

# Sistemas de Big Data

## Fundamentos de Big Data

### Clase 1

MAESTRÍA EN  
SISTEMAS DE INFORMACIÓN  
Mención Data Science

La excelencia no se improvisa



# 1. INTRODUCCIÓN DE LA CLASE

**Big Data** hace referencia al conjunto de datos cuyo **volumen, velocidad y variedad** superan la capacidad de los sistemas tradicionales de gestión de datos. Este fenómeno se ha convertido en un elemento clave en la **transformación digital** de las empresas y organizaciones. La definición de Big Data está ligada a las tres V: volumen, que describe la gran cantidad de datos generados; velocidad, que refiere al ritmo con el que estos datos son producidos y procesados; y variedad, que señala la **diversidad de tipos** de datos provenientes de **diferentes fuentes** (LaValle et al., 2011). Esta expansión de datos presenta grandes retos, pero también oportunidades para la innovación, ya que ofrece a las organizaciones la capacidad de extraer conocimiento útil a partir de grandes **volúmenes de información**. Además, la evolución de las tecnologías como la computación en la nube, el procesamiento paralelo y los algoritmos de análisis avanzados han permitido que el uso de **Big Data** se convierta en una práctica común en muchos sectores.

En cuanto a sus aplicaciones, **Big Data** no solo permite la toma de decisiones más informadas y precisas, sino que también posibilita la personalización de productos y servicios, la **optimización** de procesos y la mejora en la atención al cliente. En diversas industrias, como la salud, el **retail**, las finanzas y el transporte, Big Data ha demostrado su potencial al ofrecer soluciones innovadoras que mejoran la eficiencia y efectividad operativa (Manyika et al., 2011). El análisis de **grandes volúmenes** de datos permite a las empresas obtener *insights* en tiempo real, lo que resulta crucial para anticipar tendencias de mercado, mejorar la experiencia del usuario y desarrollar nuevas oportunidades de negocio. Sin lugar a dudas, el impacto de **Big Data** en la toma de decisiones empresariales es innegable, ya que transforma la manera en que las organizaciones gestionan la información y la aplican para lograr una **ventaja competitiva**.

**RDA 1:** Utilizar tecnologías y herramientas de Big Data para grandes volúmenes de datos.

## Clase 1: Fundamentos de Big Data Introducción a Big Data, definición y características

**Big Data** se refiere a **conjuntos de datos extremadamente grandes y complejos**, que requieren herramientas avanzadas de **procesamiento y análisis** para extraer valor. Tradicionalmente, los datos eran manejados utilizando **bases de datos relacionales** y técnicas **convencionales de análisis**, pero con la explosión de datos generados por dispositivos digitales, redes sociales, sensores y aplicaciones empresariales, la necesidad de nuevas tecnologías y métodos de **procesamiento** se hizo evidente. **Big Data** se caracteriza por las **tres V**: volumen, que se refiere a la cantidad masiva de datos generados a diario; velocidad, que destaca la rapidez con la que estos datos deben ser procesados; y variedad, que hace referencia a los diversos tipos de datos (estructurados, no estructurados y semiestructurados), que se generan en **diversas fuentes** (Laney, 2001). Estos desafíos han llevado al desarrollo de herramientas y plataformas especializadas como **Hadoop**, Spark y bases de datos **NoSQL**, que permiten almacenar y procesar grandes volúmenes de datos de manera eficiente.

La era del Big Data está marcada por dos grandes **preceptos**; por un lado, el gran caudal de datos y las altas capacidades de **cómputo** que se tiene en la actualidad, sumado a que este procesamiento se puede hacer en **cualquier** momento y en cualquier **lugar**.

