mejor BI, se pudo ver el auge de empresas como *Business Objects, Actuate, Crystal Reports y MicroStrategy*, que ofrecían informes y análisis de los datos de las empresas. Fuente: A Brief History of Decision Support Systems de D.J. Power.

1992
El primer informe de base de datos.
En 1992, Crystal Reports creó el primer informe de base de datos sencillo con Windows. Estos informes permitían a las empresas crear un informe sencillo a partir de diversos orígenes de datos con escasa programación de código. De esta forma, se redujo la presión existente sobre el panorama saturado de datos y se permitió que las empresas emplearan la inteligencia empresarial de un modo asequible.

1995
Explosión de la World Wide Web.
En la década de 1990 se produjo un crecimiento tecnológico explosivo y los datos de la Inteligencia empresarial comenzaron a apilarse en forma de documentos de Microsoft Excel.

1996
El espectacular crecimiento de la potencia informática e Internet.
El aumento desmesurado del volumen de datos supuso otros problemas para los proveedores de sistemas ERP. La necesidad de tener que diseñar de nuevo los productos ERP, y que incluía romper los límites de titularidad y de personalización, obligó a los proveedores a adoptar de forma gradual un método de negocio colaborativo, en lugar de la intranet.

1996
Inteligencia empresarial 2.0.
La influencia de la información trajo consigo un nuevo problema en la gestión de los datos, además de un aumento del coste que suponía publicarla y almacenarla. Como los datos resultaban más difíciles de mantener, para poder ofrecer más funcionalidades, el almacenamiento digital empezó a resultar más rentable que el papel para almacenar los datos y comenzaron a emerger las plataformas de BI.

Julio 1997
El problema del Big Data.
El término «*Big Data*» se empleó 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. A esto se denominó el «problema del Big Data».
Fuente: Application-Controlled Demand Paging for Out-of-Core Visualization. https://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf

Agosto 1997
El futuro del almacenamiento de datos.
Michael Lesk publicó *How much information is there in the world*? Su conclusión fue que «Puede que la cantidad de información ascienda a varios miles de petabytes, y la producción de cinta y disco alcanzará ese nivel en el año 2000. Esto significa que,

en unos años, (a) podremos guardarlo todo, no será necesario eliminar información, y que (b) la mayoría de la información jamás será consultada por un ser humano».

1998
El problema de la Inteligencia empresarial.
A finales de los 90, muchas empresas creían que sus sistemas de extracción de datos no funcionaban. Los trabajadores eran incapaces de encontrar respuestas y de acceder a los datos que necesitaban de las búsquedas. Los departamentos informáticos eran responsables del 80% del acceso. Cada vez que los empleados necesitaban acceso, tenían que llamar al departamento informático, ya que acceder a la información no resultaba tan fácil.

1999
Internet de las cosas (IoT).
El término "Internet de las cosas" o *IoT*, por sus siglas en inglés, fue acuñado por el emprendedor británico Kevin Ashton, cofundador del Auto-ID Center del MIT, durante una presentación que mostraba la idea de identificación por radiofrecuencia (RFID) en los componentes de la cadena de suministro. "Si tuviéramos equipos que supieran todo lo que hay que saber acerca de las cosas, a partir de datos que recopilados sin nuestra ayuda, seríamos capaces de monitorizar y contar todo, y reducir así considerablemente los costes, los desperdicios y las pérdidas."
Fuente: RFID Journal. http://www.rfidjournal.com/articles/view?4986

1999
Se cuantifica la información.
Peter Lyman y Hal R. Varian de UC Berkeley publicaron el primer estudio que cuantificaba, en términos de almacenamiento informático, la cantidad total de información nueva y original creada en el mundo al año. El estudio, titulado *How Much Information?*, se completó en 1999, un año en el que el mundo produjo unos 1,5 exabytes de información.

