Modulo uno

¿Qué es la ciencia de datos?

Cuando tienes un modelo o hipótesis de un problema y tratas de validar esa hipótesis o modelo con tus datos.

Es cuando traduces los datos en una historia, es decir utilizas la narración para generar información, y con esa información puedes tomar decisiones estratégicas para una empresa o una institución.

El intento de trabajar con datos para encontrar respuestas a preguntas que estamos explorando.

Si tienes datos y tienes curiosidad y estás trabajando con datos y los estás manipulando, los estás explorando, el ejercicio mismo de revisarlos, analizarlos, tratar de obtener algunas respuestas de ellos es ciencia de datos.

Fundamentos de la ciencia de datos

Cualquier persona describe ciencia de datos de manera diferente. La mayoría concuerda que tiene un componente importante de análisis de datos.

La ciencia de datos puede ayudar a las organizaciones a comprender sus entornos, analizar problemas existentes, y revelar oportunidades previamente ocultas.

Los buenos científicos de datos son personas curiosas que hacen preguntas para aclarar las necesidades del negocio. Las siguientes preguntas son: ¿qué datos necesitamos para resolver el problema? ¿Y de dónde saldrán esos datos?

Cuando los datos han revelado su potencial, el papel del científico de datos pasa a ser el de un narrador que comunica los resultados a las partes interesadas del proyecto.

Camino hacia la ciencia de datos

La *curiosidad* es absolutamente necesaria. Si no tienes curiosidad, no sabrás qué hacer con los datos. *Juzgador* porque si no tienes nociones preconcebidas sobre las cosas, no sabrías por dónde empezar. *Argumentativo* porque si puedes argumentar y si puedes, puedes defender un caso.

La otra cosa que un científico de datos necesitaría es cierta comodidad y flexibilidad con las plataformas de análisis, algún software, y alguna plataforma informática.

La capacidad de contar una historia. Una vez que tengas tus análisis, una vez que tengas tus tabulaciones, ahora deberías ser capaz de contar una gran historia a partir de ello.

Así que primero debes determinar cuál es tu interés y cuál es tu ventaja competitiva. Tu ventaja competitiva no necesariamente serán tus destrezas analíticas. Su ventaja competitiva es su comprensión de algún aspecto de la vida en el que usted supera a otros en esa comprensión. Quizás sea cine, quizás sea comercio minorista, quizás sea salud, quizás sean computadoras.

Algoritmos	Un conjunto de instrucciones paso a paso para resolver un problema o completar una tarea.

Modelo	Una representación de las relaciones y patrones encontrados en los datos para hacer predicciones o analizar sistemas complejos, manteniendo los elementos esenciales necesarios para el análisis.
Valores atípicos	Cuando un punto de datos o puntos ocurren significativamente fuera de la mayoría de los otros datos en un conjunto de datos, lo que podría indicar anomalías, errores o fenómenos únicos que podrían afectar el análisis estadístico o la modelización.
Análisis cuantitativo	Un enfoque sistemático que utiliza análisis matemáticos y estadísticos para interpretar datos numéricos.
Datos estructurados	Datos organizados y formateados en un esquema predecible, generalmente relacionados con tablas con filas y columnas.
Datos no estructurados	Datos desorganizados que carecen de un modelo de datos o de organización predefinido, lo que dificulta su análisis utilizando métodos tradicionales. Este tipo de datos a menudo incluye texto, imágenes, videos y otro contenido que no encaja perfectamente en filas y columnas como los datos estructurados.

Comprender los distintos tipos de formatos de archivo

Los archivos de texto delimitados son archivos de texto que se utilizan para almacenar datos como texto en el que cada línea o fila tiene valores separados por un delimitador, donde un delimitador es una secuencia de uno o más caracteres para especificar el límite entre entidades o valores independientes. Se puede utilizar cualquier carácter para separar los valores, pero los delimitadores más comunes son la coma el tabulador, los dos puntos, la barra vertical, y el espacio. Los valores separados por comas, o CSV, y los valores separados por tabulaciones, o TSV, son los tipos de archivos más utilizados en esta categoría. En los CSV, el delimitador es una coma, mientras que en los TSV el delimitador es un tabulador. Cuando hay comas literales en los datos de texto y, por lo tanto, no se pueden usar como delimitadores, las TSV sirven como alternativa al formato CSV. Los tabuladores son poco frecuentes en el texto continuo. Los archivos delimitados permiten valores de campo de cualquier longitud y se consideran un formato estándar para proporcionar un esquema de información sencillo.

Hoja de cálculo Open XML de Microsoft Excel, o XLSX, es un formato de archivo Open XML de Microsoft Excel que se incluye dentro del formato de archivo de hoja de cálculo. Puede haber múltiples hojas de cálculo. y cada hoja de cálculo está organizada en filas y columnas en cuya intersección se encuentra la celda. Cada celda contiene datos. XLSX utiliza el formato de archivo abierto, lo que significa que generalmente es accesible para la mayoría de las demás aplicaciones. El lenguaje de marcado extensible o XML es un lenguaje de marcado con reglas establecidas para codificar datos. El formato de archivo XML es legible tanto para humanos como para máquinas. Es un lenguaje autodescriptivo diseñado para enviar información a través de Internet.

XML es similar a HTML en algunos aspectos, pero también tiene diferencias. Por ejemplo, un XML no utiliza etiquetas predefinidas como lo hace HTML. XML es independiente de la plataforma y del lenguaje de programación y, por lo tanto, simplifica el intercambio de datos entre varios sistemas.

