Responsabilidades de un Analista de Datos

Si bien la función de un analista de datos varía según el tipo de organización y la medida en que haya adoptado prácticas basadas en datos, en las organizaciones actuales hay algunas responsabilidades que son típicas de la función de un analista de datos. Entre ellas figuran las siguientes: Adquirir datos de fuentes de datos primarias y secundarias, Crear consultas para extraer los datos necesarios de las bases de datos y otros sistemas de recopilación de datos, Filtrar, limpiar, normalizar y reorganizar los datos en preparación para el análisis de los mismos, Utilizar instrumentos estadísticos para interpretar los conjuntos de datos, Utilizar técnicas estadísticas para identificar patrones y correlaciones en los datos, Analizar patrones en conjuntos de datos complejos e interpretar tendencias, Preparar informes y gráficos que comuniquen eficazmente las tendencias y los patrones, Crear la documentación apropiada para definir y demostrar los pasos del proceso de análisis de datos. En correspondencia con estas responsabilidades, veamos algunas de las habilidades que son valiosas para un Analista de Datos. El proceso de análisis de datos requiere una combinación de aptitudes técnicas, funcionales y de programación. Veamos primero algunos de los conocimientos técnicos que necesitas en tu función como Analista de Datos. Estos incluyen: Experiencia en el uso de hojas de cálculo como Microsoft Excel o Google Sheets, Dominio de herramientas y software de análisis estadístico y visualización como IBM Cognos, IBM SPSS, Oracle Visual Analyzer, Microsoft Power BI, SAS y Tableau. Dominio de al menos uno de los lenguajes de programación como R, Python y, en algunos casos, C++, Java y MATLAB. Buen conocimiento de SQL, y capacidad para trabajar con datos en bases de datos relacionales y NoSQL. La capacidad de acceder y extraer datos de los repositorios de datos como data marts, data warehouses, data lakes y data pipelines. Familiaridad con las herramientas de procesamiento de Big Data como Hadoop, Hive y Spark. Aprenderemos más sobre las características y casos de uso de algunos de estos lenguajes de programación, bases de datos, repositorios de datos y herramientas de procesamiento de Big Data más adelante en el curso. Ahora veremos algunas de las habilidades funcionales que se requieren para el papel de Analista de Datos. Estas incluyen: Dominio de la Estadística para ayudarte a analizar los datos, validar tu análisis, e identificar las faltas y errores lógicos. Habilidades analíticas que te ayuden a investigar e interpretar datos, teorizar y hacer pronósticos. Habilidades de resolución de problemas, porque en última instancia, el objetivo final de todo análisis de datos es resolver problemas. Habilidades de indagación que son esenciales para el proceso de descubrimiento, es decir, para comprender un problema desde la perspectiva de los diversos interesados y usuarios, porque el proceso de análisis de datos comienza realmente con una clara articulación del enunciado del problema y el resultado deseado. Habilidades de visualización de datos que ayudan a decidir las técnicas e instrumentos que presentan los hallazgos de manera eficaz en función de la audiencia, el tipo de datos, el contexto y el objetivo final del análisis. Habilidades en Gestión de Proyectos para manejar el proceso, personas, dependencias y plazos de la iniciativa. Eso nos lleva a tus habilidades como Analista de Datos. El análisis de datos es tanto una ciencia como un arte. Puedes mejorar la experiencia técnica y funcional, pero uno de los diferenciadores clave para tu éxito serán las habilidades sociales. Esto incluye tu habilidad para trabajar en colaboración con equipos empresariales y multifuncionales; comunicarte de manera efectiva para informar y presentar tus hallazgos; contar una historia atractiva y convincente; y conseguir apoyo y aceptación para tu trabajo. Por encima de todo, tener curiosidad está en el corazón del análisis de datos. En el curso de tu trabajo, te tropezarás con patrones, fenómenos y anomalías que pueden mostrarte un camino diferente. La capacidad de permitir que salgan a la superficie nuevas preguntas y desafiar tus suposiciones e hipótesis hace que seas un gran analista. También oirás a los profesionales del análisis de datos hablar de la intuición como una cualidad imprescindible. Es esencial señalar que la intuición, en este contexto, es la capacidad de tener un sentido del futuro basado en el reconocimiento de patrones y experiencias pasadas. En este video, hemos aprendido sobre las responsabilidades y habilidades de un Analista de Datos. En el próximo vídeo, te guiaremos a través de un día en la vida de un Analista de Datos.

Puntos de Vista: Cualidades y habilidades para ser un Analista de Datos

En este video, vamos a escuchar a profesionales de datos en ejercicio hablar de las cualidades y habilidades necesarias para convertirse en un analista de datos. >> Las cualidades y habilidades de un analista de datos, una persona que es curiosa por naturaleza. Alguien que tiene atención a los detalles y disfruta trabajando con computadoras. Una persona curiosa a veces buscará respuestas incluso cuando no hay una pregunta, o no le importa investigar y buscar en áreas sobre las que no se ha reflexionado antes. ¿Atención al detalle o búsqueda de patrones? ¿Entras en una habitación y cuentas de forma natural las personas, cómo está configurada la habitación, prestando atención a los detalles cercanos? ¿y luego disfrutas de la informática? porque la tecnología se mueve muy rápido. Algo o una habilidad que aprendiste hoy, en dos o tres años puede no ser aplicable. Así que necesitas ser capaz de desarrollar nuevas habilidades y aprender nuevos programas dependiendo de cómo ha cambiado el mercado o la industria. >> Se requieren habilidades técnicas y habilidades sociales. Las habilidades técnicas incluyen Python, SQL, R, Tableau y Power BI, y las habilidades sociales o interpersonales significan si sabes qué datos utilizar y cuál es la herramienta correcta a utilizar, y cómo presentar los datos a las principales partes interesadas. Y estas habilidades requieren visión de negocios y habilidades de presentación. >> Tienes que estar muy orientada a los detalles, te tienen que gustar los números, te tiene que gustar la información y estar dispuesta a mirar esa información y no sólo mirarla superficialmente, sino también a sumergirte más profundamente. Así que por ejemplo, en lo que hacemos, no puedo tomar un extracto bancario a valor nominal, tengo que mirarlo y compararlo. ¿El sello tiene buen aspecto? Especialmente en el mundo actual hay mucho fraude y mala comunicación, y gente que intenta obtener tu información y usarla de forma fraudulenta. Así que un buen analista de datos debería ser capaz de comparar la información del año pasado con la de este año para ver si parece correcta. Debes tener su mirada y esa mentalidad, y no sólo tomar las cosas al pie de la letra. >> Hay muchas cualidades y habilidades requeridas para ser analista de datos y las desgloso en dos cubos, básicamente, habilidades sociales y habilidades técnicas. Creo que las habilidades sociales más importantes para un analista de datos es ser muy curioso, hacer muchas preguntas buenas, ser muy reflexivo, y escuchar cuidadosamente y entender tanto la perspectiva del usuario como la de tus compañeros de trabajo y lo que más necesitan obtener de los datos, y siempre estar dispuesto a aprender, porque el análisis es un campo que se mueve rápidamente, así que tienes que estar constantemente aprendiendo y leyendo para estar al día. Hay muchas habilidades técnicas que se requieren para ser analista de datos. La habilidad técnica más importante para cualquier nuevo analista de datos que esté aprendiendo es SQL. Es, con mucho, el más utilizado, y cada vez que extraigas datos de una base de datos, necesitarás saber SQL. Y no hay nada como un analista de datos con muy buenas habilidades en SQL. Creo que a veces la gente se adelanta y prueba un montón de tecnologías muy complicadas antes de aprender bien lo básico de SQL, y creo que eso es un gran error. Creo que siempre está bien saber Python y R, que son los dos principales lenguajes de programación que se utilizan para el análisis de datos. Creo que como analista de datos novato, no necesitas ser experto en ambos o en ninguno de ellos. Pero empezar a ser bueno en uno u otro va a ser muy útil para tu carrera. Otra habilidad técnica importante que debe tener un analista de datos es ser realmente bueno en al menos una herramienta de visualización de datos y comprender los principios generales de la visualización de datos. >> Hoy en día, el conjunto de habilidades de un analista de datos es mucho más dinámico de lo que solía ser. Así que los analistas de datos necesitan saber qué problema están tratando de resolver con los datos. Extraer de los datos a medida que lo necesiten, la estructura que necesiten, utilizando SQL, a partir del data lake en el que se encuentran. Sabes que habrá muchas tablas diferentes y necesitarán averiguar cómo unirlas y luego sacar los datos, limpiarlos, prepararlos, manipularlos, extraerlos para que sean capaces de obtener una especie de nueva visión a partir de ellos. Presentar esos conocimientos de manera concisa, usando claramente buenas visualizaciones y cuadros de mando, y, en otras palabras, ser capaz de contar una buena historia con esos datos.