Julio 1999
El análisis predictivo cambia el perfil del negocio.
ComputerWeekly cuenta con un artículo destacado en el que se explica la forma de elegir e instalar la solución ERP adecuada y que al utilizar pronósticos de análisis predictivo se cambia el método de trabajo de todo tipo de organizaciones. Fuente: Choosing and Installing the Right ERP Solution.
http://www.computerweekly.com/feature/Choosing-and-installing-the-right-ERP-solution

2001
Software como servicio (SaaS).
Las siglas SaaS aparecen por primera vez en un artículo de la división de comercio electrónico de *Software & Information Industry* (SIIA).

Febrero 2001
Las tres V.

Doug Laney, analista de Gartner, publicó un artículo titulado *3D Data Management: Controlling Data Volume, Velocity, and Variety*. A día de hoy, las tres V siguen siendo las dimensiones comúnmente aceptadas del Big Data.

2002
Sistemas ERP ampliados.
Durante la década de los 90, los proveedores de sistemas ERP añadieron más módulos y funciones como complementos de los módulos básicos, con lo que surgieron los sistemas ERP extendidos o ampliados. El número de opciones de software y hardware aumentó exponencialmente y a principios de la década del 2000, comenzaron a surgir importantes empresas de software. Oracle y SAP fueron las principales empresas de software ERP que sobrevivieron a este auge. Fuente: The Evolution of ERP Systems: A Historical Perspective.
https://faculty.biu.ac.il/~shnaidh/zooloo/nihul/evolution.pdf

Junio 2002
Servicios web y ERP.
Los principales proveedores de sistemas ERP, como SAP, *PeopleSoft*, Oracle y JD Edwards, comenzaron a centrarse en el uso de servicios web para enlazar sus propios conjuntos de aplicaciones, y en facilitar a los clientes la creación de aplicaciones nuevas a partir de datos de varias aplicaciones utilizando XML. Fuente: ComputerWeekly, ERP Giants Prepare for Web Services

Marzo 2005
El enfoque en la usabilidad del usuario final.
Las empresas de SaaS entraron en escena para ofrecer una alternativa a Oracle y SAP más centrada en la usabilidad del usuario final. Una de las primeras fusiones de empresas fue la que dio origen a Workday, Inc., fundada en marzo de 2005, como alternativa a Oracle y SAP, más económica y utilizable. El software de Workday Inc. es más intuitivo para el usuario final y funciona de la misma forma que la gente, «de forma colaborativa, sobre la marcha y en tiempo real».

Septiembre 2005
La gestión de la base de datos, el centro del universo.
Tim O'Reilly publicó *What is Web 2.0?* donde afirma que «los datos son el próximo *Intel Inside*©». En el artículo, O'Reilly afirma lo siguiente: «Como Hal Varian apuntó en una conversación personal el año pasado, "el SQL es el nuevo HTML". La gestión de la base de datos es una competencia básica de las empresas Web 2.0, tanto que a veces denominamos a estas aplicaciones "infoware" en lugar de simplemente software».

2006
Una solución de código abierto para la explosión del Big Data.
Hadoop se creó en el año 2006 a raíz de la necesidad de sistemas nuevos para gestionar la explosión de datos de la web. De descarga gratuita y libre para potenciarlo y mejorarlo, Hadoop es un método de código abierto para almacenar y procesar los datos que «permite el procesamiento paralelo distribuido de enormes

cantidades de datos en servidores estándar del sector, económicos, que almacenan y procesan los datos, y que pueden escalarse sin límite».
Fuente: HADOOP: Scalable, Flexible Data Storage and Analysis de Mike Olson

Marzo 2007
El primer estudio que calcula y prevé la cantidad de crecimiento de la información.
Los investigadores de International Data Corporation publicaron un artículo titulado *The Expanding Digital Universe: A Forecast of Worldwide Information Growth through 2010*, en el que se calcula y pronostica la cantidad de datos digitales que se crearán y reproducirán cada año. En el artículo se calcula que, solo en el año 2006, se crearon en todo el mundo 161 exabytes de datos, y prevé que, en los próximos cuatro años, la información creada aumentará hasta multiplicarse por seis (hasta los 988 exabytes). En otras palabras, predicen que la información se duplicará cada 18 meses durante los próximos cuatro años. Si se consultan los informes de los años 2010 y 2012, la cantidad de datos digitales creados cada año superó los pronósticos iniciales (1227 en 2010 y 2837 exabytes en 2012).