El Formato de Documento Portátil, o PDF, es un formato de archivo desarrollado por Adobe para presentar documentos de manera independiente del software de aplicación, hardware y sistemas operativos (sistema operativo), lo que significa que se puede ver de la misma manera en cualquier dispositivo.

ADP. Este formato se utiliza frecuentemente en documentos legales y financieros y también se puede utilizar para rellenar datos como formularios. La Notación de Objetos de JavaScript o JavaScript Object Notation, es un estándar abierto basado en texto diseñado para transmitir datos estructurados a través de la web.

JSON es fácil de usar, es compatible con una amplia gama de navegadores, y se considera una de las mejores herramientas para compartir datos de cualquier tamaño y tipo, incluso audio y vídeo.

Los datos no estructurados son básicamente datos que provienen principalmente de la web, donde no son tabulares. No está en filas y columnas. Es texto. A veces es video y audio. Por lo tanto, habría que implementar algoritmos más sofisticados (más sofisticados que la regresión) para extraer datos. Y, de hecho, muchas veces tomamos Datos no estructurados y pasamos una gran cantidad de tiempo y esfuerzo para obtener algo de Datos estructurados de ellos y luego analizarlos. Así que, si tienes algo que se adapte bien a tablas, columnas y filas, adelante. Esos son datos estructurados, pero si ves que se trata de un registro web o si estás intentando obtener información de páginas web y tienes un montón de páginas web, se trata de datos no estructurados que requerirían un poco más de esfuerzo para extraer la información de ellos.

¿Qué es la regresión?

Si alguna vez has tomado un taxi, entenderás la regresión. Así es como funciona. En el momento en que te sientas en un taxi, en un taxi ves que hay una cantidad fija allí. Dice $ 2,50. Ya sea que el taxi se mueva o usted baje, esto es lo que le debe al conductor. Desde el momento en que te subes a un taxi, eso es una constante. Si has subido a un taxi, tendrás que pagar esa cantidad. Luego, a medida que comienza a moverse, por cada metro o cien metros, la tarifa aumenta en una cierta cantidad. Entonces, hay una fracción, hay una relación entre la distancia y la cantidad que pagarías, más allá de esa constante. y si no te estás moviendo y estás atrapado en el tráfico, entonces por cada minuto adicional tienes que pagar más, así que a medida que aumentan los minutos tu tarifa aumenta, a medida que aumenta la distancia tu tarifa aumenta y mientras todo esto sucede ya has pagado una tarifa base que es la constante, esto es lo que es la regresión, la regresión te dice cuál es la tarifa

base ¿Y cuál es la relación entre el tiempo y la tarifa que has pagado y la distancia que has recorrido y la tarifa que has pagado? Porque en ausencia de conocer esas relaciones y solo saber cuánto viajó la gente y cuánto pagó, la regresión le permite calcular esa constante que no sabía que era 250 y calcularía la relación entre la tarifa y la distancia y la tarifa y el tiempo

¿Qué convierte a alguien en un científico de datos?

Alguien que encuentra soluciones a problemas analizando datos grandes o pequeños utilizando herramientas apropiadas y luego cuenta historias para comunicar sus hallazgos a las partes interesadas relevantes.

Mientras uno tenga una mente curiosa, fluidez en analítica y la capacidad de comunicar los hallazgos, considero que esa persona es un científico de datos.

Un científico de datos tiene que ser alguien con una mente muy curiosa, dispuesto a dedicar tiempo y esfuerzos significativos para explorar sus corazonadas.

Tienden a ser personas realmente curiosas, pensadores que hacen buenas preguntas y están bien al lidiar con situaciones no estructuradas y tratando de encontrar estructura en ellas.

Los científicos de datos investigan y encuentran explicaciones para muchos problemas. Por ejemplo, el Dr. Murtaza Haider encontró una explicación de por qué medio millón de clientes se quejaron del transporte público en Toronto. Después de mucha investigación, encontró una relación entre los eventos climáticos inesperados y la cantidad de quejas ese día en particular.

Granville, autor sobre ciencia de datos enumera el álgebra, el cálculo, y la formación en probabilidad y estadística como antecedentes educativos necesarios para ser un científico de datos.

Un científico de datos exitoso es una combinación de científico informático, ingeniero de software, y estadístico. Su capacidad para transformar soluciones no estructuradas en conocimientos estructurados define su destreza.

Término	Definición
Valores separados por comas (CSV) / Valores separados por tabulaciones (TSV)	Formato comúnmente utilizado para almacenar datos tabulares como texto plano donde el valor es separado por una coma o una tabulación.
Tipos de archivos de datos	Una configuración de archivo informático diseñada para almacenar datos de una manera específica.

Formato de datos	Cómo se codifican los datos para que puedan ser almacenados dentro de un tipo de archivo de datos.
Visualización de datos	Una forma visual, como un gráfico, de representar datos de manera fácilmente comprensible que facilita la identificación de tendencias en los datos.
Archivo de texto delimitado	Un archivo de texto plano donde un carácter específico separa los valores de los datos.
Lenguaje de Marcado Extensible (XML)	Un lenguaje diseñado para estructurar, almacenar y permitir el intercambio de datos entre diversas tecnologías.
Hadoop	Un marco de trabajo de código abierto diseñado para almacenar y procesar grandes conjuntos de datos en clústeres de computadoras.
Notación de Objetos de JavaScript (JSON)	Un formato de datos compatible con varios lenguajes de programación para que dos aplicaciones intercambien datos estructurados.
Notebooks de Jupyter	Un entorno computacional que permite a los usuarios crear y compartir documentos que contienen código, ecuaciones, visualizaciones y texto explicativo. Consulta los notebooks de Python.
Vecino más cercano	Un algoritmo de aprendizaje automático que predice una variable objetivo-basada en su similitud con otros valores en el conjunto de datos.