Un día en la vida de un Analista de Datos

Un día en la vida de un analista de datos puede incluir una serie de posibilidades: desde adquirir datos de diversas fuentes de datos hasta crear consultas para extraer datos de repositorios de datos, rebuscar entre las filas de datos para encontrar percepciones, crear informes y cuadros de mando e interactuar con las partes interesadas para reunir información y presentar las conclusiones, es todo un abanico. Y sí, lo más grande - limpiar y preparar los datos para que los hallazgos tengan una base creíble — que, por cierto, es una gran parte de lo que cualquier analista de datos puede tener que hacer en su trabajo. Pero si tuviera que guiarte a través de un día "tipo", voy a elegir uno en el que esté rebuscando entre los datos para obtener conocimientos. Esta es la parte de mi trabajo con la que estoy totalmente fascinado. Hola. Soy Sivaram Jaladi. Trabajo como analista de datos en Fluentgrid, una empresa de soluciones de tecnología para redes inteligentes con sede en Vishakhapatnam, India. Fluentgrid es socio de IBM y ha recibido premios Beacon de IBM por sus soluciones en las áreas de energía inteligente y ciudades inteligentes. Ofrecemos soluciones de centros de operaciones integradas para empresas de energía y ciudades inteligentes, aprovechando nuestra práctica plataforma de inteligencia conocida como Fluentgrid Actilligence. Nuestro cliente, una empresa de servicios públicos de energía del sur de la India, ha estado notando un aumento en las quejas relacionadas con sobrefacturación. Y la frecuencia de estas quejas parece sugerir que hay algo más que sucesos aleatorios. Así que me piden que mire las quejas y los datos de facturación, a ver si puedo detectar algo. Empiezo por hacer un balance de lo que tengo. Algunos de los lugares obvios que sé que voy a mirar son los datos de las quejas, los datos de información de los abonados y los datos de facturación. Ese será mi punto de partida. Antes de entrar en los detalles de los datos, voy a hacer una lista de preguntas, hipótesis iniciales, con las que voy a empezar. Como el patrón de uso de los abonados que informan sobre este tema: ¿Existe un rango de consumo para el cual se está sobrefacturando más que para otros? Concentración de quejas por zonas: ¿Las quejas se concentran en localidades específicas dentro de la ciudad? Frecuencia y ocurrencia de las quejas basadas en los abonados individuales: ¿Informan los mismos abonados sobre la sobrefacturación de manera repetitiva? En caso afirmativo, ¿cuál es la frecuencia de los casos repetidos? Si se sobrefactura a un abonado una vez, ¿la sobrefacturación se produce todos los meses desde la primera vez, o las repeticiones son esporádicas o no se producen en absoluto? A medida que me aclaro en mis hipótesis iniciales y en el conjunto de preguntas con las que voy a empezar, identifico los conjuntos de datos que voy a aislar y analizar para validar o refutar mis hipótesis. Extraigo los promedios de facturación anual, trimestral y mensual de los demandantes y busco un rango en el que las demandas se acumulen más que en otros. Luego saco los datos de ubicación de los reclamantes para ver si hay una conexión entre la sobrefacturación y los códigos postales. Aquí veo lo que parece ser una concentración de quejas en ciertas áreas. Esto parece que podría significar algo. Así que en lugar de pasar a la tercera hipótesis, decido profundizar un poco más en estos datos. A continuación, saco la fecha de los datos de conexión. Más del 95% de los reclamantes habían sido abonados nuestros durante más de siete años, aunque no todos los abonados de más de siete años tenían esta queja. Así que ahora, vemos una concentración por áreas, y vemos una concentración significativa de quejas basadas en la fecha de conexión. A continuación, saco la marca y el número de serie de los contadores. Y ahí está, los números de serie pertenecían al mismo lote de contadores proporcionados por el mismo proveedor. La concentración de estos contadores, y por lo tanto de las quejas, venía de las áreas en las que se instalaron estos contadores. En este momento, me siento confiado para presentar estos hallazgos a las partes interesadas. También voy a compartir las fuentes de datos y mi proceso para llegar a este análisis — eso siempre es muy importante para dar credibilidad a los hallazgos. Este podría ser el final de este proyecto, o podría muy bien volver. Tal vez las mismas quejas con diferentes puntos en común, o un conjunto completamente diferente de quejas para las que necesitamos encontrar respuestas.

Puntos de Vista: Aplicaciones del Análisis de Datos

En este video, profesionales de datos en activo hablan sobre algunas de las aplicaciones del análisis de datos en el mundo de hoy.

Reproduce el video desde ::17 y sigue la transcripción0:17

>> Las aplicaciones del análisis de datos en el mundo de hoy están en todas partes. Para cada publicidad que ves, alguien ha tenido que analizar e identificar, ya sea a partir del consumidor o por parte de la empresa, qué información quiere compartir. Así que ya sabes, cuatro de cada 10 dentistas, o verás información relacionada con el recuento de calorías o las reacciones a ciertas cosas, todo eso ha requerido un análisis. Esto no es algo que deba pensarse por separado y aparte, es lo que hacemos todos los días en nuestras vidas. Incluso las personas que controlan su nivel de azúcar por la diabetes, siempre hay análisis, por lo que las aplicaciones son universales. >> Así que lo mejor del análisis en esta época es que se puede aplicar ampliamente. Cada sector, cada estructura, cada función dentro de una organización dada puede beneficiarse de los datos y el análisis. Ya sea por hacer análisis de canales de ventas, ya sea por hacer análisis financieros a fin de mes, creando informes predefinidos y estandarizados con formato. . O si estás haciendo algo como la planificación o revisión del recuento de personal, todo esto a través de cada estructura vertical, como dije, ya sea para aerolíneas, farmacéuticas, bancos, todos estos y las funciones que los integran pueden beneficiarse de los análisis. >> Y en este clima en el que estamos ahora mismo con la pandemia, hay empresas que están prestando mucha atención a los hábitos de compra de sus clientes. Obviamente, estos hábitos pueden haber variado de lo que estas compañías esperaban que fueran. Así que ahora el análisis de datos es más importante porque necesitan asegurarse de que pueden virar y mantenerse al día con la demanda. Y realmente ser capaces de atender lo que sus clientes y consumidores quieren. >> Puedo hablar de las aplicaciones de la analítica de datos en las finanzas, se trata de que hemos visto más y más aplicaciones de la analítica de datos alternativa en el mundo de las finanzas. Por ejemplo, podemos utilizar el análisis de sentimiento sobre los tweets y las nuevas historias para complementar el análisis financiero tradicional y para fundamentar mejores decisiones de inversión. Además, los datos de las imágenes satelitales pueden utilizarse para seguir el desarrollo de las actividades industriales. Y los datos de geolocalización pueden utilizarse para rastrear el tráfico de las tiendas y predecir el volumen de ventas.

