
Universidad
Tecmilenio®

Universidad Tecmilenio

Ingeniería de datos masivos

Fundamentos de data science

Semana 5

Introducción

Amazon es un extraordinario ejemplo de cómo una compañía puede crear ventaja competitiva y obtener valor al extraer grandes volúmenes de información.

Lo anterior le ha permitido llegar a ser la empresa de comercio electrónico y de computación en la nube más grande del mundo, gracias a la comercialización de su tecnología enfocada en el manejo de grandes volúmenes de información a través de Amazon Web Services.

Fue pionera en el comercio electrónico e introdujo un sistema de recomendación personalizada basado en Big Data, lo cual permite analizar y realizar predicciones a partir de millones de transacciones de sus clientes.

Explicación

Gartner (s.f.) define **Big Data** como los activos de información de gran volumen, velocidad y alta variedad que necesitan formas rentables e innovadoras de procesamiento de la información para una mejor comprensión y toma de decisiones.

La exponencial creación de datos digitales, así como la disposición a una mayor capacidad de cómputo a menor costo para almacenarlos ha hecho posible tener acceso a grandes volúmenes de información y tecnologías como las siguientes:

1. **Inteligencia artificial** (IA)
2. **Aprendizaje automático** (machine learning)
3. **Aprendizaje profundo** (deep learning)
4. **Cómputo en la nube** (cloud computing)

Explicación

Universidad Tecmilenio

Mills (2019) enlista algunos de los beneficios de utilizar Big Data en las organizaciones:

Acceso y visibilidad de información valiosa en el momento adecuado para tomar mejores decisiones.
Segmentación de poblaciones para personalizar acciones.
Habilita la experimentación para descubrir necesidades y exponer variabilidades, mejorando la ejecución.
Habilita la toma de decisiones humanas con algoritmos automatizados.
Habilita la innovación de nuevos modelos de negocio, productos y servicios.

Big Data proporciona datos para proyectos de ciencia y minería de datos, mediante los cuales extrae el valor de estos para revelar tendencias, generar información para mejorar la toma de decisiones, habilitar procesos de innovación en la operación o para crear productos y servicios más innovadores.

Explicación

Ciencia de datos

Ciencia multidisciplinaria que combina el conocimiento del negocio, la computación y las matemáticas, mediante una metodología científica para extraer valor de los datos, revelar tendencias y generar información para mejorar la toma de decisiones.

Ingeniería de datos

Se encarga de definir, diseñar, implementar y operar formas rentables de la infraestructura de datos, así como de diseñar y administrar la arquitectura de flujo de datos.

Minería de datos

Consiste en extraer patrones e identificar relaciones para hacer predicciones en un conjunto de datos determinados.

Explicación

En **data science**, **data mining** y **data engineering** se utilizan las tecnologías de inteligencia artificial (IA), aprendizaje automático (ML) y aprendizaje profundo (DL) para desarrollar proyectos.

La **inteligencia de negocios** (BI) consiste en hacer visualizaciones y *dashboards* (tableros dinámicos), así como administrar, organizar y guardar los datos para crear información a partir de estos.

Data science utiliza Big Data, data mining y data engineering para sus proyectos.

Data science y data mining se apoyan en el uso de **estadísticas** para procesar datos en algoritmos de aprendizaje automático (ML).

La **ciencia de datos** es la evolución de la analítica de los negocios que con el tiempo ha integrado tecnologías como bases de datos, cómputo en la nube, estadística, aprendizaje automático, aprendizaje profundo e inteligencia artificial.

Explicación

La **ciencia de datos** trata de hacer análisis más complejos, así como realizar algoritmos y modelos de datos para descubrir, aprender y responder a la pregunta: ¿por qué? Además, analiza lo que podemos hacer en el futuro.

Analítica descriptiva

Patrones/relaciones. Se trata de analizar datos, encontrar patrones, descubrir relaciones entre variables o diferencias significativas entre grupos.

Analítica predictiva

Tendencia/comportamiento. Analiza la información para predecir tendencias y patrones de comportamiento. En el pasado, presente o futuro.