2008
La explosión de datos continúa.
Bret Swanson y George Gilder proyectaron que el tráfico IP estadounidense podría alcanzar el zettabyte en el año 2015, y que la Internet estadounidense del 2015 será, como mínimo, 50 veces más grande que lo era en el 2006.
Fuente: whatsthebigdata.com, Gil Press

Junio 2008
El aluvión de datos hace que el método científico quede obsoleto.
El término «Big Data» comenzó a utilizarse cada vez con más frecuencia en los círculos tecnológicos. La revista *Wired* publicó un artículo en el que se presentaba el impacto positivo y negativo del aluvión de datos reciente. En este artículo, *Wired* anunció que éste era el «principio de la era del petabyte». A pesar de que era una buena hipótesis, la clasificación de «petabyte» era demasiado técnica para el público en general. Inevitablemente, un petabyte, que equivale a 1.000.000.000.000.000 bytes de datos, dará paso dentro de poco a tamaños de datos todavía mayores: exabytes, zettabytes y yottabytes.
Fuente: Wired Magazine, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

Noviembre 2008
SAP desvela su estrategia de SaaS.
SAP realizó un movimiento estratégico de cara al mercado de SaaS mediante el desarrollo de una estrategia de software como servicio destinada a las grandes empresas. Como parte del mismo, se contrató a John Wookey (antiguo empleado de Oracle) en noviembre de 2008 como nuevo jefe de aplicaciones de software a la carta para grandes empresas. Tras presentar su propuesta a la junta en el mes de enero, desarrollaron un plan para lanzar las nuevas ofertas de productos de SaaS

en series de aplicaciones de software concretas para cada función. Estas aplicaciones, disponibles por suscripción, se conectan con los sistemas SAP Business Suite in situ que SAP alojará en régimen de tenencia múltiple.
Fuente: InformationWeek, SAP Unveils SaaS Strategy.
http://www.informationweek.com/cloud/software-as-a-service/sap-unveils-saas-strategy/d/d-id/1080351

Diciembre 2008
Avances revolucionarios.
Un grupo de investigadores científicos en el ámbito de la informática publicó el artículo titulado *Big Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science, and Society*. En el artículo se afirma lo siguiente: «De la misma forma que los motores de búsqueda han cambiado la forma de acceder a la información, otras formas de informática de Big Data pueden transformar y transformarán las actividades de empresas, investigadores científicos, médicos y las operaciones de defensa e inteligencia de nuestra nación…. Probablemente, la informática para Big Data sea la mayor innovación informática de la última década. A día de hoy, tan solo hemos visto el potencial que tiene para recopilar, organizar y procesar los datos en todos los aspectos de nuestras vidas. Si el gobierno federal efectuara una modesta inversión, su desarrollo e implantación podrían acelerarse enormemente». Este apoyo hizo que el Big Data finalmente lograra la credibilidad intelectual que necesitaba.

Enero 2009
La Inteligencia empresarial (*Business Intelligence*) pasa a ser una prioridad.
En el año 2009, la Inteligencia empresarial pasó a ser una de las principales prioridades para los directores de tecnologías de la información.
Fuente: Gartner.com

1 Febrero 2009
Linked Data.
Tim Berners-Lee, director del World Wide Web Consortium (W3C) e inventor del World Wide Web, fue el primero en usar el término "linked data" (datos enlazados) durante una presentación sobre el tema en el congreso TED de 2009. Linked Data describe un método de publicación de datos estructurados, basado en protocolos web estándar, para que puedan ser interconectados, leídos automáticamente por ordenadores y enlazados desde otros conjuntos de datos externos.
Fuente: Linked Data - The Story so far.
http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

