
SKILLING
CENTER
TECMILENIO

SKILLING CENTER TECMILENIO

Fundamentos del Big Data

Introducción al Big Data

Introducción

Hoy en día, la cantidad de información de diversas fuentes, contenidos y formatos está experimentando un crecimiento exponencial. Por ejemplo, los negocios basados en transacciones, como los bancos, las tiendas en e-commerce, los servicios de salud generan y analizan patrones de comportamiento para la toma de decisiones. Incluso las grandes corporaciones internacionales están muy ocupadas en mejorar las tecnologías de la información para lograr mayor alcance con menor costo, además de agilizar y automatizar este proceso.

Considera el caso de Jerónimo, que es el responsable de la obtención y organización del contenido de un noticiero por televisión. Aunque tiene bien definidos el perfil de la empresa y convenios con agencias de noticias internacionales, siempre tiene una gran variedad de información que seleccionar para su programación. ¿Qué herramienta necesita Jerónimo para hacer más eficiente su trabajo y cuál sería la ruta recomendable para lograr ese objetivo?

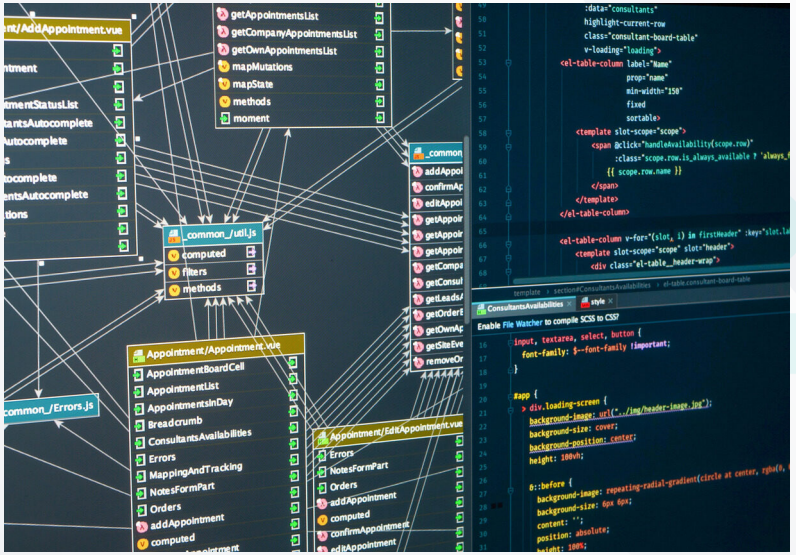
Explicación

Big Data es una herramienta de trabajo para empoderar la toma de decisiones con base en grandes volúmenes de información, en muchos casos, en tiempo real. La consultora en tecnología Gartner, Inc. (s.f.) define al Big Data como “activos de información de gran volumen, alta velocidad y/o alta variedad que exigen formas innovadoras y rentables de procesamiento que permiten una mejor comprensión, toma de decisiones y automatización de procesos”.

El campo de las finanzas es un gran usuario del Big Data, por ejemplo, para decisiones de otorgamiento de créditos, para analizar las transacciones diarias de los clientes de un banco, y detectar transacciones fraudulentas, por mencionar algunos.

Explicación

La generación de información se realiza en múltiples fuentes y contenidos. La disposición de esta información contempla grandes cantidades, diversidad de formatos y crecimiento acelerado. Dasgupta (2018) señala que la Biblioteca del Congreso de los Estados Unidos, es la más grande del mundo con 164 millones de artículos en su colección.



Explicación

Big Data requiere de herramientas tecnológicas de alta capacidad como procesadores y almacenamiento de memoria para facilitar el proceso de disponerla e interpretarla, lo cual se convierte en conocimiento. Adicionalmente se trata de automatizar este proceso.

Explicación

Tipos de datos según su estructura:

Datos estructurados

- Tablas con filas y columnas que combinan datos. Por ejemplo, el comportamiento del precio de una acción en el mercado de valores, las filas son periodos de tiempo como días y las columnas son variables, como el precio máximo, precio mínimo, precio de cierre y el volumen operado para cada fecha.

Datos semiestructurados

- No tienen estructura fija, pueden ser presentados a través de un esquema. Por ejemplo, una factura.

Datos no estructurados

- No tienen esquema organizacional ni estructura fija. Por ejemplo, noticias financieras en periódicos digitales, contenido de una página web, aplicaciones en dispositivos digitales, etcétera.