Redes neuronales	Un modelo computacional utilizado en el aprendizaje profundo que imita la estructura y el funcionamiento de las vías neuronales del cerebro humano. Toma una entrada, la procesa utilizando el aprendizaje previo y produce una salida.
Pandas	Una biblioteca de Python de código abierto que proporciona herramientas para trabajar con datos estructurados y que se utiliza a menudo para la manipulación y análisis de datos.
Notebooks de Python	También conocido como un notebook “Jupyter”, este entorno computacional permite a los usuarios crear y compartir documentos que contienen código, ecuaciones, visualizaciones y texto explicativo.
R	Un lenguaje de programación de código abierto utilizado para computación estadística, análisis de datos y visualización de datos.
Motor de recomendación	Un programa informático que analiza la entrada del usuario, como comportamientos o preferencias, y hace recomendaciones personalizadas basadas en ese análisis.
Regresión	Un modelo estadístico que muestra una relación entre una o más variables predictoras y una variable de respuesta.
Datos tabulares	Datos que están organizados en filas y columnas.
XLSX	El formato de archivo de hoja de cálculo de Microsoft Excel.

Modulo dos

¿Como los macrodatos impulsan la transformación digital?

La disponibilidad de grandes cantidades de datos y la ventaja competitiva que supone su análisis han desencadenado transformaciones digitales en muchas industrias. Netflix pasó de ser un sistema de préstamo de DVD por correo a uno de los principales proveedores de transmisión de video del mundo. El equipo MBA de los Houston Rockets utilizó datos recopilados por cámaras aéreas para analizar las jugadas más productivas, y Lufthansa analizó los datos de los clientes para mejorar su servicio.

La transformación digital no es simplemente duplicar procesos existentes en forma digital. El análisis en profundidad de cómo funciona el negocio ayuda a las organizaciones a descubrir cómo mejorar sus procesos y operaciones y aprovechar los beneficios de integrar la ciencia de datos en sus flujos de trabajo.

La transformación digital impacta cada aspecto de la organización, por lo que es manejada por los tomadores de decisiones en los niveles más altos para garantizar el éxito. El apoyo del director ejecutivo es crucial para el proceso de transformación digital, como lo es el apoyo del director de información y el rol emergente del director de datos

Introducción a la nube

La computación en la nube, también conocida como nube, es la entrega de recursos informáticos bajo demanda, como redes, servidores, almacenamiento, aplicaciones, servicios y centros de datos, a través de Internet mediante un pago por uso.

El término computación en la nube se puede utilizar para describir aplicaciones y datos a los que los usuarios acceden a través de Internet en lugar de hacerlo en su computadora local. Algunos ejemplos de computación en la nube incluyen usuarios que utilizan aplicaciones web en línea, empleados que utilizan aplicaciones comerciales en línea seguras para realizar su trabajo, y usuarios que almacenan archivos personales en plataformas de almacenamiento basadas en la nube, como Google Drive, OneDrive, y Dropbox.

Comencemos por comprender las cinco características esenciales de la nube. El autoservicio a pedido significa que usted obtiene acceso a los recursos de la nube como la potencia de procesamiento, el almacenamiento, y la red que necesita mediante una interfaz simple sin necesidad de interacción humana con cada proveedor de servicios. El acceso amplio a la red significa que se puede acceder a los recursos de computación en la nube a través de la red mediante mecanismos y plataformas estándar como teléfonos móviles, tabletas, computadoras portátiles, y estaciones de trabajo. La puesta en común de recursos es lo que brinda a los proveedores de la nube economías de escala, que luego transmiten a sus clientes, haciendo que la nube sea rentable. Al utilizar un modelo multiinquilino, los recursos informáticos se agrupan para atender a múltiples consumidores, y los recursos de la nube se asignan y reasignan dinámicamente según la demanda, sin que los clientes necesiten conocer la ubicación física de estos recursos. La elasticidad rápida implica que usted puede acceder a más recursos cuando los necesita y reducirlos cuando no los necesita porque los recursos se aprovisionan y liberan de forma elástica. Y un servicio medido significa que usted sólo paga por lo que usa o reserva a medida que avanza.

Los modelos de implementación de la nube indican dónde reside la infraestructura, quién la posee y la administra, y cómo los recursos y servicios de la nube se ponen a disposición de los usuarios. Hay tres tipos de modelos de implementación de nube: pública, privada e híbrida. La nube pública

es cuando se aprovechan los servicios en la nube a través de Internet abierta, en hardware propiedad del proveedor de la nube, pero su uso es compartido por otras empresas. Nube privada significa que la infraestructura de la nube está destinada para uso exclusivo de una sola organización. Podría ejecutarse localmente, o podría ser propiedad, de un proveedor de servicios y estar administrado y operado por él. Y cuando usas una mezcla de nubes públicas y nube privada, trabajando juntas de manera fluida, eso se clasifica como el modelo de nube híbrida.

Ahora, veamos los tres modelos de servicios en la nube que se basan en las tres capas de una pila informática Infraestructura, Plataforma, y Aplicación. Estos modelos de computación en la nube se denominan acertadamente Infraestructura como Servicio o IaaS, Plataforma como Servicio o PaaS y Software como Servicio o SaaS.