Mayo 2009
Análisis ERP
Gartner predijo que los datos empresariales crecerían un 650 % durante los próximos cinco años. Estos datos representan el conglomerado de todos los datos operativos de ERP internos, además de los datos externos que tienen interconexión con las operaciones de la empresa; más allá de los datos de proveedores, hasta llegar a los datos económicos globales (estadísticas macroeconómicas o

microeconómicas). Jon Reed, analista independiente, mentor de SAP y bloguero en JonERP.com, no ve demasiado claro que Google se lance a crear un conjunto de aplicaciones ERP o a comprar una empresa de sistemas ERP —«sería una decisión extrema», afirma. Sin embargo, «si una empresa tipo Google fuera capaz de presentar una forma de recopilar toda esta información de forma conjunta en un entorno basado en la nube, para posteriormente conectarla de alguna forma a una plataforma estructurada —uniendo la información estructurada y la información no estructurada— estaríamos ante un hito muy importante».
Fuente: CIO, The Future of ERP, Part II

Diciembre 2009
¿Cuánta información?
El estudio *How Much Information*? 2009 *Report on American Consumers*, realizado por Global Information Industry Center, revela que, en el año 2008, «los americanos consumieron la información equivalente a unos 1,3 billones de horas, lo que supone una media de 12 horas al día. El consumo total fue de 3,6 zettabytes y de 10.845 billones de palabras, lo que equivale a una media de 100.500 palabras y 34 gigabytes por persona al día». Este estudio tuvo su continuación en un informe, realizado en enero de 2011, titulado «*How Much Information? 2010 Report on Enterprise Server Information*», en el que se calculó que, en el año 2008, «los servidores del mundo procesaron 9,57 zettabytes de información, casi 10 elevado a la 22ª potencia, o diez millones de millones de gigabytes. Esto equivale a 12 gigabytes de información al día de un trabajador medio, o a unos 3 terabytes de información por trabajador al año. Las distintas empresas del mundo procesaron, de media, 63 terabytes de información al año».

Febrero 2010
Datos y más datos.
The Economist publicó el informe titulado *Data, Data Everywhere*. En él, su autor Kenneth Cukier escribe: «…el mundo contiene una cantidad de información digital de una magnitud inimaginable, cuyo ritmo de crecimiento es frenético… El efecto es patente en todos los ámbitos de nuestra vida, desde los negocios hasta la ciencia, los gobiernos o el arte».

Abril 2010
La aparición del ERP en la nube.
Netsuite y Lawson Software, entre otras empresas, fueron las primeras que adoptaron las tecnologías de nube para los sistemas ERP. Comenzaron ofreciendo a medianas empresas y organizaciones soluciones de sistemas ERP ligeros, flexibles y asequibles.
Fuente: ComputerWeekly, Putting ERP in the Cloud

Julio 2010
Coordinación del ERP con los procesos de negocio.
No todas las organizaciones que han invertido en sistemas ERP han logrado el éxito de sus iniciativas. Son numerosos los casos de implementaciones erróneas y, en

algún caso, su fracaso total. La implementación de ERP es un problema sociotécnico que necesita una perspectiva diferente a la de las innovaciones informáticas, depende profundamente de una perspectiva equilibrada de toda la organización. Entre los principales factores de éxito estratégicos se puede encontrar la coordinación de los procesos de negocio y de los procesos ERP integrados, que se encuentran bajo la influencia de la cultura de la organización. Fuente: The Challenge of Enterprise Systems: Harmonization of ERP Systems with Business Processes

Enero 2011
Tendencias de la Inteligencia empresarial (BI).
En 2011, las principales tendencias emergentes de Inteligencia empresarial fueron los servicios en la nube, la visualización de datos, el análisis predictivo y el Big Data. Fuente: EnterpriseManagement.com, Top 10 Trends in Business Intelligence and Analytics in 2011

2011
#IBMBigData
En 2011, IBM introdujo la etiqueta de Twitter, #IBMbigdata, que tenía como objetivo desarrollar el sitio web temático del Big Data que crearon en 2008 con intención de integrarlo en sus acciones de marketing.
Fuente: Twitter Search, #IBMBigData

Febrero 2011
El crecimiento real de los datos.
En un artículo titulado *The World's Technological Capacity to Store, Communicate, and Compute Information de Science Magazine*, se calculó que la capacidad mundial de almacenamiento de información creció a una tasa anual del 25% anual desde 1987 hasta 2007. En el mismo sentido, se afirmó que, en el año 1986, el 99,2% del almacenamiento de datos era analógico, pero en 2007 el 94% de dicho almacenamiento era digital. Esto supone un cambio radical en un periodo de tiempo de tan solo 20 años (en 2002, el almacenamiento digital superó al no digital por primera vez).

