



Data Science



Definición de Data Science y Big Data

Big data es un término general para cualquier colección de conjuntos de datos tan grandes o complejos que se vuelve difícil procesarlos utilizando técnicas tradicionales de gestión de datos como, por ejemplo, los RDBMS (sistemas de gestión de bases de datos relacionales).

Los RDBMS ampliamente se han considerado durante mucho tiempo como una solución única para todos, pero las demandas de manejo de macrodatos han demostrado lo contrario. La ciencia de datos implica el uso de métodos para analizar cantidades masivas de datos y extraer el conocimiento que contienen. Puede pensar en la relación entre big data y ciencia de datos como si fuera la relación entre el petróleo crudo y una refinería de petróleo. Ciencia de datos y big data son la evolución de las estadísticas y la gestión de datos tradicionales, pero ahora se considera que son disciplinas distintas.



Definición de Data Science y Big Data

Las características de big data a menudo se denominan las tres V:

- Volumen: ¿Cuántos datos hay?
- Variedad: ¿Qué tan diversos son los diferentes tipos de datos?
- Velocidad: ¿A qué velocidad se generan nuevos datos?



Definición de Data Science y Big Data

Las características de big data a menudo se denominan las tres V:

- Volumen: ¿Cuántos datos hay?
- Variedad: ¿Qué tan diversos son los diferentes tipos de datos?
- Velocidad: ¿A qué velocidad se generan nuevos datos?

A menudo estas características se contemplan con una cuarta V:

- ¿Qué tan precisa son los datos?



Definición de Data Science y Big Data

La ciencia de datos es una extensión evolutiva de la estadística capaz de abordar cantidades masivas de datos producidos hoy. Agrega métodos desde la informática al repertorio de la estadística.

Las principales cosas que define a un científico de datos además de un estadístico, están la capacidad de trabajar con big data y la experiencia en aprendizaje automático, informática y creación de algoritmos. Algunas de las herramientas que utilizan son: Hadoop, Pig, Spark, R, Python y Java, entre otros.

Python es un gran lenguaje para la ciencia de datos porque tiene muchas bibliotecas de ciencia de datos disponibles y cuenta con un amplio respaldo de software. Por ejemplo, casi todas las bases de datos NoSQL populares tienen un código específico de la API de Python. Debido a estas características y la capacidad de crear prototipos, rápidamente con Python ha manteniendo un rendimiento aceptable, su influencia crece constantemente en el mundo de la ciencia de datos.



Beneficios y usos de la Data Science y Big Data

Las empresas comerciales de casi todas las industrias utilizan ciencia de datos y big data para obtener información sobre sus clientes, procesos, personal, finalización y productos. Muchas de las empresas utilizan la ciencia de datos para ofrecer a los clientes una mejor experiencia de usuario, así como para realizar ventas cruzadas, aumentar las ventas y personalizar sus ofertas.

Un buen ejemplo de esto es Google AdSense, que recopila datos de los usuarios de Internet para que los mensajes comerciales relevantes puedan coincidir con la persona que navega por Internet.

Las instituciones financieras utilizan la ciencia de datos para predecir mercados de valores, determinar el riesgo de prestar dinero y aprender a atraer nuevos clientes por sus servicios.



Beneficios y usos de la Data Science y Big Data

Las organizaciones gubernamentales también son conscientes del valor de los datos. Muchos gobiernos no solo confían en los científicos de datos internos para descubrir información valiosa, sino también compartir sus datos con el público.

Puede utilizar estos datos para obtener información o crear aplicaciones basadas en datos. Data.gov es solo un ejemplo; es el hogar de los EE. UU. Un científico de datos de una organización gubernamental se pone a trabajar en diversos proyectos como la detección de fraudes y otras actividades delictivas o la optimización financiera de proyectos. Edward Snowden proporcionó un ejemplo bien conocido, que filtró documentos internos de la Agencia de Seguridad Nacional Estadounidense y del Gobierno Británico.

Oficinas centrales de comunicaciones que muestran claramente cómo utilizaron la ciencia de datos y big data para monitorear a millones de personas. Esas organizaciones recolectaron 5 mil millones registros de datos de aplicaciones generalizadas como Google Maps, Angry Birds, correo electrónico y mensajes de texto, entre muchas otras fuentes de datos. Luego aplicaron la ciencia de datos técnicas para destilar información.



Beneficios y usos de la Data Science y Big Data

Las organizaciones no gubernamentales (ONG) tampoco son ajenas al uso de datos. Ellos lo utilizan para recaudar fondos y defender sus causas. El Fondo Mundial para la Naturaleza (WWF), por ejemplo, emplea científicos de datos para aumentar la eficacia de su recaudación de fondos. Muchos científicos de datos dedican parte de su tiempo a ayudar a las ONG, porque las ONG a menudo carecen de los recursos para recopilar datos y emplear científicos de datos.

DataKind es un grupo de científicos de datos que dedica su tiempo al beneficio de la humanidad. Las universidades utilizan la ciencia de datos en sus investigaciones, pero también para mejorar la experiencia de estudio de sus alumnos.

El auge de los cursos online abiertos masivos (MOOC) produce una gran cantidad de datos, lo que permite a las universidades estudiar cómo este tipo de aprendizaje puede complementar clases tradicionales. Los MOOC son un activo invaluable si quieres convertirte en un científico de datos y profesional de big data, así que definitivamente mire algunos de los más conocidos: Coursera, Udacity y edX.



Facetas de los datos

En ciencia de datos y macrodatos, se encuentran muchos tipos de datos diferentes, y cada uno de ellos tiende a requerir diferentes herramientas y técnicas. Las principales categorías de datos son:

- Estructurado
- No estructurado
- Lenguaje natural
- Generado por máquina
- Basado en gráficos
- Audio, video e imágenes
- Streaming