En un modelo IaaS, puede acceder a la infraestructura y a los recursos informáticos físicos, como servidores redes, almacenamiento, y espacio del centro de datos, sin la necesidad de administrarlos ni operarlos. En un modelo PAAS, se puede acceder a la plataforma que comprende las herramientas de hardware y software que normalmente se necesitan para desarrollar e implementar aplicaciones para los usuarios a través de Internet. Y un SAAS es un modelo de entrega y licencia de software en el que el software y las aplicaciones se alojan de forma centralizada y se licencian mediante suscripción. A veces se le denomina software bajo demanda.

Nube para científicos de datos

La nube es una bendición para los científicos de datos principalmente porque les permite tomar sus datos, tomar su información y colocarla en la nube, colocarla en el sistema de almacenamiento central.

La nube no sólo permite almacenar grandes cantidades de datos en servidores en algún lugar de California o Nevada, sino que también permite implementar algoritmos informáticos muy avanzados y la capacidad de realizar cálculos de alto rendimiento utilizando máquinas que no son suyas.

Permite que múltiples entidades trabajen con los mismos datos al mismo tiempo. Así que puedes estar trabajando con los mismos datos que tus colegas en, digamos, Alemania, y otro equipo en India, y otro equipo en Ghana; ellos están trabajando colectivamente y pueden hacerlo porque la información, los algoritmos, las herramientas, las respuestas y los resultados, lo que sea que necesiten, está disponible en un lugar central.

El uso de la nube le permite obtener acceso instantáneo a tecnologías de código abierto como Apache Spark sin la necesidad de instalarlas y configurarlas localmente. El uso de la nube también le brinda acceso a las herramientas y bibliotecas más actualizadas sin la preocupación de mantenerlas y garantizar que estén actualizadas. La nube es accesible desde cualquier lugar y en cualquier zona horaria. Puede utilizar tecnologías basadas en la nube desde su computadora portátil, desde su tableta e incluso desde su teléfono, lo que permite colaborar con más facilidad que nunca.

Algunas grandes empresas tecnológicas ofrecen plataformas en la nube, que le permiten familiarizarse con las tecnologías basadas en la nube en un entorno prediseñado. IBM ofrece IBM Cloud, Amazon ofrece AWS y Google Cloud Platform.

Fundamento de macrodatos

En este mundo digital, cada uno deja huella. Desde nuestros hábitos de viaje hasta nuestros entrenamientos y entretenimiento, la creciente cantidad de dispositivos conectados a Internet con los que interactuamos a diario registran enormes cantidades de datos sobre nosotros. Incluso hay un nombre para ello: Big Data.

Big Data se refiere a los volúmenes de datos dinámicos grandes y dispares que crean las personas, las herramientas y las máquinas. Éstas son las V del big data:

La *velocidad* es la velocidad a la que se acumulan los datos. Los datos se generan extremadamente rápido en un proceso que nunca se detiene. Las tecnologías de transmisión en tiempo real o casi real, locales y basadas en la nube pueden procesar información muy rápidamente. El *volumen* es la escala de los datos o el aumento en la cantidad de datos almacenados. Los impulsores del volumen son el aumento de las fuentes de datos, los sensores de mayor resolución y la infraestructura escalable.

La *variedad* es la diversidad de los datos. Los datos estructurados encajan perfectamente en filas y columnas en Base de datos relacional, mientras que los datos no estructurados no están organizados de una manera predefinida como los tweets, las publicaciones de blogs, las imágenes, los números y los videos. La variedad también refleja que los datos provienen de diferentes fuentes, máquinas, personas, y procesos, tanto internos como externos a las organizaciones.

Los impulsores son las tecnologías móviles, las redes sociales, las tecnologías portátiles, las geotecnologías, el vídeo, y muchos, muchos más. La *veracidad* es la calidad y origen de los datos y su conformidad con los hechos y la exactitud. Los atributos incluyen consistencia, completitud, integridad y ambigüedad. Los factores que impulsan esta tendencia son el costo y la necesidad de trazabilidad.

El *valor* es nuestra capacidad y necesidad de convertir los datos en valor. El valor no es sólo ganancia. Puede tener beneficios médicos o sociales, así como satisfacción del cliente, del empleado o personal. La razón principal por la que las personas invierten tiempo en comprender el big data es para obtener valor de él.

BigData

Son datos que son lo suficientemente grandes y tienen suficiente volumen y velocidad como para que no se puedan manejar con los sistemas de bases de datos tradicionales. Algunos de nuestros estadísticos creen que el big data es algo que no cabe en una memoria USB.

Hadoop

Tradicionalmente, en el cálculo y procesamiento de datos, llevábamos los datos a la computadora. Ejecutarías un programa y llevarías los datos al programa. En un clúster de big data, lo que Larry Page y Sergey Brin idearon es bastante simple: tomaron los datos y los dividieron en pedazos. Y distribuyeron cada uno, y replicaron cada pieza, o triplicaron cada pieza, y la enviaban.

Los fragmentos de estos archivos a miles de computadoras. Al principio eran cientos, ahora son miles, ahora son decenas de miles. Y luego enviarían el mismo programa. a todas estas computadoras en el clúster. Y cada computadora ejecutaría el programa en su pequeña porción del archivo y enviaría los resultados. Los resultados se ordenarían y luego se redistribuirían a otro proceso. El primer proceso se llama mapa o proceso mapeador, y el segundo se llama proceso de reducción.

Yahoo entonces se subió a bordo. Yahoo contrató a alguien llamado Doug Cutting, que había estado trabajando en un clon o una copia de la arquitectura de big data de Google, que ahora se llama Hadoop.