2012
Capacidad de la información.
En 2012, en el artículo *Tracking the Flow of Information into the Home del International Journal of Communication*, se calculó que el suministro de información por parte de los medios de comunicación a los hogares estadounidenses había pasado de ser de unos 50.000 minutos al día en el año 1960, a cerca de 900.000 en 2005. Igualmente, se calculó que los Estados Unidos «se estaban aproximando a los mil minutos de contenido a través de medios de comunicación disponibles por cada minuto disponible para su consumo».

Marzo 2012

Las dudas existenciales del Big Data.
El artículo *Critical Questions for Big Data*, publicado en Information, Communications, and Society Journal, define el Big Data como «un fenómeno cultural, tecnológico e intelectual que aparece por la interconexión de los siguientes elementos: (1) Tecnología: optimización de la capacidad informática y de la precisión de los algoritmos para recopilar, analizar, enlazar y comparar grandes conjuntos de datos. (2) Análisis: uso de grandes conjuntos de datos para identificar patrones con el fin de realizar afirmaciones económicas, sociales, técnicas y legales. (3) Mitología: la creencia popular de que los grandes conjuntos de datos ofrecen una forma superior de inteligencia y conocimientos que pueden generar datos que anteriormente no eran posibles, con un aura de verdad, objetividad y exactitud».

1 Noviembre 2014
El año del Internet de las cosas (IoT).
El IoT se ha convertido en una fuerza poderosa para la transformación de negocios, y su enorme impacto afectará en los próximos años a todos los sectores y todas las áreas de la sociedad. Existen enormes redes de objetos físicos dedicados (cosas) que incorporan tecnología para detectar o interactuar con su estado interno o medio externo. Según Gartner, había 3700 millones de "cosas" conectadas en uso en 2014 y esa cifra se elevará hasta los 4900 millones en 2015.
Fuente: Impact of IoT on Business at the Gartner Symposium/ITxpo 2014

2015
Smart Cities o ciudades inteligentes.
Una ciudad inteligente (*smart city*) hace uso del análisis de información contextual en tiempo real para mejorar la calidad y el rendimiento de los servicios urbanos, reduce costos, optimiza recursos e interactúa de forma activa con los ciudadanos. Según estimaciones de Gartner habrá más de 1100 millones de dispositivos conectados y en uso en diversas ciudades en 2015, incluyendo sistemas de iluminado LED inteligentes, de monitorización de salud, cerraduras inteligentes y numerosas redes de sensores para detección de movimiento, estudio de contaminación atmosférica, etc.
Fuente: Impact of IoT on Business at the Gartner Symposium/ITxpo 2014

2020
El futuro del Big Data.
La producción de datos aumenta a un ritmo espectacular. Los expertos apuntan a un aumento estimado del 4300% en la generación de datos anuales para 2020. Entre los principales motivos que llevan a este cambio se incluyen el cambio de tecnologías analógicas a digitales y el rápido aumento en la generación de datos, tanto por particulares como por grandes empresas.
Fuente: CSC.com, Big Data Just Beginning to Explode

Km 12+000 Carretera Estatal 431 “El colorado-Galindo”
Parque Tecnológico San Fandila
Mpio. Pedro Escobedo, Querétaro, México
CP 76703
Tel +52 (442) 216 9777 ext. 2610
Fax +52 (442) 216 9671

publicaciones@imt.mx

http://www.imt.mx/

Esta publicación fue desarrollada en el marco de un sistema de gestión de calidad certificada bajo la norma ISO 9001:2015