Resumen y Aspectos Resaltantes

En esta lección, has aprendido la siguiente información:

El rol de un Analista de Datos se extiende a lo largo de todo:

● Adquirir los datos que mejor se adapten al caso de uso.

● Preparar y analizar los datos para entender lo que representan.

● Interpretar y comunicar eficazmente el mensaje a las partes interesadas que deben actuar en función de las conclusiones.

● Asegurarse de que el proceso esté documentado para su futura referencia y repetibilidad.

Para desempeñar esta función con éxito, los analistas de datos necesitan una combinación de conocimientos técnicos, funcionales y de programación.

● Los conocimientos técnicos incluyen diversos niveles de competencia en el uso de hojas de cálculo, instrumentos estadísticos, instrumentos de visualización, lenguajes de programación y de consulta al igual que la capacidad de trabajar con diferentes tipos de repositorios de datos y grandes plataformas de datos.

● La comprensión de la Estadística, las Técnicas analíticas, la capacidad de investigar una situación desde múltiples perspectivas, la visualización de datos y las aptitudes de gestión de proyectos–son habilidades funcionales que un analista de datos necesita para desempeñar un rol eficaz.

● Las aptitudes funcionales incluyen la capacidad de trabajar en colaboración, comunicarse con eficacia, contar una historia convincente con datos y obtener el apoyo y la aceptación de los interesados. La curiosidad por explorar diferentes caminos y la intuición que ayuda a dar un sentido del futuro basado en las experiencias pasadas son también aptitudes esenciales para ser un buen Analista de Datos.

Visión General del Ecosistema del Analista de Datos

El ecosistema de un analista de datos incluye infraestructura, software, herramientas, frameworks y procesos utilizados para recopilar, limpiar, analizar, extraer y visualizar datos. En este vídeo daremos una rápida visión general del ecosistema antes de entrar en los detalles de cada uno de estos temas en los vídeos siguientes. Hablemos primero de los datos. Basándonos en lo bien definida que esté la estructura de los datos, éstos se pueden clasificar en estructurados, semiestructurados o no estructurados. Los datos que siguen un formato rígido y que pueden organizarse ordenadamente en filas y columnas son datos estructurados. Estos son los datos que se ven habitualmente en las bases de datos y las hojas de cálculo, por ejemplo. Los datos semiestructurados son una mezcla de datos que tienen características consistentes y datos que no se ajustan a una estructura rígida. Por ejemplo, los correos electrónicos. Un correo electrónico tiene una mezcla de datos estructurados, como el nombre del remitente y el destinatario, pero también tiene el contenido del correo electrónico, que es un dato no estructurado. Y luego hay datos no estructurados: Datos que son complejos, y en su mayoría información cualitativa que es imposible reducir a filas y columnas. Por ejemplo, fotos, videos, archivos de texto, PDFs, y contenidos de redes sociales. El tipo de datos determina el tipo de repositorios de datos en los que se pueden recoger y almacenar los datos, y también las herramientas que se pueden utilizar para consultar o procesar esos datos. Los datos también se presentan en una amplia variedad de formatos de archivo que se recopilan de diversas fuentes de datos, que van desde las bases de datos relacionales y no relacionales, hasta APIs, servicios web, flujos de datos, plataformas de redes sociales y dispositivos con sensores. Esto nos lleva a los repositorios de datos: Un término que incluye bases de datos, data warehouses ("almacenes de datos"), data marts, data lakes, y almacenes de big data. El tipo, formato y fuentes de datos influyen en el tipo de repositorios de datos que se pueden utilizar para recopilar, almacenar, limpiar, analizar y extraer los datos para su análisis. Si estás trabajando con big data, por ejemplo, necesitarás grandes data warehouses, que te permitan almacenar y procesar datos de gran volumen a alta velocidad, y también frameworks que te permitan realizar análisis complejos en tiempo real sobre el big data. El ecosistema también incluye lenguajes que pueden clasificarse como lenguajes de consulta, lenguajes de programación, y lenguajes de shell y de programación de scripts. Desde la consulta y la manipulación de datos con SQL hasta el desarrollo de aplicaciones de datos con Python, y la escritura de secuencias de comandos de shell para tareas operativas repetitivas, estos son componentes importantes en la caja de herramientas de un analista de datos. Las herramientas, frameworks y procesos automatizados son parte del ecosistema de los analistas de datos en todas las etapas del proceso de análisis. Desde las herramientas utilizadas para recopilar, extraer, transformar y cargar datos en los repositorios de datos, hasta las herramientas para data wrangling ("preparación de los datos"), limpieza de datos, data mining ("extracción de datos"), análisis y visualización de datos - es un ecosistema muy diverso y rico. Las hojas de cálculo, cuadernos Jupyter y Cognos de IBM son sólo algunos ejemplos. Abordaremos algunas de las herramientas de análisis de datos con mayor detalle en las siguientes secciones del curso.