Explicación

Escala de las dimensiones de la información:

Medida de almacenamiento	Tamaño (decimal)	Ejemplo
Bit	Un dígito binario	1 o 0.
Byte	8 bits	Una letra.
Kilobyte (KB)	1,000 bytes	Un párrafo corto.
Megabyte (MB)	1,000 kilobytes	Una novela corta.
Gigabyte (GB)	1,000 megabytes	7 minutos de video en HD.
Terabyte (TB)	1,000 gigabytes	Un disco duro comercial.
Petabyte (PB)	1,000 terabytes	2,000 años de archivos MP3.
Exabyte (EB)	1,000 petabytes	Todo el contenido de Netflix visto 3 mil veces.
Zettabyte (ZB)	1,000 exabytes	250 mil millones de DVD.

Fuente: UC Berkeley School of Information. (2013). *Data Size Matters [Infographic].* Recuperado de https://ischoolonline.berkeley.edu/blog/big-data-infographic/

Con los avances en las tecnologías de la información se incrementa la capacidad de procesamiento, logrando mayor alcance en información y menor costo de procesamiento.

Actividad

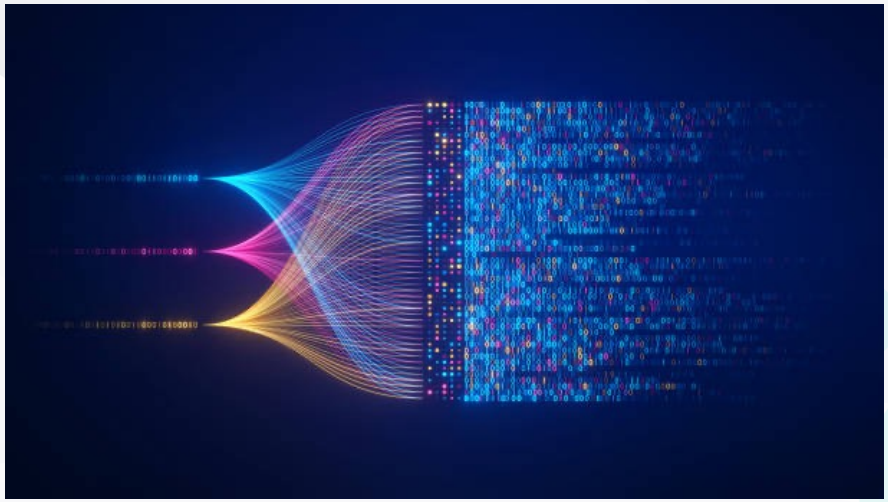
Para reforzar tu comprensión del tema, responde las siguientes preguntas:

1. Investiga y menciona tres campos de acción en donde se utiliza el Big Data y sus beneficios.
2. Describe brevemente la aplicación del Big Data en las finanzas de una empresa global dedicada a una actividad específica (la que tú selecciones).
3. Menciona un ejemplo real de cada tipo de datos según su estructura, y menciona sus características.

Cierre

Con esta introducción a los conceptos básicos del Big Data, estás entrando a un mundo de herramientas tecnológicas en el que siempre podrás encontrar aplicación para múltiples campos de acción. Ahora tienes conocimiento de las dimensiones y características de la información que requiere de herramientas poderosas para aprovecharla, dado que en todo momento está disponible en diversas fuentes, pero para la mente humana, es imposible procesar tal cantidad de información.

Con el conocimiento de este tema, mediante la ciencia del Big Data, Jerónimo cuenta con la herramienta adecuada para seleccionar y procesar los contenidos para su programación del noticiero televisivo.



Bibliografía

- Dasgupta, N. (2018). *Practical Big Data Analytics : Hands-on Techniques to Implement Enterprise Analytics and Machine Learning Using Hadoop, Spark, NoSQL and R.* Reino Unido: Packt Publishing.
- Gartner, Inc. (s.f.). *Big Data.* Recuperado de https://www.gartner.com/en/information-technology/glossary/big-data
- UC Berkeley School of Information. (2013). *Data Size Matters Infographic.* Recuperado de https://ischoolonline.berkeley.edu/blog/big-data-infographic/