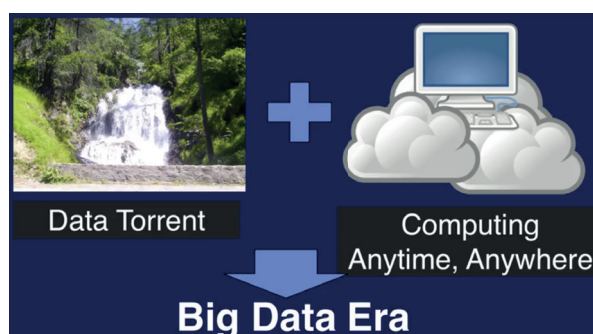


Figura N.º 1. Big Data Era

El impacto de Big Data es profundo y se extiende a diversas **industrias** y **sectores**. En el ámbito **empresarial**, permite la toma de decisiones más informadas y la personalización de productos y servicios de manera que se adapten mejor a las necesidades del consumidor. Además, en sectores como la salud, la educación y la **manufactura**, Big Data facilita la mejora de procesos operativos, la optimización de recursos y la predicción de tendencias. Las tecnologías de análisis de **Big Data** permiten obtener información relevante de grandes volúmenes de datos que, a menudo, no son accesibles a través de métodos tradicionales. Mediante técnicas de análisis –como minería de datos, **machine learning** y **análisis predictivo**– las organizaciones pueden detectar patrones ocultos, identificar oportunidades de negocio y mejorar la eficiencia en la toma de decisiones estratégicas (Mayer-Schönberger & Cukier, 2013).

## Describir las características y componentes claves de Big Data

Big Data se distingue por sus características fundamentales que permiten diferenciarlo de los datos tradicionales. Estas características son conocidas como las **tres V (volumen, velocidad y variedad)** y cada una de ellas resalta un desafío particular, que las organizaciones deben superar para **gestionar** y aprovechar **adecuadamente** los datos a gran escala. El **volumen** hace referencia a la enorme cantidad de datos que se generan y almacenan a diario, **alcanzando niveles** que superan la **capacidad** de los sistemas tradicionales de bases de datos. Por otro lado, la **velocidad** describe la rapidez con la que estos datos deben ser procesados, a menudo en tiempo real o casi en tiempo real, para que sean útiles. Finalmente, la **variedad** se refiere a los diferentes tipos de datos que existen (estructurados, semiestructurados y no estructurados), provenientes de **diversas fuentes** como redes sociales, dispositivos IoT, **transacciones comerciales** y más (Laney, 2001). Además de las **tres V**, algunos autores incluyen otras características, como la **veracidad** (calidad y confiabilidad de los datos) y el **valor** (la capacidad de extraer información útil de los datos recopilados), lo que da forma al concepto completo de **Big Data**.

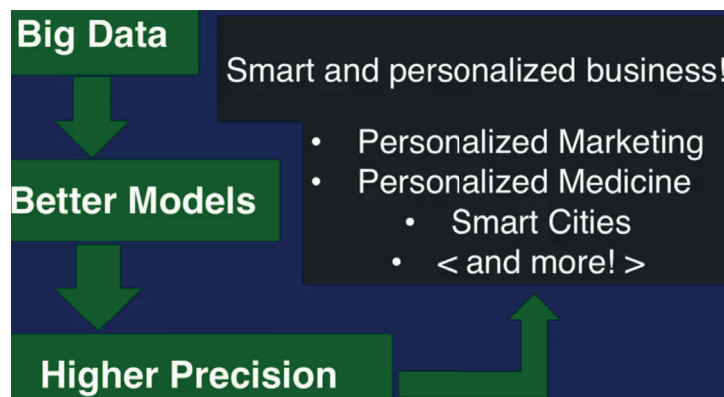


Figura N.º 2. Aplicaciones del Big Data

Los componentes clave del ecosistema de **Big Data** son esenciales para abordar los desafíos mencionados y para poder extraer valor de los datos. En primer lugar, el **almacenamiento** de **Big Data** se logra mediante sistemas distribuidos como el Hadoop Distributed File System (**HDFS**), que permite almacenar grandes volúmenes de datos de manera eficiente y escalable. A su vez, se utilizan bases de datos **NoSQL** como **MongoDB** y **Cassandra**, para gestionar datos **no estructurados o semiestructurados**. En segundo lugar, el **procesamiento** de los datos se realiza utilizando tecnologías como **Hadoop** y **Apache Spark**, que facilitan el procesamiento paralelo y distribuido de datos. **Hadoop** utiliza el modelo MapReduce, para dividir tareas complejas en fragmentos más pequeños; mientras que **Spark**, al trabajar en memoria, mejora la velocidad del procesamiento (Zaharia et al., 2010). Finalmente, el **análisis** de Big Data se centra en identificar patrones y generar predicciones a través de técnicas como el aprendizaje automático, la minería de datos y el análisis predictivo. Estas herramientas, más el **visualizado** de datos a través de plataformas como **Tableau o Power BI**, permiten a los usuarios interpretar los resultados de forma efectiva y tomar decisiones informadas (Few, 2006). El conjunto de estos componentes,

desde el almacenamiento hasta la visualización, habilita a las organizaciones para transformar grandes volúmenes de datos en información valiosa y procesable.