Tipos de Datos

Los datos son información no organizada que se procesa para convertirla en significativa. Por lo general, los datos abarcan hechos, observaciones, percepciones, números, caracteres, símbolos e imágenes que pueden interpretarse para obtener un significado. Una de las formas en que los datos se pueden categorizar es por su estructura. Los datos pueden ser: Estructurados; Semiestructurados, o No Estructurados. Los datos estructurados tienen una estructura bien definida o se adhieren a un modelo de datos especificado, se pueden almacenar en esquemas bien definidos como bases de datos, y en muchos casos se pueden representar de manera tabular con filas y columnas. Los datos estructurados son hechos y números objetivos que se pueden recopilar, exportar, almacenar y organizar en bases de datos típicas. Algunas de las fuentes de datos estructurados podrían ser: Bases de datos SQL y Sistemas de Procesamiento de Transacciones Online (u OLTP) que están enfocados a transacciones empresariales, Hojas de cálculo como Excel y Google Spreadsheets, Formularios Online, Sensores como Sistemas de Posicionamiento Global (o GPS) y etiquetas de Identificación por Radiofrecuencia (o RFID); y logs de Redes y servidores Web. Normalmente se pueden almacenar datos estructurados en bases de datos relacionales o SQL. También puedes examinar fácilmente los datos estructurados con métodos y herramientas estándar de análisis de datos. Los datos semiestructurados son datos que tienen algunas propiedades de organización pero carecen de un esquema fijo o rígido. Los datos semiestructurados no se pueden almacenar en forma de filas y columnas como en las bases de datos. Contienen etiquetas y elementos, o metadatos, que se utilizan para agrupar los datos y organizarlos en una jerarquía. Algunas fuentes de datos semiestructurados podrían ser: Correos electrónicos, XML y otros lenguajes de marcado, Ejecutables binarios, paquetes TCP/IP, Archivos comprimidos, Integración de datos de diferentes fuentes. XML y JSON permiten que los usuarios definan etiquetas y atributos para almacenar datos en forma jerárquica y se utilizan de manera muy amplia para almacenar e intercambiar datos semiestructurados. Los datos no estructurados son los que no tienen una estructura fácilmente identificable y, por lo tanto, no se pueden organizar en una base de datos relacional convencional en forma de filas y columnas. No siguen ningún formato, secuencia, semántica o reglas particulares. Los datos no estructurados pueden afrontar fuentes heterogéneas y tienen diversas aplicaciones en materia de inteligencia de negocios y análisis. Algunas de las fuentes de datos no estructurados podrían incluir: Páginas web, Feeds de redes sociales, Imágenes en diversos formatos de archivo (como JPEG, GIF y PNG), archivos de vídeo y audio, documentos y archivos PDF, presentaciones en PowerPoint, logs multimedia; y encuestas. Los datos no estructurados se pueden almacenar en archivos y documentos (como Word doc) para su análisis manual o en bases de datos NoSQL, que tienen sus propios instrumentos de análisis para examinar este tipo de datos. En resumen, los datos estructurados son los que están bien organizados en formatos que pueden almacenarse en bases de datos y se prestan a métodos y herramientas de análisis de datos estándar; los datos semiestructurados son los que están de alguna manera organizados y dependen de metaetiquetas para agruparlos y jerarquizarlos; y los datos no estructurados son los que no están organizados convencionalmente en forma de filas y columnas en un formato determinado. En el próximo video, aprenderemos acerca de los diferentes tipos de estructuras de archivos.

Entendiendo los Diferentes Tipos de Formatos

Como profesional de los datos, trabajarás con diversos tipos de archivos de datos y formatos. Es importante comprender la estructura subyacente de los formatos de archivo junto con sus beneficios y limitaciones. Esta comprensión te ayudará a tomar las decisiones correctas sobre los formatos más adecuados para satisfacer tus requisitos de datos y rendimiento. Algunos de los formatos de archivo estándar que cubriremos en este vídeo incluyen: Formatos de archivo de texto delimitado, Hoja de cálculo Microsoft Excel Open XML, o XLSX, Lenguaje de Marcado Extensible, o XML, Formato de Documento Portátil, o PDF, Notación de Objeto JavaScript, o JSON, Los archivos de texto delimitado son archivos de texto utilizados para almacenar datos como texto en los 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 archivo más utilizados en esta categoría. En los CSV, el delimitador es una coma mientras que en los TSV, el delimitador es una tabulación. Cuando en los datos de texto hay comas literales y, por lo tanto, no se pueden utilizar como delimitadores, los TSV sirven como alternativa al formato CSV. Los tabuladores son poco frecuentes en texto continuo. Cada fila, o línea horizontal, del archivo de texto tiene un conjunto de valores separados por el delimitador, y representa un registro. La primera fila funciona como un encabezado de columna, donde cada columna puede tener un tipo de datos diferente. Por ejemplo, una columna puede ser de tipo fecha, mientras que otra puede ser de tipo cadena o datos de tipo entero. Los archivos delimitados permiten valores de campo de cualquier longitud y se consideran un formato estándar para proporcionar un esquema de información sencillo. Pueden procesarse con casi todas las aplicaciones existentes. Los delimitadores también representan uno de los diversos medios para especificar los límites en un flujo de datos. La Hoja de cálculo Microsoft Excel Open XML, o XLSX, es un formato de archivo Microsoft Excel Open XML que corresponde a un formato de archivo de hoja de cálculo. Es un formato de archivo basado en XML creado por Microsoft. En un .XLSX, también conocido como libro de trabajo, puede haber múltiples hojas de trabajo. Y cada hoja de trabajo está organizada en filas y columnas, y en cada intersección de ambas se encuentra la celda. Cada celda contiene datos. XLSX utiliza el formato de archivo abierto, lo que significa que suele ser accesible para la mayoría de las otras aplicaciones. Puede utilizar y guardar todas las funciones disponibles en Excel y también se sabe que es uno de los formatos de archivo más seguros, ya que no puede guardar código malicioso. El Lenguaje de Marcado Extensible, o XML, es un lenguaje de marcado con reglas establecidas para codificar los datos. El formato de archivo XML es legible tanto por humanos como por 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 el .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 independientemente del software de aplicación, el hardware y los sistemas operativos, lo que significa que se puede ver de la misma manera en cualquier dispositivo. Este formato se utiliza con frecuencia en documentos jurídicos y financieros, y también se puede utilizar para rellenar datos, por ejemplo, para formularios. La Notación de Objetos JavaScript, o JSON, es un estándar abierto basado en texto diseñado para transmitir datos estructurados a través de la web. El formato de archivo es un formato de datos independiente del lenguaje que puede leerse en cualquier lenguaje de programación. JSON es fácil de usar, es compatible con una amplia gama de navegadores y está considerado como una de las mejores herramientas para compartir datos de cualquier tamaño y tipo, incluso de audio y vídeo. Esa es una razón, muchas APIs y Servicios Web devuelven los datos como JSON. En este vídeo, hemos visto algunos formatos de archivos y de datos populares. En el próximo video, aprenderemos acerca de las diferentes fuentes de datos.

Fuentes de Datos

Como hemos mencionado en uno de nuestros videos anteriores, las fuentes de datos nunca han sido tan dinámicas y diversas como lo son hoy en día. En este vídeo, veremos algunas fuentes comunes como: Bases de Datos Relacionales, Archivos Planos y Conjuntos de Datos XML, APIs y Servicios Web, Extracción de Datos Web, Flujos de Datos y Canales de Datos.

Reproduce el video desde ::32 y sigue la transcripción0:32

Por lo general, las organizaciones cuentan con aplicaciones internas que les ayudan a gestionar sus actividades comerciales diarias, las transacciones con los clientes, las actividades de recursos humanos y sus flujos de trabajo. Estos sistemas utilizan bases de datos relacionales como SQL Server, Oracle, MySQL e IBM DB2, para almacenar datos de forma estructurada. Los datos almacenados en bases de datos y data warehouses se pueden utilizar como fuente para el análisis. Por ejemplo, los datos procedentes de un sistema de transacciones de venta minorista se pueden utilizar para analizar las ventas en diferentes regiones, y los datos de un sistema de gestión de relaciones con los clientes se pueden utilizar para hacer proyecciones de ventas. Fuera de la organización, existen otros conjuntos de datos disponibles, tanto públicos como privados. Por ejemplo, las organizaciones gubernamentales que publican continuamente conjuntos de datos demográficos y económicos. También hay empresas que venden datos específicos, por ejemplo, datos de Puntos de Venta o datos Financieros, o datos Meteorológicos, que las empresas pueden utilizar para definir la estrategia, predecir la demanda y tomar decisiones relacionadas con la distribución o las promociones de marketing, entre otras cosas. Estos conjuntos de datos suelen estar disponibles en forma de archivos planos, archivos de hojas de cálculo o documentos XML. Los archivos planos almacenan datos en formato de texto plano, con un registro o fila por línea, y cada valor separado por delimitadores como comas, punto y coma o tabuladores. Los datos de un archivo plano se asignan a una sola tabla, a diferencia de las bases de datos relacionales que contienen múltiples tablas. Uno de los formatos de archivo plano más común es el CSV, en el que los valores están separados por comas. Los archivos de hoja de cálculo son un tipo especial de archivos planos, que también organizan los datos en un formato tabular: filas y columnas. Pero una hoja de cálculo puede contener múltiples hojas de trabajo, y cada hoja de trabajo puede asignarse a una tabla diferente. Aunque los datos de las hojas de cálculo están en texto plano, los archivos pueden almacenarse en formatos personalizados e incluir información adicional como el formato, las fórmulas, etc. Microsoft Excel, que almacena datos en formato .XLS o .XLSX es probablemente la hoja de cálculo más habitual.