Hadoop es un framework de código abierto que permite almacenar y procesar grandes cantidades de datos. Se usa para aplicaciones de big data y se basa en clústeres de computadoras.

Herramientas de procesamiento de BigData

Las tecnologías de procesamiento de big data proporcionan formas de trabajar con grandes conjuntos de datos estructurados, semiestructurados, y no estructurados para poder obtener valor de los big data.

Hadoop es una colección de herramientas que proporciona almacenamiento distribuido y procesamiento de grandes cantidades de datos. Hive es un almacén de datos para consulta y análisis de datos construido sobre Hadoop.

Spark es un marco de análisis de datos distribuido diseñado para realizar análisis de datos complejos en tiempo real.

Hadoop, un marco de código abierto basado en Java, permite el almacenamiento y procesamiento distribuido de grandes conjuntos de datos a través de clústeres de computadoras. En el sistema distribuido Hadoop, un nodo es una sola computadora y una colección de nodos forma un clúster. Hadoop puede escalar desde un solo nodo a cualquier número de nodos Hadoop ofrece almacenamiento y computación local y proporciona una solución confiable, escalable, y rentable para almacenar datos sin requisitos de formato.

Usando Hadoop, se puede incorporar formatos de datos emergentes como audio en streaming, video, Sentimiento en redes sociales, y datos de clics, junto con Datos estructurados, semi-estructurados, y Datos no estructurados que no se utilizan tradicionalmente en un Almacén de datos. Proporcionar acceso de autoservicio en tiempo real para todas las partes interesadas.

Uno de los cuatro componentes principales de Hadoop es Hadoop Distributed File System, o HDFS, que es un sistema de almacenamiento para big data que se ejecuta en múltiples hardware básicos conectados a través de una red.

HDFS proporciona almacenamiento de big data escalable y confiable al particionar archivos en múltiples nodos Divide archivos grandes en varias computadoras, lo que permite el acceso paralelo a ellos. Por lo tanto, los cálculos pueden ejecutarse en paralelo en cada nodo donde se almacenan los datos. También replica bloques de archivos en diferentes nodos para evitar la pérdida de datos, lo que lo hace tolerante a fallas.

Ejemplo. Considere un archivo que incluye números de teléfono de todos en los Estados Unidos. Los números de las personas con apellidos que comienzan con A podrían estar almacenados en el Servidor 1, B en el Servidor 2, y así sucesivamente. Con Hadoop, partes de esta libreta telefónica se almacenarían en todo el clúster. Para reconstruir la libreta telefónica completa, su programa necesitará los bloques de cada servidor del clúster. HDFS también replica estas piezas más pequeñas en dos servidores adicionales de forma predeterminada, lo que garantiza la disponibilidad cuando un servidor falla.

Otros beneficios que se obtienen al usar HDFS incluyen una recuperación rápida de fallas de hardware porque HDFS está diseñado para detectar fallas y recuperarse automáticamente. Acceso a datos en streaming porque HDFS admite altas tasas de transferencia de datos Alojamiento de grandes conjuntos de datos porque HDFS puede escalar a cientos de nodos o computadoras en un solo clúster Portabilidad porque HDFS es portátil en múltiples plataformas de hardware y compatible con una variedad de sistemas operativos subyacentes

Hive es un software de almacenamiento de datos de código abierto para leer, escribir y administrar archivos de grandes conjuntos de datos que se almacenan directamente en HDFS u otros sistemas

de almacenamiento de datos como Apache HBase. Hadoop está diseñado para exploraciones secuenciales largas y debido a que Hive se basa en Hadoop, las consultas tienen una latencia muy alta, lo que significa que Hive es menos apropiado para aplicaciones que necesitan tiempos de respuesta muy rápidos.

Hive es más adecuado para tareas de Almacenamiento de datos como ETL, informes y análisis de datos, e incluye herramientas que permiten un fácil acceso a los datos a través de SQL.

Esto nos lleva a Spark, un motor de procesamiento de datos de propósito general diseñado para extraer y procesar grandes volúmenes de datos para una amplia gama de aplicaciones, incluidos análisis interactivos, procesamiento de flujos, aprendizaje automático, integración de datos, y ETL. Aprovecha el procesamiento en memoria para aumentar significativamente la velocidad de los cálculos y el volcado al disco solo cuando la memoria está limitada. Spark tiene interfaces para los principales lenguajes de programación, como Java, Scala, Python, R, y SQL. Puede ejecutarse utilizando su tecnología de clustering independiente, así como también sobre otras infraestructuras como Hadoop. Además, puede acceder a datos de una gran variedad de fuentes, incluidas HDFS y Hive, lo que lo hace muy versátil. La capacidad de procesar datos de transmisión rápidamente y realizar análisis complejos en tiempo real es el caso de uso clave de Apache Spark.

Término	Definición
Analítica	El proceso de examinar datos para extraer conclusiones y tomar decisiones informadas es un aspecto fundamental de la ciencia de datos, que implica análisis estadístico e información basada en datos.
Big Data	Cantidades vastas de datos estructurados, semi-estructurados y no estructurados se caracterizan por su volumen, velocidad, variedad y valor, que, al ser analizados, pueden proporcionar ventajas competitivas y impulsar transformaciones digitales.
Clúster de Big Data	Un entorno de computación distribuida que comprende miles o decenas de miles de computadoras interconectadas que almacenan y procesan grandes conjuntos de datos de manera colectiva.