Analítica prescriptiva

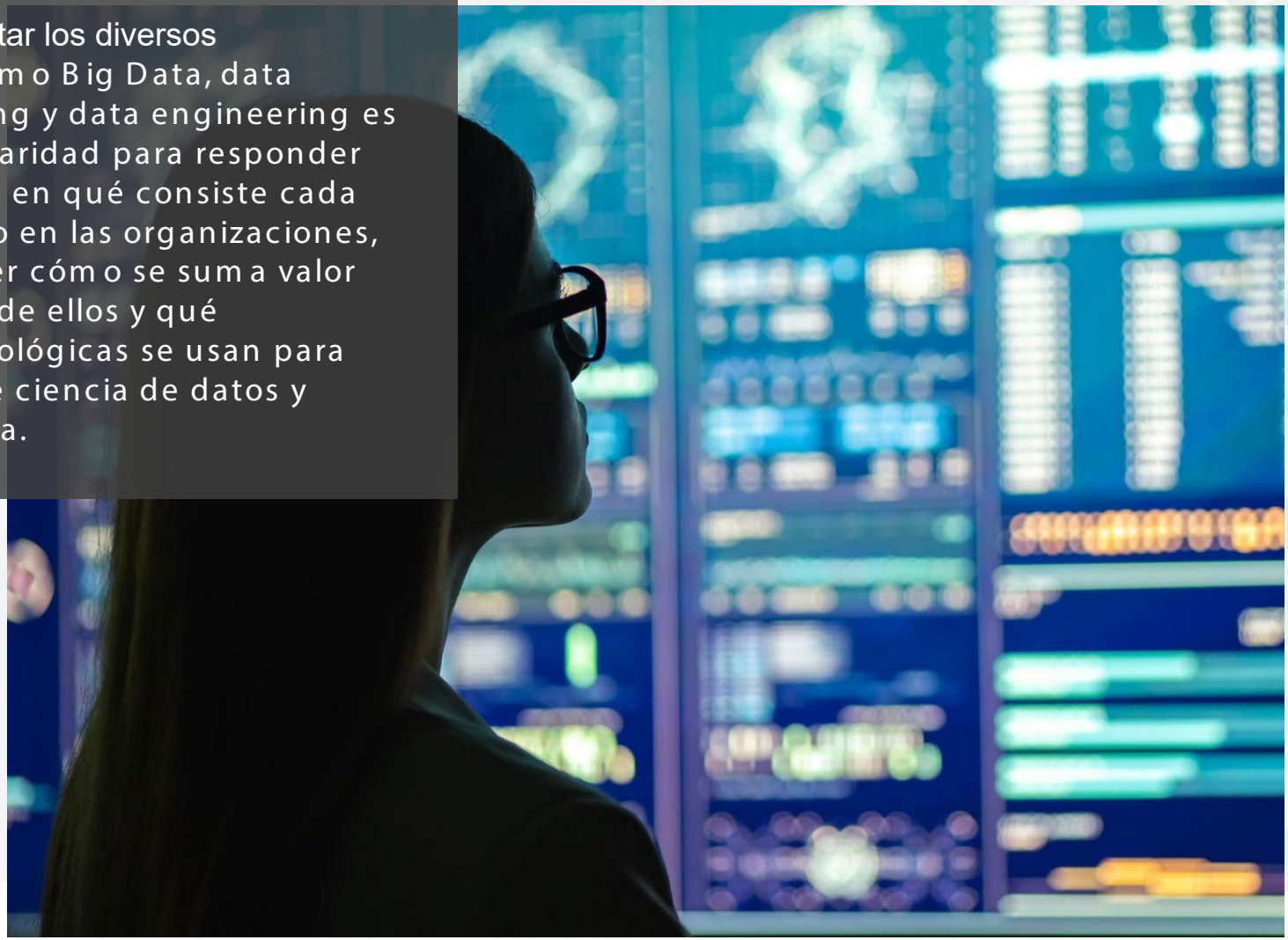
Futuro/acciones. Se centra en informar acerca de lo que debería suceder y busca mejorar el resultado esperado mediante la toma de acciones.

Existe un conjunto de proveedores de tecnología y herramientas que permiten el manejo de grandes volúmenes de datos distribuidos, por ejemplo, **Amazon Web Services**, **Microsoft Azure** y **Google Cloud**, que ofrecen servicios de cómputo y analítica en la nube.

Asimismo, existen proveedores especializados en diferentes áreas y tipos de tratamiento o procesamiento de datos como ETL (Extract Transform Load), bases de datos, *data warehouse*, etcétera.

Cierre

El objetivo de presentar los diversos conceptos tales como Big Data, data science, data mining y data engineering es que logres tener claridad para responder de forma detallada en qué consiste cada uno de ellos, su uso en las organizaciones, así como para saber cómo se suma valor agregado a través de ellos y qué herramientas tecnológicas se usan para hacer proyectos de ciencia de datos y almacenar Big Data.