Reproduce el video desde :3:1 y sigue la transcripción3:01

Existen otras como Google sheets, Numbers de Apple y LibreOffice. Los archivos XML, contienen valores de datos que se identifican o marcan mediante etiquetas. Mientras que los datos de los archivos planos son "planos" o se asignan a una sola tabla, los archivos XML pueden admitir estructuras de datos más complejas, como las jerárquicas. Algunos usos comunes de XML incluyen datos de encuestas en línea, extractos bancarios y otros conjuntos de datos no estructurados.

Reproduce el video desde :3:35 y sigue la transcripción3:35

Muchos proveedores de datos y sitios web proporcionan API, o Interfaces de Programas de Aplicación, y Servicios Web, con los que múltiples usuarios o aplicaciones pueden interactuar y obtener datos para su procesamiento o análisis. Las API y los servicios web suelen escuchar las solicitudes entrantes, que pueden venir en forma de solicitudes web de los usuarios o solicitudes de red de las aplicaciones, y devuelven los datos en texto plano, XML, HTML, JSON o archivos de medios.

Reproduce el video desde :4:11 y sigue la transcripción4:11

Veamos algunos ejemplos populares de API que se utilizan como fuente de datos para el análisis de datos: El uso de las APIs de Twitter y Facebook para obtener datos de tweets y posts para realizar tareas como la extracción de opiniones o el análisis de sentimientos, que consiste en resumir el grado de apreciación y crítica sobre un tema determinado, como las políticas de un gobierno, un producto, un servicio o la satisfacción del cliente en general. Las APIs del Mercado de Valores se utilizan para extraer datos como los precios de las acciones y los productos básicos, las ganancias por acción y los precios históricos, para el comercio y el análisis. Las APIs de Búsqueda y Validación de Datos, que pueden ser muy útiles para que los Analistas de Datos limpien y preparen los datos, así como para cotejar los datos; por ejemplo, para comprobar a qué ciudad o estado pertenece un código postal. Las API también se utilizan para extraer datos de bases de datos, tanto internas como externas a la organización. La extracción de datos web ("web scraping") se utiliza para extraer datos relevantes de fuentes no estructuradas. También conocido como screen scraping y web harvesting, el web scraping permite descargar datos específicos de páginas web en base a parámetros definidos. Los extractores web pueden, entre otras cosas, extraer texto, información de contacto, imágenes, vídeos, artículos de productos y mucho más de un sitio web. Algunos usos comunes del web scraping son: recolectar detalles de productos de minoristas, fabricantes y sitios web de comercio electrónico para proporcionar comparaciones de precios, generar oportunidades de venta a través de fuentes de datos públicas, extraer datos de publicaciones y autores en diversos foros y comunidades, y recolectar conjuntos de datos de entrenamiento y prueba para modelos de machine learning. Algunas de las herramientas de scraping web más populares son BeautifulSoup, Scrapy, Pandas y Selenium. Los flujos de datos son otra fuente ampliamente utilizada para agregar flujos constantes de datos que fluyen de fuentes como instrumentos, dispositivos y aplicaciones IoT, datos GPS de automóviles, programas informáticos, sitios web y publicaciones en redes sociales. Estos datos suelen tener una marca de tiempo y también están etiquetados geográficamente con su localización. Algunas de los flujos de datos y las formas en que se pueden aprovechar incluyen: dicadores bursátiles y de mercado para el comercio financiero, flujos de transacciones de comercio minorista para predecir la demanda y la gestión de la cadena de suministro, flujos de vídeo y vigilancia para la detección de amenazas, canales de redes sociales para el análisis de sentimientos, flujos de datos de sensores para la supervisión de maquinaria industrial o agrícola, flujos de clics en la web para supervisar el rendimiento de la web y mejorar el diseño, y eventos de vuelos en tiempo real para hacer nuevas reservas y reprogramarlas. Algunas aplicaciones populares utilizadas para procesar flujos de datos son Apache Kafka, Apache Spark Streaming y Apache Storm. Los canales RSS (o Really Simple Syndication), son otra fuente de datos popular. Normalmente se utilizan para captar datos actualizados de foros en línea y sitios de noticias en los que se actualizan los datos de forma continua. Mediante un lector de feeds, que es una interfaz que convierte los archivos de texto RSS en un flujo de datos actualizados, las actualizaciones se transmiten a los dispositivos del usuario.

Lenguajes para los Profesionales de los Datos

En este video, vamos a aprender sobre algunos de los lenguajes relevantes en el trabajo de los profesionales de los datos. Se pueden clasificar como: lenguajes de consulta, lenguajes de programación y secuencias de comandos de shell. El dominio de al menos un lenguaje en cada categoría es esencial para cualquier profesional de los datos. En pocas palabras: Los lenguajes de consulta están diseñados para acceder y manipular datos en una base de datos; por ejemplo, SQL, los lenguajes de programación están diseñados para desarrollar aplicaciones y controlar el comportamiento de las mismas; por ejemplo, Python, R y Java; y los lenguajes de shell y de secuencias de comandos, como Unix/Linux Shell y PowerShell, son ideales para tareas operativas repetitivas y que llevan mucho tiempo. En el resto del vídeo, examinaremos estos lenguajes con mayor profundidad. SQL, o SQLLAGO Estructurado, es un lenguaje de consulta diseñado para acceder y manipular información de, principalmente, aunque no exclusivamente, bases de datos relacionales. Utilizando SQL, podemos escribir un conjunto de instrucciones para realizar operaciones como Insertar, actualizar y eliminar registros en una base de datos; Crear nuevas bases de datos, tablas y vistas; y Escribir procedimientos almacenados, lo que significa que se puede escribir un conjunto de instrucciones y llamarlas para utilizarlas posteriormente. He aquí algunas ventajas del uso de SQL:

Reproduce el video desde :1:40 y sigue la transcripción1:40

SQL es portable y se puede usar independientemente de la plataforma, puede usarse para consultar datos en una amplia variedad de bases de datos y repositorios de datos, aunque cada fabricante puede introducir algunas variaciones y extensiones especiales, tiene una sintaxis simple similar a la del idioma inglés, su sintaxis permite que los desarrolladores escriban programas con menos líneas que con algunos otros lenguajes de programación, usando palabras clave básicas como select, insert, into, y update, Puede extraer grandes cantidades de datos de forma rápida y eficiente, se ejecuta sobre un intérprete, lo que significa que el código se puede ejecutar tan pronto como se escribe, haciendo que la creación de prototipos sea rápida y fácil. SQL es uno de los lenguajes de consulta más populares. Gracias a su gran comunidad de usuarios y al enorme volumen de documentación acumulada a lo largo de los años, sigue proporcionando una plataforma uniforme, en todo el mundo, a todos sus usuarios. Python es un lenguaje de programación muy utilizado, de código abierto, de uso general y de alto nivel. Su sintaxis permite que los programadores expresen sus conceptos en menos líneas de código, en comparación con algunos de los lenguajes más antiguos. Python se percibe como uno de los lenguajes más fáciles de aprender y cuenta con una gran comunidad de desarrolladores. Por su énfasis en la simplicidad y la legibilidad, y su baja curva de aprendizaje, es una herramienta ideal para programadores principiantes. Es genial para realizar tareas de alta computación sobre grandes cantidades de datos, que de otra manera pueden ser extremadamente largas y engorrosas. Python proporciona librerías como Numpy y Pandas, que facilitan esta tarea gracias al procesamiento en paralelo. Tiene funciones integradas para casi todos los conceptos de uso frecuente. Python soporta múltiples paradigmas de programación, como el orientado a objetos, el imperativo, el funcional y el procedimental, lo que lo hace adecuado para una amplia variedad de casos de uso. Ahora veamos algunas de las razones que hacen de Python uno de los lenguajes de programación de más rápido crecimiento en el mundo actualmente. Es fácil de aprender - Con Python, tienes la ventaja de usar menos líneas de código para realizar tareas en comparación con otros lenguajes. Es de código abierto - Python es gratuito y utiliza un modelo de desarrollo basado en la comunidad. Funciona en entornos Windows y Linux, y puede portarse a múltiples plataformas. Tiene un amplio soporte por parte de la comunidad, y dispone de muchas librerías de análisis útiles. Tiene varias bibliotecas de código abierto para la manipulación de datos, visualización de datos, estadística y matemáticas, por nombrar sólo algunas. Su vasta gama de bibliotecas y funcionalidades también incluyen: Pandas para la limpieza y análisis de datos, Numpy y Scipy, para el análisis estadístico, Beautifulsoup y Scrapy para web scraping, Matplotlib y Seaborn para representar visualmente los datos en forma de gráficos de barras, histogramas y gráficas de tarta Opencv para el procesamiento de imágenes.

Reproduce el video desde :5: y sigue la transcripción5:00

R es un lenguaje y entorno de programación de código abierto para el análisis de datos, visualización de datos, machine learning y estadísticas. Ampliamente utilizado para desarrollar software estadístico y realizar análisis de datos, es especialmente conocido por su capacidad para crear visualizaciones convincentes, lo que le da una ventaja sobre algunos de los otros lenguajes en esta área. Algunos de los beneficios clave de R son los siguientes: Es un lenguaje de programación de código abierto independiente de la plataforma, puede emparejarse con muchos lenguajes de programación, incluido Python, es muy extensible, lo que significa que los desarrolladores pueden seguir añadiendo funcionalidades definiendo nuevas funciones, facilita el manejo de datos estructurados y no estructurados, lo que significa que tiene una capacidad sobre los datos más completa, tiene bibliotecas como Ggplot2 y Plotly que ofrecen gráficos atractivos a los usuarios, puede hacer informes con los datos y secuencias de comandos integrados en ellas; también, aplicaciones web interactivas que permiten que los usuarios jueguen con los resultados y los datos, Predomina entre otros lenguajes de programación en el desarrollo de herramientas estadísticas. Java es un lenguaje de programación orientado a objetos, basado en clases e independiente de plataformas, desarrollado originalmente por Sun Microsystems. Se encuentra entre los lenguajes de programación mejor valorados que se utilizan hoy en día. Java se utiliza en diversos procesos relacionados con el análisis de datos, incluyendo la limpieza de datos, importación y exportación de datos, análisis estadístico y visualización de datos. De hecho, la mayoría de los frameworks y herramientas populares que se usan para big data están habitualmente escritos en Java, como Hadoop, Hive y Spark. Es perfectamente adecuado para proyectos de velocidad crítica.

Reproduce el video desde :6:59 y sigue la transcripción6:59

Un Shell de Unix/Linux es un programa informático escrito para el shell de UNIX. Es una serie de comandos UNIX en un archivo de texto plano destinados a realizar una tarea específica. Escribir una secuencia de comandos de shell es rápido y fácil. Es muy útil para tareas repetitivas que pueden llevar mucho tiempo de ejecución escribiendo una línea cada vez. Las operaciones típicas que realizan las secuencias de comandos de shell incluyen: manipulación de archivos, ejecución de programas, tareas de administración del sistema como copias de seguridad del disco y evaluación de registros del sistema, secuencias de comandos de instalación para programas complejos, ejecución de copias de seguridad rutinarias, ejecución por lotes, PowerShell es una herramienta de automatización multiplataforma y un framework de configuración de Microsoft optimizado para trabajar con formatos de datos estructurados, como JSON, CSV, XML y API REST, sitios web y aplicaciones ofimáticas. Consiste en un shell de línea de comandos y un lenguaje de secuencias de comandos. PowerShell está basado en objetos, lo que permite filtrar, clasificar, medir, agrupar, comparar y muchas más acciones sobre los objetos a medida que pasan por una cadena de procesado de datos. También es una buena herramienta para la extracción de datos, la creación de interfaces gráficas de usuario y la creación de gráficas, cuadros de mando e informes interactivos.

Resumen y Aspectos Resultants

En esta lección, has aprendido la siguiente información:

Un ecosistema de analistas de datos incluye la infraestructura, el software, las herramientas, los marcos y los procesos utilizados para recopilar, limpiar, analizar extraer y visualizar datos.

Basándose en lo bien definida que está la estructura de los datos, los datos pueden ser categorizados como:

● Datos Estructurados, es decir, datos que están bien organizados en formatos que pueden ser almacenados en bases de datos.

● Datos Semi-Estructurados, es decir, datos que están bien parcialmente organizados y parcialmente en forma libre.

● Datos No Estructurados, es decir, datos que no pueden organizarse convencionalmente en filas y columnas.

Los datos se presentan en una amplia variedad de formatos de archivo, como archivos de texto delimitados, hojas de cálculo, XML, PDF, y JSON, cada uno con su propia lista de ventajas y limitaciones.

Los datos se extraen de múltiples fuentes de datos, que van desde las bases de datos relacionales y no relacionales hasta las API, los servicios web, los flujos de datos, las plataformas sociales y los dispositivos de sensores.

Una vez que los datos se identifican y se reúnen de diferentes fuentes, es necesario ponerlos en un repositorio de datos para poder prepararlos para el análisis. El tipo, el formato y las fuentes de los datos influyen en el tipo de repositorio de datos que puede utilizarse.

Los profesionales de los datos necesitan una serie de lenguajes que les ayuden a extraer, preparar y analizar los datos. Estos pueden clasificarse como:

● Lenguajes de Consulta, como el SQL, utilizado para acceder y manipular los datos de bases de datos.

● Lenguajes de Programación, como Python, R, y Java, para desarrollar aplicaciones y controlar el comportamiento de las mismas.