SKILLING CENTER TECMILENIO

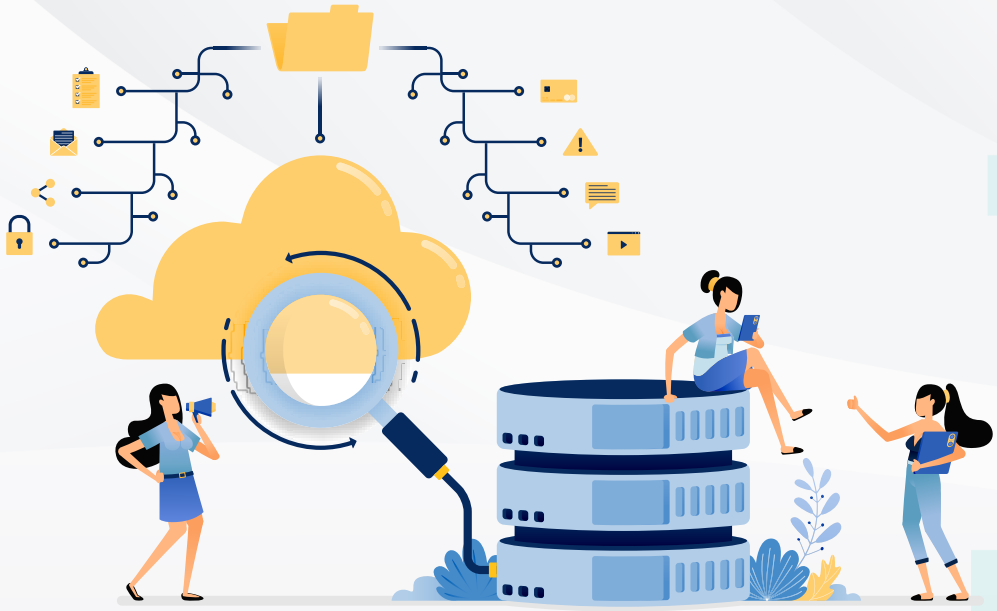
Fundamentos del Big Data

Orígenes de los grandes volúmenes de datos

Introducción

Juan es analista de un corporativo financiero y recientemente aprendió sobre cómo analizar Big Data para generar valor en el negocio. Para empezar a trabajar en dichos análisis, Juan necesita tener primero el contexto de los datos con los que podría trabajar y cómo es que estos son generados. Conocer el proceso no implica que vaya a encargarse de la recolección de los datos ni de su extracción, pero le permite ganar confianza en la calidad y comprensión de los datos con los que trabajará.

En el presente tema acompañarás a Juan en el entendimiento de cómo se generan los datos. Lo anterior te permitirá conocer las maneras en las que puedes obtener o generar nuevos datos para la empresa o industria en la que te desempeñas.



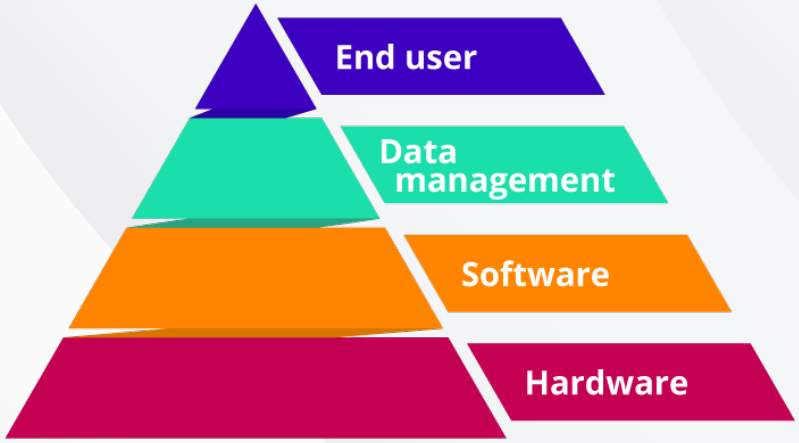
Explicación

La información ha tenido un crecimiento exponencial debido a la digitalización de los sistemas y procesos de diversos tipos, como entretenimiento, educación, turismo, finanzas, seguridad, etc. Este comportamiento ha sido soportado por la capacidad de procesamiento y almacenamiento de información y el desarrollo de dispositivos inteligentes.

Por ejemplo, el conjunto de transacciones de los clientes de un banco tiene enormes dimensiones, pero es de mayor valor cuando se cruza con datos sociodemográficos y el comportamiento histórico de los clientes. Con este análisis, los directivos identifican segmentos de clientes y sus respectivas estrategias de servicios para incrementar la competitividad del banco.

Explicación

El ciclo del Big Data, según Dasgupta (2018), se compone de cuatro bloques o etapas: *hardware*, *software*, *data management* y *end user*.