## Aplicaciones de Big Data

### Marketing personalizado y sistemas de recomendación

En la era digital, el **marketing personalizado** ha emergido como una estrategia clave para las empresas que buscan optimizar su relación con los clientes. A través del análisis de datos, las empresas pueden segmentar a sus clientes y ofrecerles productos o servicios adaptados a sus **preferencias y comportamientos específicos**. Esto se logra mediante el uso de **sistemas de recomendación**, que son algoritmos diseñados para prever las preferencias de los consumidores basándose en datos históricos, actividades previas y patrones de compra. Estos sistemas, comúnmente utilizados por plataformas como **Amazon** y **Netflix**, ayudan a los usuarios a descubrir productos o contenidos que podrían interesarles, mejorando la experiencia del cliente y, por ende, las tasas de conversión (Ricci et al., 2015). Los sistemas de recomendación pueden ser colaborativos, basados en el comportamiento de otros usuarios similares, o basados en contenido, recomendando productos similares a aquellos que el cliente ha mostrado interés previamente.



Figura N.º 3. Aplicaciones del Big Data. Sistemas de recomendación.

Además, el marketing personalizado también se extiende a través de la **publicidad móvil**, donde los datos recolectados de dispositivos móviles permiten a las empresas segmentar de manera más precisa a sus **audiencias**. Con el uso de tecnologías de **geocalización** y datos de comportamiento en tiempo real, las marcas pueden adaptar sus campañas publicitarias para que sean más relevantes y eficaces, optimizando así los recursos invertidos en publicidad. Este tipo de marketing dirigido ha **demostrado** ser más efectivo en comparación con los **métodos tradicionales**, ya que permite una conexión más cercana y personalizada con el consumidor, aumentando la satisfacción y la lealtad del cliente (Shankar et al., 2017).



"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Figura N.º 4. Aplicaciones del Big Data. Perfiles de usuario.

En la Figura N.º4 se aprecia un ejemplo 'algo exagerado' del uso y aplicación del Big Data en el **scoring** y determinación del **perfil** de un **cliente** en función de su consumo.

### Análisis de sentimientos y crecimiento del cliente

El análisis de sentimientos se refiere al uso de técnicas de **procesamiento de lenguaje natural** (PLN) para extraer, analizar y comprender las opiniones y emociones de los usuarios expresadas a través de **redes sociales, reseñas, comentarios** y otros puntos de contacto digitales. Este tipo de análisis permite a las empresas conocer cómo se sienten los clientes respecto de productos o servicios, esto ofrece información valiosa para mejorar la **experiencia del cliente** y la toma de decisiones empresariales (Pang & Lee, 2008). La combinación del análisis de sentimientos con el **crecimiento del cliente** es una estrategia poderosa para identificar oportunidades de mejora en productos y servicios, así como para predecir **comportamientos futuros**, como la probabilidad de que un cliente continúe comprando o abandone la marca. Esto se traduce en la creación de **estrategias de retención personalizadas** y de desarrollo de productos que se alinean con las expectativas y necesidades de los consumidores.



Figura N.º 5. Aplicaciones del Big Data. Análisis de sentimientos.

### Captura del comportamiento del cliente y agendamiento de vuelos



Las aplicaciones biomédicas, impulsadas por el análisis de Big Data, están revolucionando el campo de la salud. Con el uso de algoritmos avanzados y la recopilación de grandes volúmenes de datos médicos, se pueden identificar patrones y predicciones sobre enfermedades y tratamientos. En el caso del **tratamiento personalizado de cáncer**, el análisis de datos genómicos y clínicos permite diseñar tratamientos específicos para cada paciente, aumentando las probabilidades de éxito y reduciendo efectos secundarios. Este enfoque personalizado en oncología mejora la calidad de vida del paciente y optimiza los recursos médicos, marcando un avance significativo hacia una medicina más precisa y eficiente (Topol, 2019).