Acceso Amplio a la Red	La capacidad de acceder a recursos en la nube a través de mecanismos y plataformas estándar como dispositivos móviles, laptops y estaciones de trabajo a través de redes.
Director de Datos (CDO)	Un rol emergente responsable de supervisar iniciativas, gobernanza y estrategias relacionadas con los datos, asegurando que los datos desempeñen un papel central en los esfuerzos de transformación digital.
Director de Información (CIO)	Un ejecutivo responsable de gestionar la tecnología de la información y los sistemas informáticos de una organización, contribuyendo a los aspectos tecnológicos de la transformación digital.
Computación en la Nube	La entrega de recursos informáticos bajo demanda, incluidos redes, servidores, almacenamiento, aplicaciones, servicios y centros de datos, a través de Internet en una base de pago por uso.
Modelos de Implementación en la Nube	Categorías que indican dónde reside la infraestructura en la nube, quién la gestiona y cómo se ponen a disposición de los usuarios los recursos y servicios en la nube, incluidos modelos públicos, privados e híbridos.
Modelos de Servicio en la Nube	Modelos basados en las capas de una pila de computación, incluidos Infraestructura como Servicio (IaaS), Plataforma como Servicio (PaaS) y Software como Servicio (SaaS), representan diferentes ofertas de computación en la nube.
Hardware de Commodidad	Componentes de hardware estándar y comerciales utilizados en un clúster de big data, que ofrecen soluciones rentables para almacenamiento y procesamiento sin depender de hardware especializado.

Algoritmos de Datos	Procedimientos computacionales y modelos matemáticos utilizados para procesar y analizar datos accesibles en la nube para que los científicos de datos los implementen de manera eficiente en grandes conjuntos de datos.
Replicación de Datos	Una estrategia en la que los datos se duplican en múltiples nodos en un clúster para garantizar la durabilidad y disponibilidad de los datos, reduciendo el riesgo de pérdida de datos debido a fallas de hardware.
Ciencia de Datos	Un campo interdisciplinario que implica extraer información y conocimiento de los datos utilizando diversas técnicas, incluidas programación, estadísticas y herramientas analíticas.
Aprendizaje Profundo	Un subconjunto del aprendizaje automático que involucra redes neuronales artificiales inspiradas en el cerebro humano, capaces de aprender y tomar decisiones complejas a partir de los datos por sí solas.
Cambio Digital	La integración de tecnología digital en los procesos y operaciones comerciales conduce a mejoras e innovaciones en cómo las organizaciones operan y entregan valor a los clientes.
Transformación Digital	Un cambio organizacional estratégico y cultural impulsado por la ciencia de datos, especialmente Big Data, para integrar la tecnología digital en todas las áreas de la organización, resultando en cambios fundamentales en las operaciones y en la entrega de valor.
Datos Distribuidos	La práctica de dividir datos en fragmentos más pequeños y distribuirlos a través de múltiples computadoras dentro de un clúster permite el procesamiento paralelo para el análisis de datos.

Hadoop	Un marco de almacenamiento y procesamiento distribuido utilizado para manejar y analizar grandes conjuntos de datos, particularmente bien adaptado para análisis de big data y aplicaciones de ciencia de datos.
Sistema de Archivos Distribuidos de Hadoop (HDFS)	Un sistema de almacenamiento dentro del marco de Hadoop que particiona y distribuye archivos a través de múltiples nodos, facilitando el acceso paralelo a los datos y la tolerancia a fallos.
Infraestructura como Servicio (IaaS)	Un modelo de servicio en la nube que proporciona acceso a infraestructura informática, incluidos servidores, almacenamiento y redes, sin que los usuarios necesiten gestionarlos u operarlos.
Marco Basado en Java	Hadoop se implementa en Java, un lenguaje de programación de alto nivel y de código abierto, que proporciona la base para construir soluciones de almacenamiento y procesamiento distribuidas.
Proceso Map	El paso inicial en el modelo de programación MapReduce de Hadoop, donde los datos se procesan en paralelo en nodos individuales del clúster, a menudo utilizado para tareas de transformación de datos.
Servicio Medido	Una característica donde los usuarios son facturados por los recursos en la nube según su uso real, con la utilización de recursos monitorizada, medida e informada de manera transparente.
Autoservicio Bajo Demanda	La capacidad de los usuarios para acceder y provisionar recursos en la nube, como potencia de procesamiento, almacenamiento y redes, utilizando interfaces simples sin interacción humana con los proveedores de servicios.

Elasticidad Rápida	La capacidad de escalar rápidamente los recursos en la nube hacia arriba o hacia abajo según la demanda, permitiendo a los usuarios acceder a más recursos cuando los necesitan y liberarlos cuando no están en uso.
Proceso Reduce	El segundo paso en el modelo MapReduce de Hadoop, donde los resultados del proceso de mapeo se agregan y procesan más para producir la salida final, utilizado típicamente para análisis.
Replicación	El acto de crear copias de piezas de datos dentro de un clúster de big data mejora la tolerancia a fallos y garantiza la disponibilidad de los datos en caso de fallas de hardware o nodos.
Agrupamiento de Recursos	Una característica de la nube donde los recursos informáticos se comparten y se asignan dinámicamente a múltiples consumidores, promoviendo economías de escala y eficiencia de costos.
Labs de Skills Network (SN Labs)	Recursos de aprendizaje proporcionados por IBM, incluidos herramientas como Jupyter Notebooks y clústeres de Spark, disponibles para los estudiantes para proyectos de ciencia de datos en la nube y desarrollo de habilidades.
Volcado en Disco	Una técnica utilizada en situaciones con limitaciones de memoria donde los datos se escriben temporalmente en almacenamiento en disco cuando los recursos de memoria se agotan, asegurando un procesamiento ininterrumpido.
Clases STEM	Cursos de Ciencia, Tecnología, Ingeniería y Matemáticas (STEM) que se enseñan típicamente en las escuelas secundarias preparan a los estudiantes para carreras técnicas, incluida la ciencia de datos.

