www.elementos.buap.mx

Internet, *big data* y ciencia de datos:
una triada transformadora

María Josefa **Somodevilla García**

Internet, *big data* y ciencia de datos forman una tripleta poderosa que está revolucionando a toda la sociedad. Su proceso de interacción permite la generación, procesamiento y análisis de grandes cantidades de datos para brindar una comprensión adicional y poder tomar decisiones informadas. En este artículo se tratarán aspectos importantes de cada componente de esta triada, como lo son sus competencias, sus áreas de aplicación, así como los desafíos que esta representa.

INTERNET: LA FUENTE INAGOTABLE DE DATOS

Internet es una red de computadoras que se encuentran interconectadas a nivel mundial para compartir información y, debido al incremento de los dispositivos conectados a la misma, facilita la generación masiva de datos. Esta impresionante actividad en línea genera diariamente una cantidad enorme de datos, la cual proviene principalmente de las redes sociales, el comercio electrónico y los dispositivos IoT (Internet de las Cosas).

Las redes sociales generan datos a través de publicaciones, comentarios y otras interacciones de sus usuarios. Por ejemplo, Rahul Shewale señala que cada minuto se realizan 6.3 millones de búsquedas en Google y se envían 41.6 millones de mensajes de WhatsApp (Shewale, 2024). Por otra parte, el comercio electrónico comprende transacciones

en línea, reseñas de productos y comportamientos de compra, y su análisis proporciona datos valiosos sobre las preferencias del consumidor. Los sistemas de recomendación son herramientas esenciales en el comercio electrónico para mejorar la experiencia del usuario y aumentar las ventas al sugerir productos relevantes. Estos sistemas emplean algoritmos que analizan el comportamiento del consumidor, los artículos comprados con anterioridad y otros datos para generar recomendaciones sugestivas, ejemplo en Amazon, eBay y Netflix. Finalmente, los dispositivos IoT suministran datos diariamente sobre las actividades realizadas en línea a través de dispositivos conectados, desde sensores de fábrica hasta artefactos inteligentes. El hecho de que el usuario también pueda completar transacciones en línea significa que la fuente de estos datos, conocidos como *big data,* sea altamente compleja, y aumenta cada día el potencial de Internet.

BIG DATA: LA GESTIÓN DE VOLÚMENES MASIVOS DE DATOS

Big data se refiere a conjuntos de datos extremadamente grandes y complejos que superan la capacidad de las herramientas de procesamiento de datos tradicionales, como los sistemas de análisis estadístico (SAS) o las bases de datos relacionales, por mencionar algunas. Para describir *big data* se hace referencia a sus características principales, conocidas como sus cinco V (Figura 1).

De acuerdo con Big data Analytics News, "en 2024 alcanzó 2.5 quintillones de bytes, a una velocidad de procesamiento promedio de alrededor de 45.6 Mbps" (Volumen y Velocidad), lo cual es fundamental para aplicaciones en tiempo real (BDAN, 2024). La calidad y precisión de los datos es crucial para obtener resultados confiables (Veracidad). *Big data* tiene una naturaleza heterogénea, ya que los datos, provienen de múltiples fuentes (Variedad). Este tipo de datos puede presentarse de manera estructurada, no estructurada o semiestructurada, y son multimodales (texto, audio, imagen, video). La V de Valor representa el objetivo final de *big data*: "se estima que el mercado está creciendo a un ritmo astronómico, listo para alcanzar un valor de $103 mil millones en 2027" (BDAN, 2024).

CIENCIA DE DATOS: LA EXTRACCIÓN DE CONOCIMIENTOS VALIOSOS

La ciencia de datos es una disciplina que se nutre de la integración y colaboración de diversas áreas del conocimiento para abordar problemas complejos y extraer valor de grandes volúmenes de datos generados por Internet y gestionados por *big data*. Por lo tanto, se pueden identificar tres dimensiones fundamentales que hacen de la ciencia de datos una ciencia interdisciplinaria en la que participan las matemáticas, la estadística, las ciencias de la computación, y la ética y legislación.

Las matemáticas y la estadística son necesarias para el análisis de datos. Por un lado, la estadística aporta los métodos para la inferencia, la probabilidad y la modelización de datos. Por otro lado, las matemáticas proporcionan las bases teóricas para entender y desarrollar algoritmos complejos, los cuales hacen posible el desarrollo de modelos de aprendizaje automático. Las ciencias de la computación abarcan una gran variedad de áreas de estudio que se centran en la teoría, el desarrollo y la aplicación de *software* y sistemas de computación para el procesamiento de grandes volúmenes de datos. Entre estas áreas destacan los algoritmos, el estudio de las estructuras de datos, la programación, la ingeniería de *software*, la inteligencia artificial, el aprendizaje automático y la visualización de datos. Los algoritmos, las estructuras de datos y la programación son esenciales para la implementación de modelos de ciencia de datos, ya que son imprescindibles para desarrollar sistemas escalables y eficientes para el análisis de datos masivos. Por su parte, las metodologías de la ingeniería de *software* son vitales para desarrollar aplicaciones y herramientas que soporten el análisis de datos a gran escala, como, por ejemplo, el diseño de sistemas para la gestión y la integración de datos, y la creación de interfaces de usuario.