## Publicidad en teléfonos celulares: estrategias e impacto

La **publicidad en teléfonos celulares** se ha convertido en un componente esencial de las estrategias de **marketing** digital modernas, debido al auge del uso de dispositivos móviles. Con el crecimiento del acceso a internet a través de **smartphones**, las **empresas** tienen la oportunidad de llegar a sus audiencias en tiempo real y en cualquier lugar. La publicidad **móvil** no solo permite la segmentación avanzada de los **consumidores**, sino que también ofrece una variedad de formatos, como anuncios en **aplicaciones**, **notificaciones push**, anuncios en **redes sociales** y **búsquedas** móviles, entre otros. Esta diversificación de formatos permite a las marcas crear campañas más dinámicas y personalizadas que maximizan la interacción del usuario y mejoran las tasas de **conversión** (Shankar et al., 2017). La capacidad de dirigir anuncios específicos basados en datos de comportamiento del usuario, ubicación **geográfica** y preferencias **individuales** ha revolucionado el marketing, haciendo que los anuncios sean mucho más relevantes y menos intrusivos, lo que genera una mejor experiencia para el **consumidor**.

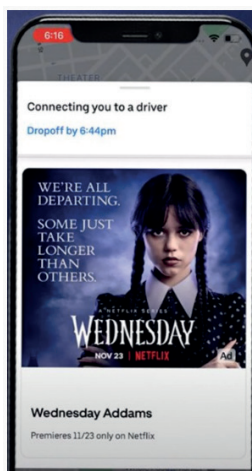


Figura N.º 8. Publicidad de teléfonos celulares.

Además, la publicidad en teléfonos celulares se ha visto impulsada por tecnologías como el **marketing basado en la localización (location-based marketing)**, que utiliza los datos de geolocalización para mostrar anuncios en tiempo real a los consumidores en función de su ubicación física. Esto es especialmente útil para negocios locales que buscan atraer a clientes cercanos. Según un estudio de Google (2017), más del 50% de las búsquedas móviles tienen una intención local, lo que subraya la importancia de los anuncios dirigidos por ubicación. Esta capacidad de personalizar los anuncios no solo aumenta la probabilidad de conversión, sino que también mejora la relevancia de la marca en la vida del consumidor; lo que genera una mayor lealtad a largo plazo. A su vez, el análisis de los datos generados a través de estas interacciones móviles permite a las empresas medir el rendimiento de sus campañas en tiempo



rectamente, pueden proporcionar una visión más precisa y completa de las **dinámicas del mercado**, el comportamiento del consumidor y la eficiencia operativa. Las empresas que integran Big Data en sus **procesos de toma de decisiones** pueden anticiparse a las tendencias del mercado, mejorar la personalización de productos y servicios, e identificar oportunidades para optimizar operaciones. El análisis de grandes volúmenes de datos provenientes de **diversas fuentes** –como transacciones de clientes, redes sociales y dispositivos de **IoT**– ofrece una visión más profunda y dinámica de los procesos empresariales (McKinsey & Company, 2011). Esto no solo permite una mayor **agilidad y precisión** en las decisiones estratégicas, sino también la capacidad de adaptar las operaciones de la empresa en tiempo real para mantenerse competitiva.



Figura N.º 10. Aplicaciones del Big Data.

El impacto del Big Data en las decisiones **empresariales** se puede observar en diversas áreas funcionales dentro de una **organización**. Por ejemplo, en el área de marketing, las empresas pueden usar Big Data para desarrollar campañas publicitarias más efectivas, mediante el análisis de datos de clientes, comportamientos y **preferencias** individuales. Esto permite la **segmentación avanzada** y el **marketing personalizado**, lo que aumenta la eficacia de las campañas y mejora el retorno de inversión (ROI). En términos de operaciones, Big Data permite la optimización de la cadena de **suministro**, el análisis predictivo para la gestión de inventarios y la mejora en la logística mediante el análisis de datos en tiempo real. En el **ámbito** de la toma de **decisiones financieras**, los análisis de Big Data pueden ayudar a predecir fluctuaciones del mercado, identificar riesgos potenciales y tomar decisiones de inversión más **informadas** (Chui et al., 2012). De esta manera, Big Data no solo facilita la toma de decisiones más informadas, sino que transforma la forma en que las empresas operan y se **posicionan** en un entorno competitivo.