Variedad	La diversidad de tipos de datos, incluidos datos estructurados y no estructurados de diversas fuentes como texto, imágenes, video y más, que plantea desafíos en la gestión de datos.
Velocidad	La velocidad a la que se acumulan y generan los datos, a menudo en tiempo real o casi en tiempo real, impulsa la necesidad de procesamiento y análisis de datos rápidos.
Veracidad	La calidad y precisión de los datos, asegurando que se ajusten a los hechos y sean consistentes, completos y libres de ambigüedad, impacta en la fiabilidad y confianza de los datos.
Sistema de Seguimiento de Video	Un sistema utilizado para capturar y analizar datos de video de juegos, permitiendo un análisis profundo de los movimientos de los jugadores y la dinámica del juego, contribuyendo a la toma de decisiones basada en datos en el deporte.
Volumen	La escala de datos generados y almacenados es impulsada por el aumento de fuentes de datos, sensores de mayor resolución e infraestructura escalable.
Las V de Big Data	Un conjunto de características comunes en las definiciones de Big Data, que incluye Velocidad, Volumen, Variedad, Veracidad y Valor, destacando la rápida generación, escala, diversidad, calidad y valor de los datos.

Inteligencia artificial y ciencia de datos

El término big data se refiere a conjuntos de datos que son tan masivos, se construyen tan rápidamente y son tan variados que desafían los métodos de análisis tradicionales, como los que se podrían realizar con una base de datos relacional. El big data suele describirse en términos de cinco V: velocidad, volumen, variedad, veracidad y valor.

La minería de datos es el proceso de búsqueda y análisis automático de datos. Descubriendo patrones previamente no revelados. Implica preprocesar los datos para prepararlos y transformarlos en un formato apropiado. Una vez hecho esto, se extraen conocimientos y patrones utilizando

diversas herramientas y técnicas que van desde simples herramientas de visualización de datos hasta aprendizaje automático y modelos estadísticos.

El aprendizaje automático es un subconjunto de la IA que utiliza algoritmos informáticos para analizar datos y tomar decisiones inteligentes basadas en lo que ha aprendido sin estar programado explícitamente. Los algoritmos de aprendizaje automático se entrenan con grandes conjuntos de datos y aprenden a partir de ejemplos. No siguen algoritmos basados en reglas. El aprendizaje automático es lo que permite a las máquinas resolver problemas por sí solas y hacer predicciones precisas utilizando los datos proporcionados.

El aprendizaje profundo es un subconjunto especializado del aprendizaje automático que utiliza redes neuronales en capas para simular la toma de decisiones humanas. Los algoritmos de aprendizaje profundo pueden etiquetar y categorizar información e identificar patrones. Es lo que permite a los sistemas de IA aprender continuamente en el trabajo y mejorar la calidad y precisión de los resultados al determinar si las decisiones fueron correctas.

Las redes neuronales artificiales, a menudo denominadas simplemente redes neuronales, se inspiran en las redes neuronales biológicas, aunque funcionan de forma bastante diferente. Una red neuronal en IA es una colección de pequeñas unidades informáticas llamadas neuronas que toman datos entrantes y aprenden a tomar decisiones a lo largo del tiempo. Las redes neuronales suelen tener capas profundas y son la razón por la que los algoritmos de aprendizaje profundo se vuelven más eficientes a medida que los conjuntos de datos aumentan en volumen. a diferencia de otros algoritmos de aprendizaje automático que pueden estancarse a medida que aumentan los datos.

La ciencia de datos puede utilizar muchas de las técnicas de IA para obtener información de los datos. Por ejemplo, podría utilizar algoritmos de aprendizaje automático e incluso modelos de aprendizaje profundo para extraer significado y sacar inferencias de los datos. Existe cierta interacción entre la IA y la ciencia de datos, pero una no es un subconjunto de la otra. Más bien, la ciencia de datos es un término amplio que abarca toda la metodología de procesamiento de datos, mientras que la IA incluye todo lo que permite a las computadoras aprender a resolver problemas y tomar decisiones inteligentes. Tanto la IA como la ciencia de datos pueden implicar el uso de big data, es decir, volúmenes significativamente grandes de datos.

IA generativa y ciencia de datos

La IA generativa es un subconjunto de la inteligencia artificial que se centra en producir nuevos datos en lugar de simplemente analizar datos existentes. Permite que las máquinas creen contenido, incluyendo imágenes, música, lenguaje, código informático y más, imitando las creaciones de las personas.

Los científicos de datos pueden ampliar sus conjuntos de datos utilizando IA generativa para crear datos sintéticos. Crea estos datos con propiedades similares a los datos reales, como su distribución, agrupamiento, y muchos otros factores que la IA aprendió sobre el conjunto de datos reales. Los científicos de datos pueden luego utilizar estos datos sintéticos junto con datos reales para el entrenamiento y prueba de modelos.

Con la IA generativa, los científicos de datos pueden aprovecharla para generar y probar código de software para construir modelos analíticos. La automatización de la codificación tiene el potencial de revolucionar el campo de la analítica, permitiendo al científico de datos centrarse en tareas de

nivel superior, como identificar y aclarar el problema que los modelos pretenden resolver y evaluar hipótesis de una gama más amplia de fuentes de datos. La IA generativa puede generar información empresarial precisa e informes completos, lo que permite actualizar esta información a medida que evolucionan los datos. Además, puede explorar datos de forma autónoma para descubrir patrones ocultos y conocimientos que podrían pasar desapercibidos durante el análisis manual y mejorar la toma de decisiones.