La inteligencia artificial (IA) y el aprendizaje automático (ML) son campos interrelacionados de las

Figura 1. Las cinco V de *big data*

ciencias de la computación. Dichos campos se centran en crear sistemas capaces de realizar tareas que requieren de habilidades humanas, tales como razonamiento, aprendizaje, percepción, comprensión del lenguaje y toma de decisiones. Estos campos han revolucionado la tecnología y se aplican en asistentes virtuales y diagnósticos médicos, por ejemplo. Los sistemas inteligentes usan algoritmos que hacen predicciones y encuentran relaciones basadas en datos, y son capaces de identificar fraudes o ataques cibernéticos, entre otros. Los algoritmos de aprendizaje por refuerzo optimizan decisiones en entornos dinámicos, como los vehículos autónomos que navegan en escenarios complejos.Una etapa fundamental de la ciencia de datos es la comunicación de los hallazgos de manera efectiva, para lo cual se emplean herramientas de visualización de datos (Simplilearn, 2024). Dichas herramientas combinan las habilidades de diseño, estadística y programación, tales como gráficos de barras, de pastel, etc., facilitando la comprensión y la toma de decisiones basadas en datos. Con el uso creciente de datos personales y sensibles, la ética y la legislación son componentes críticos de la ciencia de datos. Para garantizar que el uso de datos se realice de manera responsable, los científicos de datos deben trabajar colaborativamente con expertos legales.

En México, el derecho a la protección de datos personales estaba regulado con diversas normativas por el Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI), el organismo constitucional autónomo responsable de asegurar el cumplimiento de estas leyes. En la actualidad, el INAI ha concluido un proceso de transición hacia la nueva Secretaría Anticorrupción y Buen Gobierno.

En el sector privado, la Ley Federal de Protección de Datos Personales en Posesión de los Particulares (LFPDPPP) establece las reglas y lineamientos (*Ley Federal de Protección de Datos Personales en Posesión de los Particulares*, s.f.); por otro lado, en el sector público, la Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados (LGPDPPSO) se encarga de esta regulación (*Ley General de Protección de Datos Personales En Posesión de Sujetos Obligados*, s.f.). A pesar de la transición, las leyes mencionadas se encuentran aún vigentes

APLICACIONES PRÁCTICAS

La integración de Internet, *big data* y ciencia de datos ha permitido desarrollar una amplia gama de aplicaciones prácticas que están transformando diversos sectores para lo cual se requiere un conocimiento profundo del dominio específico en el que se aplica (Rice, 2024). Los expertos en el dominio colaboran con científicos de datos para asegurar que los modelos y análisis sean relevantes y aplicables al contexto específico, proporcionando interpretaciones precisas de los resultados que conllevan a acciones concretas.

Entre las aplicaciones más relevantes podemos mencionar que, en el área de salud y medicina, la telemedicina y los registros electrónicos de salud generan datos que facilitan el diagnóstico y tratamiento de enfermedades. En el campo de la educación, las plataformas de aprendizaje en línea recopilan datos sobre el comportamiento de los estudiantes, permitiendo mejorar los métodos de enseñanza y personalizar los procesos de enseñanza-aprendizaje al considerar sus necesidades.

© **Rafael Pareja**. De la serie *Nuestras manos*, 2012.

© **Rafael Pareja**. De la serie *Nuestras manos*, 2012.

Para las empresas de mercadeo digital, es posible analizar el comportamiento del consumidor en línea y personalizar campañas de publicidad, aumentando la rentabilidad. Al mismo tiempo, el procesamiento de grandes cantidades de datos puede identificar y mitigar amenazas de seguridad cibernética en tiempo real.

Otras aplicaciones importantes son las del transporte y la logística, especialmente en la optimización de rutas y la gestión de flotas. En tiempo real se utilizan datos de tráfico y condiciones de la carretera para reducir la duración de viajes; esto disminuye los costos de operación, aumenta la eficiencia y reduce las emisiones de carbono.