**Big Data, la nueva frontera para competencia, innovación y productividad** [<https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation>]

En las próximas clases se hablará mucho de **Hadoop**, por ello se añade una guía práctica de cómo leer su **documentación**.

**Accede** al sitio oficial de Hadoop. Visita la documentación oficial de **Hadoop**, donde encontrarás guías sobre su instalación, configuración y uso.

**Elige** la **versión adecuada** de Hadoop. Asegúrate de seleccionar la versión correspondiente a tu instalación, ya que las características pueden variar entre versiones.

**Comprende** la estructura de la **documentación**:

- **Guías** de inicio rápido. Útiles para entender los conceptos básicos y comenzar rápidamente con Hadoop.
- **Referencia** de la **API**. Si necesitas detalles técnicos sobre las clases y métodos de Hadoop.

- **Tutoriales** y ejemplos. Recomendados para aprender casos de uso comunes.
- **Documentación de configuración.** Te ayudará a configurar los componentes de Hadoop según tus necesidades.

Lee conceptos clave: familiarízate con conceptos básicos como **HDFS** (sistema de archivos distribuidos), MapReduce (modelo de procesamiento) y **YARN** (gestor de recursos).

Busca soluciones a problemas comunes: utiliza las secciones de solución de problemas y preguntas frecuentes (**FAQ**) para resolver dudas comunes.

Si tienes una pregunta específica o una tarea que quieras realizar con Hadoop, también puedes buscar en foros o en la comunidad de Apache para obtener ejemplos y soluciones más detalladas.

## Referencias citadas en la Clase 1

- Chui, M., Manyika, J., & Miremadi, M. (2012). *The impact of big data on business decisions*. McKinsey Quarterly.
- Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety*. META Group Research Note, 6(70), pp. 1-4.
- LaValle, S., Lesser, E., Shapiro, D., & Scott, A. (2011). *Big data, analytics and the path from insights to value*. MIT Sloan Management Review, 52(2), pp. 21-30.
- Manyika, J., Chui, M., Brown, B., & Bughin, J. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey & Company.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
- McKinsey & Company. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.
- White, T. (2012). *Hadoop: The definitive guide*. O'Reilly Media.
- Zhang, X., Zhao, K., & Chen, C. (2014). *Geo-marketing in mobile commerce: Applications, opportunities, and challenges*. Journal of Electronic Commerce Research, 15(3), pp. 238-250.

## Definición de los términos citados en la Clase 1

**Hadoop.** Es un marco de trabajo de código abierto diseñado para almacenar y procesar grandes volúmenes de datos distribuidos eficientemente. Su arquitectura está basada en el modelo de **procesamiento en paralelo**, usa un sistema de archivos distribuidos, conocido como **Hadoop Distributed File System (HDFS)**, que permite almacenar datos en múltiples nodos de una red, lo cual asegura **escalabilidad** y **resiliencia** frente a **fallos**. Además, Hadoop emplea el modelo de programación **MapReduce** para procesar **grandes cantidades de datos de manera distribuida**, dividiendo las tareas en pequeñas subtareas, que se ejecutan simultáneamente en varios **nodos** del **clúster**, de ese modo optimiza el tiempo de procesamiento. Hadoop es ampliamente utilizado en entornos de Big Data debido a su capacidad de manejar datos **no estructurados**, a su flexibilidad para trabajar con diversas fuentes de datos y a su capacidad de escalar de manera económica en función de las necesidades del negocio (White, 2012).

**Marketing usando geolocalización.** Es una estrategia que utiliza la ubicación geográfica de un usuario, obtenida a través de dispositivos móviles o tecnologías de ubicación, para enviarle mensajes de marketing personalizados en tiempo real. Este tipo de marketing se aprovecha de la capacidad de los smartphones para rastrear la posición del usuario por GPS, wifi o redes celulares. Gracias a esta información, las empresas pueden enviar promociones, descuentos, recomendaciones o alertas espe-





La excelencia no se improvisa

síguenos