Redes neuronales y aprendizaje profundo

Una Red neuronal intenta usar un programa de computadora que imite cómo las neuronas, cómo nuestros cerebros utilizan las neuronas para procesar cosas, neuronas a sinapsis, y construir estas redes complejas que pueden ser entrenadas. Así comienza una red neuronal. con algunas entradas y algunas UO. Y sigues introduciendo estas entradas para intentar ver qué tipos de transformaciones darán lugar a estas salidas. Y sigues haciendo esto una y otra y otra vez de manera que esta red converja para que estas entradas, las transformaciones eventualmente obtengan estos resultados.

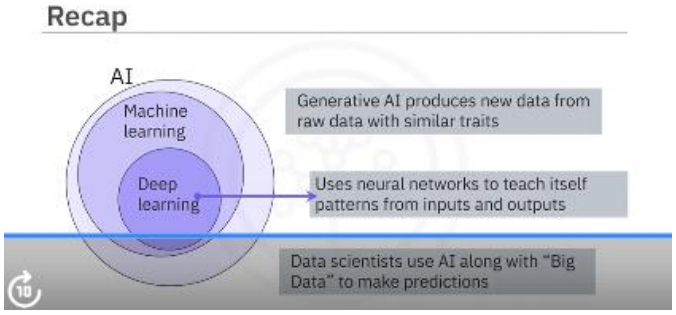
El aprendizaje profundo son redes neuronales con esteroides. Lo que hicieron fue tener múltiples capas de redes neuronales, y usaron mucha, mucha, mucha potencia de cómputo para resolverlas.

Reconocer el habla, reconocer personas, imágenes, clasificar imágenes, casi todas las tareas tradicionales que las redes neuronales solían hacer en cosas pequeñas ahora pueden hacer cosas realmente grandes. Aprenderá por sí solo la diferencia entre un gato y un perro y diferentes tipos de objetos. No tiene que ser enseñado. No, simplemente aprende. Por eso lo llaman aprendizaje profundo.

Aplicaciones del aprendizaje profundo

Algunas aplicaciones del aprendizaje automático en FinTech son probablemente un par de cosas diferentes de las que puedo hablar allí. Una de ellas son las recomendaciones, ¿verdad? Entonces, cuando usas Netflix o usas Facebook o muchos otros servicios de software, las recomendaciones te llegan, es decir, oye, eres un usuario, has visto este programa, así que tal vez te gustaría ver este otro programa. correcto o sigues a esta persona así que tal vez deberías seguir a esa otra persona en realidad es algo similar en Fintech.

Otro tema del que la gente habla y que es importante, especialmente en el comercio minorista, en los aspectos minoristas de la banca y las finanzas, es la detección de fraudes. Intentando determinar si un cargo que llega a través de una tarjeta de crédito es fraudulento o no en tiempo real Es un problema de aprendizaje automático, ¿verdad? Tienes que aprender de todas las transacciones que han ocurrido previamente y construir un modelo.



Término	Definición
Redes neuronales artificiales	Colecciones de pequeñas unidades de computación (neuronas) que procesan datos y aprenden a tomar decisiones a lo largo del tiempo.
Análisis bayesiano	Una técnica estadística que utiliza el teorema de Bayes para actualizar las probabilidades basándose en nueva evidencia.
Perspectivas empresariales	Los conocimientos y los informes precisos generados por la IA generativa se pueden actualizar a medida que evolucionan los datos, lo que mejora la toma de decisiones y descubre patrones ocultos.
Análisis de conglomerados	El proceso de agrupar puntos de datos similares en función de determinadas características o atributos.
Automatización de codificación	Uso de IA generativa para generar y probar automáticamente código de software para construir modelos analíticos, liberando a los científicos de datos para que se concentren en tareas de nivel superior.
Minería de datos	El proceso de buscar y analizar datos automáticamente para descubrir patrones y conocimientos que antes eran desconocidos.
Árboles de decisión	Un tipo de algoritmo de aprendizaje automático que se utiliza para la toma de decisiones mediante la creación de una estructura de decisiones en forma de árbol.
Modelos de aprendizaje profundo	Incluye redes generativas antagónicas (GAN) y autocodificadores variacionales (VAE) que crean nuevas instancias de datos aprendiendo patrones de grandes conjuntos de datos.
Las cinco V del Big Data	Características utilizadas para describir big data: velocidad, volumen, variedad, veracidad y valor.
IA generativa	Un subconjunto de IA que se centra en crear nuevos datos, como imágenes, música, texto o código, en lugar de simplemente analizar datos existentes.

Análisis de la cesta de la compra	El análisis de qué bienes tienden a comprarse juntos se utiliza a menudo para obtener información de marketing.
Bayes ingenuo	Un algoritmo de clasificación probabilística simple basado en el teorema de Bayes.
Procesamiento del lenguaje natural (PLN)	Un campo de la IA que permite a las máquinas comprender, generar e interactuar con el lenguaje humano, revolucionando la creación de contenido y los chatbots.
Precisión vs. recuperación	Las métricas se utilizan para evaluar el rendimiento de los modelos de clasificación.
Análisis predictivo	Utilizando técnicas de aprendizaje automático para predecir resultados o eventos futuros.
Datos sintéticos	Datos generados artificialmente con propiedades similares a los datos reales, utilizados por científicos de datos para aumentar sus conjuntos de datos y mejorar el entrenamiento de modelos.