La interacción entre Internet, *big data* y ciencia de datos tiene un gran impacto en el desarrollo y funcionamiento de modelos de IA, porque Internet proporciona los datos, *big data* maneja su almacenamiento y procesamiento y la ciencia de datos desarrolla y perfecciona los modelos. Algunos los aspectos clave de esta interacción pueden verse de la siguiente manera: 1) la mejora de la capacidad de respuesta a las consultas del usuario, en el sentido de hacerlas más precisas y contextualmente relevantes; 2) la escalabilidad y eficiencia, que permite atender a un gran número de usuarios simultáneamente sin perder calidad en las respuestas; 3) la innovación continua, que enriquece las capacidades del modelo para adaptarse a nuevas aplicaciones y usos. Al mismo tiempo, los elementos de esta tripleta hacen posible que los modelos de lenguaje grande (LLM), como ChatGPT, Deepseek y Gemini, entre otros, se conviertan en una herramienta poderosa y útil en una amplia gama de aplicaciones, desde la asistencia virtual hasta la generación automática de texto.

COMENTARIOS FINALES

La convergencia entre Internet, *big data* y ciencia de datos está revolucionando la forma en que los datos son generados, procesados y aplicados en los ámbitos público, privado y social. A medida que las tecnologías continúan avanzando, esta triada seguirá evolucionando, adoptando enfoques innovadores para enfrentar retos, dentro de los cuales destacan: la gestión de datos no estructurados, la privacidad y seguridad, y la gestión de la infraestructura y los recursos. La mayoría de los datos generados son no estructurados y pueden ser de cualquier naturaleza (imágenes, audio, datos de sensores, datos de texto, etc.).

Los datos no estructurados no están predefinidos mediante modelos de datos, como las bases de datos relacionales, por ejemplo, lo que presenta un desafío significativo para su almacenamiento y análisis. Es muy importante considerar, que el aumento en los ciberataques demanda un reforzamiento en las políticas de seguridad de los datos y, finalmente, entender que procesar grandes volúmenes de datos requiere de una infraestructura robusta y de recursos computacionales y ecológicos significativos, lo que puede ser costoso para muchas organizaciones y sobre todo para el medio ambiente.

© Rafael Pareja. De la serie *Nuestras manos*, 2012.

El impacto de los LLM en la conservación de los recursos naturales, específicamente, el impacto de la huella de carbono (Dhar, 2020) y agua (Li et al., 2023), es significativo y va en ascenso, lo cual plantea retos en términos de sostenibilidad energética, a medida que los modelos de IA se tornan más complejos. Las investigaciones actuales (Sarkar *et al.*, 2024) se encaminan a racionalizar los centros de datos y trabajar en la reducción de la huella de carbono, generada por la utilización de estos modelos mediante el uso de fuentes de energía renovable y tecnologías más eficientes (Li *et al*., 2023).

Por lo tanto, el verdadero desafío será equilibrar la innovación con la responsabilidad, garantizando que estas tecnologías se utilicen de manera que beneficien a toda la sociedad, al mismo tiempo que salvaguarden nuestro planeta.

REFERENCIAS

BDAN (2024). Big data Analytics News. 50+ incredible Big data statistics for 2024: Facts, market size & industry growth. *Big Data Analytics News*. Recuperado de: https://bigdataanalyticsnews.com/big-data-statistics/.

Dhar P (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence 2*(8):423-425. DOI: https://doi.org/10.1038/s42256-020-0219-9.

Ley Federal de Protección de Datos Personales en Posesión de los Particulares. (s. f.). Cámara de Diputados. Recuperado el 18 de febrero de 2025, de: https://www.diputados.gob.mx/LeyesBiblio/ref/lfpdppp.htm.

Ley General de Protección de Datos Personales en posesión de Sujetos Obligados. (s. f.). Cámara de Diputados. Recuperado el 18 de febrero de 2025, de: https://www.diputados.gob.mx/LeyesBiblio/ref/lgpdppso.htm.

Li P Yang J, Islam MA and Ren S (2023). Making AI Less «Thirsty»: Uncovering and Addressing the Secret Water Footprint of AI Models. *arXiv.* Cornell University. DOI: https://doi.org/10.48550/arxiv.2304.03271.

Rice M (2024). 30 Data Science Applications and Examples. Recuperado de: https://builtin.com/data-science/data-science-applications-examples.

Shewale R (2024). Big data Statistics for 2024 (growth, market size & more). Recuperado de: https://www.demandsage.com/big-data-statistics/.

Simplilearn (2024). 23 Best Data Visualization Tools You Can't Miss! Recuperado de: https://www.simplilearn.com/data-visualization-tools-article.

Sarkar S, Naug A, Luna R *et al*. (2024). Carbon Footprint Reduction for Sustainable Data Centers in Real-Time. *arXiv*:2403:14092. DOI: https://doi.org/10.48550/arXiv.2403.14092.

María Josefa Somodevilla García
Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla
maria.somodevilla@correo.buap.mx

© **Rafael Pareja**. De la serie *Nuestras manos*, 2012.