● Lenguajes de Shell y Scripting, como Unix/Linux Shell, y PowerShell, para automatizar tareas operacionales repetitivas.DAPara qué se utilizan más comúnmente los lenguajes de shell y scripts

# Visión General de los Repositorios de Datos

Un repositorio de datos es un término general que se utiliza para referirse a los datos que se han recopilado, organizado y aislado de manera que puedan utilizarse para operaciones empresariales o explotarse para la presentación de informes y el análisis de datos. Puede ser una infraestructura de base de datos pequeña o grande, con una o más bases de datos que reúnan, administren y almacenen conjuntos de datos. En este vídeo, vamos a proporcionar una visión general de los diferentes tipos de repositorios en los que pueden residir los datos, tales como bases de datos, almacenes de datos y almacenes de big data, y los examinaremos con mayor detalle en otros vídeos. Comencemos con las bases de datos. Una base de datos es una colección de datos, o información, diseñada para su introducción, almacenamiento, búsqueda y consulta, y modificación de datos. Y un Sistema de Administración de Bases de Datos, o DBMS, es un conjunto de programas que crea y mantiene la base de datos. Permite almacenar, modificar y extraer información de la base de datos mediante una función llamada consulta. Por ejemplo, si quieres encontrar clientes que han estado inactivos durante seis meses o más, utilizando la función de consulta, el sistema de gestión de base de datos recuperará los datos de todos los clientes de la base de datos que han estado inactivos durante seis meses o más. Aunque base de datos y SGBD signifiquen cosas diferentes, los términos se suelen utilizar indistintamente. Existen diferentes tipos de bases de datos. Varios factores influyen en la elección de la base de datos, como el tipo y la estructura de los datos, mecanismos de consulta, requisitos de latencia, velocidades de transacción y uso previsto de los datos. Es importante mencionar aquí dos tipos principales de bases de datos: las bases de datos relacionales y las no relacionales. Las bases de datos relacionales, también denominadas RDBMS, se basan en los principios organizativos de los archivos planos, con datos organizados en un formato tabular con filas y columnas siguiendo una estructura y un esquema bien definidos. Sin embargo, a diferencia de los archivos planos, las RDBMS están optimizadas para operaciones y consultas de datos que implican muchas tablas y volúmenes de datos mucho mayores. El Lenguaje de Consulta Estructurado, o SQL, es el lenguaje de consulta estándar para las bases de datos relacionales. Luego tenemos las bases de datos no relacionales, también conocidas como NoSQL, o "No sólo SQL". Las bases de datos no relacionales surgieron en respuesta al volumen, diversidad y velocidad con que se generan los datos hoy en día, principalmente influenciadas por los avances en la computación en la nube, la Internet de las Cosas y la proliferación de las redes sociales. Construidas para ofrecer velocidad, flexibilidad y escala, las bases de datos no relacionales permitieron almacenar datos sin esquemas o de forma libre. NoSQL se utiliza ampliamente para procesar big data. Un data warehouse funciona como un repositorio central que fusiona la información procedente de fuentes dispares y la consolida mediante el proceso de extracción, transformación y carga, también conocido como proceso ETL, en una base de datos integral para el análisis y la inteligencia de negocios. A un nivel muy alto, el proceso ETL ayuda a extraer datos de diferentes fuentes de datos, transformar los datos a un estado limpio y utilizable, y cargar los datos en el repositorio de datos de la empresa. Relacionados con los Data Warehouses están los conceptos de Data Marts y Data Lakes, que cubriremos más adelante. Los Data Marts y los Data Warehouses han sido históricamente relacionales, ya que muchos de los datos tradicionales de la empresa han residido en las RDBMS. Sin embargo, con la aparición de las tecnologías NoSQL y las nuevas fuentes de datos, los repositorios de datos no relacionales también se están utilizando ahora para Almacenamiento de Datos. Otra categoría de repositorios de datos son los Almacenes de Big Data, que incluyen una infraestructura distribuida de computación y almacenamiento para almacenar, escalar y procesar conjuntos de datos muy grandes. En general, los repositorios de datos ayudan a aislar los datos y a hacer que la presentación de informes y el análisis sea más eficiente y creíble, a la vez que sirven como archivo de datos.

RDBMS (Sistema de Gestión de Bases de Datos Relacionales)