Bibliografía

- Gartner. (s.f.). *Gartner Glossary.* Recuperado de https://www.gartner.com/en/information-technology/glossary/big-data

- Mills, T. (2019). *Five Benefits Of Big Data Analytics And How Companies Can Get Started.* Recuperado de https://www.forbes.com/sites/forbestechcouncil/2019/11/06/five-benefits-of-big-data-analytics-and-how-companies-can-get-started/#18e174fb17e4

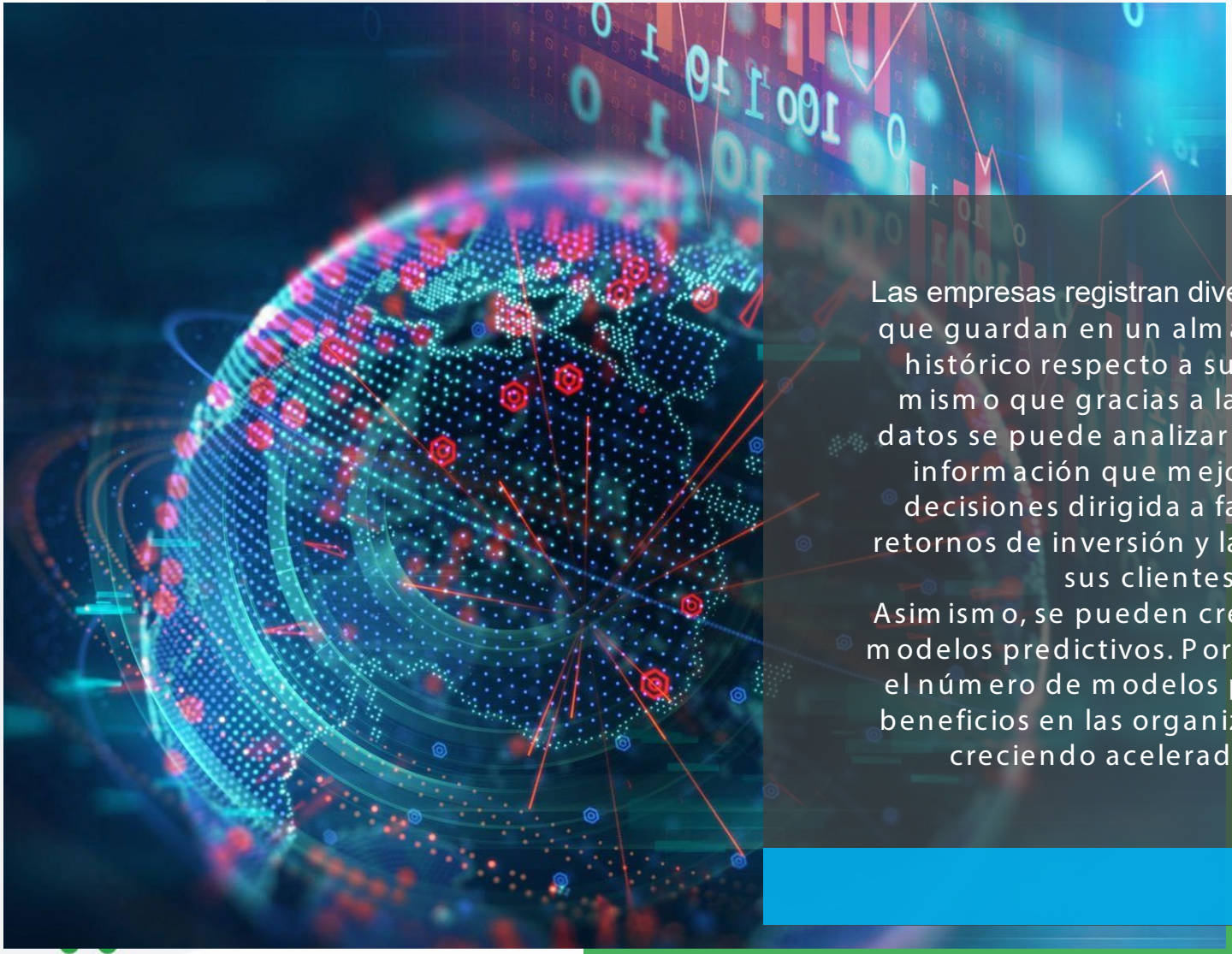
Universidad Tecmilenio

Ingeniería de datos masivos

Data mining para Big Data

Semana 5

Introducción



Las empresas registran diversos datos que guardan en un almacenamiento histórico respecto a su operación, mismo que gracias a la minería de datos se puede analizar para generar información que mejore la toma decisiones dirigida a favorecer los retornos de inversión y la relación con sus clientes.