Fuente: Dasgupta, N. (2018). *Practical Big Data Analytics: Hands-on Techniques to Implement Enterprise Analytics and Machine Learning Using Hadoop, Spark, NoSQL, and R*. Packt Publishing.

1. Hardware: equipo computacional para procesamiento, almacenamiento y conectividad.
2. Software: algoritmos en lenguajes de programación para la extracción, transformación y análisis de la información.
3. Data management: se refiere a la configuración del almacenamiento y disposición de la información, incluye la gobernanza, la encriptación, clasificación y administración de la información.
4. End user: el proceso concluye con las necesidades o requerimientos del usuario que utiliza la información para tomar decisiones.

Explicación

Participantes en el ecosistema de la información:

- *Data devices*: al ser utilizados, generan información, como teléfonos inteligentes, sensores, puntos de venta.
- *Data collectors*: recolectan datos de los data devices, como el gobierno, bancos, proveedores de internet, líneas de producción, hospitales.
- *Data aggregators*: enriquecen la información de los data collectors, como *brokers* de información, marketing digital, información estadística del INEGI.
- *Data buyers and users*: consumidores de información, calificadoras de riesgo crediticio, intermediarios financieros .

Los participantes pueden estar involucrados en uno o más componentes del ecosistema.

Datos en la web:

- Redes sociales: plataformas de datos para interactuar profesional y socialmente.
- Medios de comunicación: noticieros, servicios de videollamadas (Zoom, Teams, Google Chat, plataformas de música (Spotify, iTunes), de *streaming* (Netflix, Disney).
- Comercio electrónico: *e-commerce* (Mercado Libre, Amazon).

Recolección de datos:

- *Log files*, que monitorea los movimientos en sitios web, proporcionando información de uso y preferencias.
- Motores de búsqueda, como Google, combinan métodos como *web crawlers* que segmentan palabras e índices para mapear la internet y ofrecer resultados coherentes (Pehcevski, 2019).
- Obtención de información en sitios web, conocido como *web scraping*, utilizando algoritmos de programación como Python.

Explicación

Contexto de las fuentes de información en los negocios:

- Sistemas de información para funciones específicas, como el sistema contable, nómina, ventas, logística, finanzas, plataformas integradas como el *Enterprise Resource Planning* (ERP), gestión de relaciones con clientes (*Customer Relationship Management* o CRM). Generan y almacenan información de forma estructurada, almacenada en servidores y bases de datos.

- *Data warehouse*: es un repositorio que concentra la información de todos los sistemas de la empresa, para luego utilizarla en modelos analíticos. La conexión entre los diferentes sistemas es por medio de una o más variables en común, como fechas, clientes, sucursales, departamento, empleado, etc. Como resultado de estos modelos analíticos se encuentran patrones, anomalías, tendencias y comunicarlos a través de dashboards.

- La ciencia de datos se complementa con especialistas en la infraestructura y flujos de información dentro de la organización, que regularmente forman las funciones de Tecnologías de la Información (TI).

Actividad

Para reforzar tu comprensión del tema, contesta las siguientes preguntas:

1. Investiga y menciona el tipo de información que ha crecido exponencialmente en el área de la educación.
2. Describe de manera general los componentes en el ciclo del Big Data para el caso de un hospital.
3. Comenta las características de la información que puedes recolectar de las redes sociales, por ejemplo, en LinkedIn.

Cierre

Juan ahora cuenta con el conocimiento de las principales estrategias en que se originan datos, dentro y fuera de su empresa. Lo anterior le permite identificar sin esfuerzo cómo podría obtener datos, incluso si estos no existen, para posteriormente explorarlos.

Este conocimiento sobre las fuentes y el origen de datos representa los cimientos para comenzar a pensar como un científico de datos, ya que si se desconoce de dónde provienen o dónde pueden obtenerse, prácticamente no hay insumos para trabajar: no se puede explotar lo que no se tiene (datos).

Bibliografía

- Dasgupta, N. (2018). *Practical Big Data Analytics: Hands-on Techniques to Implement Enterprise Analytics and Machine Learning Using Hadoop, Spark, NoSQL, and R*. Packt Publishing.
- Pehcevski, J. (2019). *Big Data Analytics - Methods and Applications*. Arcler Press.

Tecmilenio no guarda relación alguna con las marcas mencionadas como ejemplo. Las marcas son propiedad de sus titulares conforme a la legislación aplicable, estas se utilizan con fines académicos y didácticos, por lo que no existen fines de lucro, relación publicitaria o de patrocinio.