Asimismo, se pueden crear diferentes modelos predictivos. Por consiguiente, el número de modelos para generar beneficios en las organizaciones está creciendo aceleradamente.

Explicación

La **minería de datos** consiste en extraer patrones e identificar relaciones para hacer predicciones en un conjunto de datos determinado. Por lo tanto, este proceso implica un almacenamiento intensivo de datos, haciendo uso de tecnologías computacionales, tales como el aprendizaje automático, el aprendizaje profundo y el cómputo en la nube.

El objetivo principal de la minería de datos es **el descubrimiento de información que mejore la toma de decisiones.**

Explicación

Algunas ventajas y desventajas de la minería de datos (Itelligent, 2016):

Ventajas

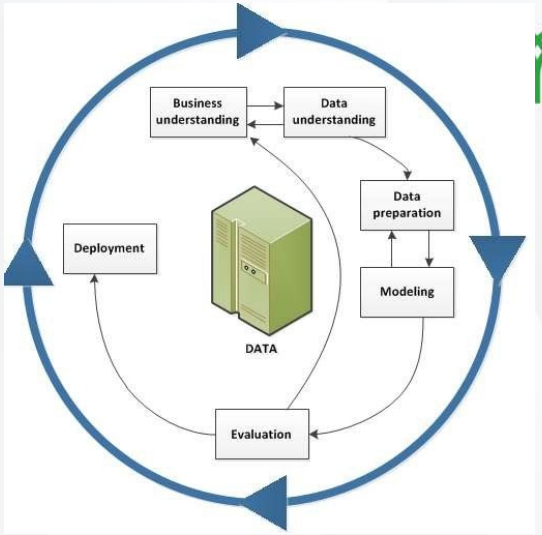
- Permite descubrir información que no se esperaba encontrar, generando un valor añadido para las organizaciones.
- Analiza bases de datos de gran tamaño.
- Obtiene resultados que pueden ser comprendidos de manera sencilla.
- Ayuda en la toma de decisiones estratégicas en una organización.
- La organización logra ofrecer los productos y servicios que necesitan sus clientes.
- Genera modelos de forma rápida.

Desventajas

- Existen dificultades en la recopilación de los datos.
- Ocasionalmente requiere de una gran inversión en las tecnologías para llevarla a cabo.
- Carece de un sistema de seguridad adecuado para la información.

Explicación

La metodología CRISP -DM (Cross Industry Process for Data Mining) es un estándar para desarrollar proyectos de minería de datos, el cual consiste en las siguientes etapas (Villena, 2016):



Fuente: IBM Knowlegde Center. (s.f.). *Conceptos básicos de ayuda de CRISP - DM*. Recuperado de https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crisp_ddita/clementine/crisp_help/crisp_overview.html

Entendimiento del negocio

Comprender los objetivos del proyecto y elaborar un plan para alcanzarlos.

1

Entendimiento de datos

Recolección de datos inicial y actividades para lograr un mejor entendimiento de los mismos.

2

Preparación de datos

Transformar los datos en bruto de inicio para edificar el conjunto final de los mismos.

3

Creación de modelos

Se eligen y utilizan las técnicas de modelado adecuadas al problema.

4

Evaluación de modelos

Evaluar a detalle y revisar los pasos para llevarlo a cabo.

5

Implementación

Se pone en marcha el modelo.

6

Cierre

La minería de datos pertenece a una rama existente dentro de la ciencia de datos, la cual permite realizar la exploración y el análisis de grandes volúmenes de datos para descubrir patrones sobresalientes en la información, mediante una serie de pasos conocidos como la metodología CRISP-DM.

Su relevancia ha crecido debido a que cada año la cantidad de datos aumenta de manera exponencial.

Bibliografía

- IBM Knowlegde Center. (s.f.). *Conceptos básicos de ayuda de CRISP - DM.* Recuperado de https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html

- Itelligent. (2016). *10 ventajas de la minería de datos.* Recuperado de https://itelligent.es/es/10-ventajas-la- mineria-web/

- Villena, J. (2016). *CRISP-DM: La metodología para poner orden en los proyectos.* Recuperado de https://www.sngular.com/es/data- science-crisp-dm-metodologia/