Una base de datos relacional es un conjunto de datos organizados en una estructura de tablas, en la que las tablas pueden vincularse, o relacionarse, sobre la base de los datos comunes a cada una de ellas. Las tablas están formadas por filas y columnas, en las que las filas son los "registros" y las columnas los "atributos". Tomemos el ejemplo de una tabla de clientes que contiene datos sobre cada cliente de una empresa. Las columnas, o atributos, de la tabla de clientes son el ID de la Empresa, el Nombre de la Empresa, la Dirección de la Empresa y el Teléfono Principal de la Empresa; y Cada fila es el registro de un cliente. Ahora entendamos lo que queremos decir con que las tablas están vinculadas, o relacionadas, en base a los datos comunes a cada una. Junto con la tabla de clientes, la empresa también mantiene tablas de transacciones que contienen datos que describen múltiples transacciones individuales pertenecientes a cada cliente. Las columnas de la tabla de transacciones pueden incluir la Fecha de la Transacción, el ID del Cliente, el Importe de la Transacción y el Método de Pago. La tabla de clientes y las tablas de transacciones pueden estar relacionadas en base al campo común ID de Cliente. Puedes realizar una consulta en la tabla de clientes para generar informes tales como un extracto para un cliente que consolide todas las transacciones en un período determinado. Esta capacidad de relacionar tablas basadas en datos comunes permite obtener una tabla completamente nueva a partir de los datos de una o más tablas con una sola consulta. También permite comprender las relaciones entre todos los datos disponibles y obtener nuevos conocimientos para tomar mejores decisiones. Las bases de datos relacionales utilizan un lenguaje de consulta estructurado, o SQL, para consultar los datos. Aprenderemos más sobre SQL más adelante en este curso. Las bases de datos relacionales se basan en los principios organizativos de los archivos planos tales como las hojas de cálculo, con los datos organizados en filas y columnas siguiendo una estructura y un esquema bien definidos. Pero aquí es donde termina la similitud. Las bases de datos relacionales, por su diseño, son ideales para el almacenamiento, obtención y procesamiento optimizado de datos para grandes volúmenes de datos, a diferencia de las hojas de cálculo, que tienen un número limitado de filas y columnas. Cada tabla de una base de datos relacional tiene un conjunto único de filas y columnas, y se pueden definir relaciones entre las tablas, lo que minimiza la redundancia de los datos. Además, se pueden restringir los campos de la base de datos a tipos y valores de datos específicos, lo que minimiza las irregularidades y da lugar a una mayor coherencia e integridad de los datos. Las bases de datos relacionales utilizan SQL para consultar datos, lo que le da la ventaja de procesar millones de registros y recuperar grandes cantidades de datos en cuestión de segundos. Además, la arquitectura de seguridad de las bases de datos relacionales proporciona un acceso controlado a los datos y también garantiza que se puedan aplicar las normas y políticas que rigen los datos. Las bases de datos relacionales van desde pequeños sistemas de escritorio hasta sistemas masivos basados en la nube. Pueden ser: de código abierto y con soporte interno, de código abierto con soporte comercial o sistemas comerciales de código cerrado. IBM DB2, Microsoft SQL Server, MySQL, Oracle Database, y PostgreSQL son algunas de las bases de datos relacionales más populares. Las bases de datos relacionales basadas en la nube, también conocidas como Base de Datos como Servicio ("DBaas"), están consiguiendo un amplio uso ya que tienen acceso a las ilimitadas capacidades de computación y almacenamiento que ofrece la nube. Some of the popular cloud relational databases include Amazon Relational Database Service (RDS), Google Cloud SQL, IBM DB2 on Cloud, Oracle Cloud y SQL Azure. RDBMS es una tecnología madura y bien documentada, que facilita el aprendizaje y la búsqueda de talento cualificado. Una de las ventajas más significativas del enfoque de la base de datos relacional es su capacidad de crear información significativa al unir tablas. Algunas de sus otras ventajas incluyen: Flexibilidad: Usando SQL, se pueden añadir nuevas columnas, añadir nuevas tablas, renombrar relaciones y hacer otros cambios mientras la base de datos está en funcionamiento y las consultas se están ejecutando. Reducción de la redundancia: Las bases de datos relacionales minimizan la redundancia de datos. Por ejemplo, la información de un cliente aparece en una sola entrada en la tabla de clientes, y la tabla de transacciones correspondiente al cliente almacena un enlace a la tabla de clientes. Facilidad de copia de seguridad y recuperación de desastres: Las bases de datos relacionales ofrecen opciones de exportación e importación sencillas, lo que facilita las copias de seguridad y la restauración. Las exportaciones pueden realizarse mientras la base de datos está en funcionamiento, lo que facilita la restauración en caso de fallo. Las bases de datos relacionales basadas en la nube realizan una duplicación continua, lo que significa que la pérdida de datos durante una restauración puede medirse en segundos o menos. Cumplimiento ACID: ACID significa Atomicidad, Consistencia, Aislamiento y Durabilidad. Y el cumplimiento del ACID implica que los datos de la base de datos siguen siendo precisos y consistentes a pesar de los fallos, y que las transacciones de la base de datos se procesan de forma fiable. Ahora veremos algunos casos de uso de bases de datos relacionales: Procesamiento de Transacciones Online (OLTP): Las aplicaciones OLTP se centran en tareas orientadas a las transacciones que se ejecutan a gran ritmo. Las bases de datos relacionales se adaptan bien a las aplicaciones OLTP porque pueden dar cabida a un gran número de usuarios; admiten la capacidad de insertar, actualizar o eliminar pequeñas cantidades de datos; y también admiten consultas y actualizaciones frecuentes, así como tiempos de respuesta rápidos. Data Warehouses: En un entorno de almacenamiento de datos, las bases de datos relacionales pueden optimizarse para el procesamiento analítico online (u OLAP), en el que se analiza el histórico de los datos para la inteligencia de negocios. Soluciones IoT: Las soluciones de la Internet de las Cosas (IoT) requieren velocidad así como la capacidad de recopilar y procesar datos de los dispositivos edge, que necesitan una solución de base de datos ligera. Esto nos lleva a las limitaciones de la RDBMS: la RDBMS no funciona bien con datos semiestructurados y no estructurados y, por lo tanto, no es adecuada para un análisis exhaustivo de dichos datos. Para la migración entre dos RDBMS, los esquemas y el tipo de datos tienen que ser idénticos entre las tablas de origen y destino. Las bases de datos relacionales tienen un límite en la longitud de los campos de datos, lo que significa que si se intenta introducir en un campo más información de la que puede contener, la información no se almacenará. A pesar de las limitaciones y la evolución de los datos en estos tiempos de big data, computación en la nube, dispositivos IoT y redes sociales, la RDBMS sigue siendo la tecnología predominante para trabajar con datos estructurados.

# NoSQL

NoSQL, que significa "no sólo SQL", o a veces "no SQL" es un diseño de base de datos no relacional que proporciona esquemas flexibles para el almacenamiento y la recuperación de datos. Las bases de datos NoSQL existen desde hace muchos años, pero sólo se han hecho más populares recientemente en la era de la nube, el big data y las aplicaciones web y móviles de gran volumen. Hoy en día se escogen por sus atributos en torno a la escala, el rendimiento y la facilidad de uso. Es importante destacar que el "No" en "NoSQL" es una abreviatura de "no sólo" y no la palabra real "No". Las bases de datos NoSQL están creadas para modelos de datos específicos y tienen esquemas flexibles que permiten que los programadores creen y manejen aplicaciones modernas. No utilizan un diseño tradicional de base de datos de fila/columna/tabla con esquemas fijos y, por lo general, no utilizan el lenguaje de consulta estructurado (o SQL) para consultar los datos, aunque algunos pueden admitir interfaces de tipo SQL o similares. El NoSQL permite almacenar datos sin esquema o de forma libre. Cualquier dato, ya sea estructurado, semiestructurado o no estructurado, puede almacenarse en cualquier registro. Según el modelo que se utiliza para almacenar los datos, hay cuatro tipos comunes de bases de datos NoSQL. Almacén de clave-valor, basada en documentos, basada en columnas y basada en gráficos. Almacén de clave-valor. Los datos de una base de datos de clave-valor se almacenan como una colección de pares de clave-valor. La clave representa un atributo de los datos y es un identificador único. Tanto las claves como los valores pueden ser cualquier cosa, desde simples enteros o cadenas hasta complejos documentos JSON. Los almacenes de clave-valor son excelentes para guardar los datos de la sesión y las preferencias de los usuarios, hacer recomendaciones en tiempo real y publicidad dirigida, y el almacenamiento en memoria de datos. Sin embargo, si se quiere poder consultar los datos sobre un valor de datos específico, se necesitan relaciones entre los valores de los datos, o se necesita tener múltiples claves únicas, un almacén de claves-valores puede no ser lo más adecuado. Redis, Memcached y DynamoDB son algunos ejemplos bien conocidos en esta categoría. Basada en documentos: Las bases de datos de documentos almacenan cada registro y sus SUBCONJUNTO DE DATOS

. Permiten una indización flexible, potentes consultas ad hoc y análisis de las colecciones de documentos. Las bases de datos de documentos son más adecuadas para las plataformas de comercio electrónico, almacenamiento de registros médicos, plataformas de CRM y plataformas de análisis. Sin embargo, si lo que se desea es ejecutar consultas de búsqueda complejas y transacciones de múltiples operaciones, es posible que una base de datos de documentos no sea la mejor opción. MongoDB, DocumentDB, CouchDB y Cloudant son algunas de las bases de datos basadas en documentos más populares.

Reproduce el video desde :3:25 y sigue la transcripción3:25

Basada en columnas: Los modelos basados en columnas almacenan los datos en celdas agrupadas como columnas de datos en lugar de en filas. La agrupación lógica de las columnas, es decir, las columnas a las que se suele acceder juntas, se denomina familia de columnas. Por ejemplo, lo más probable es que se acceda juntos al nombre y a la información del perfil de un cliente, pero no a su historial de compras. Por lo tanto, los datos de información del nombre y el perfil del cliente pueden agruparse en una familia de columnas. Dado que las bases de datos de columnas almacenan todas las celdas correspondientes a una columna como una entrada continua en disco, el acceso y la búsqueda de los datos se hace muy rápido. Las bases de datos de columnas pueden ser excelentes para los sistemas que requieren intensas solicitudes de escritura, almacenando datos de series temporales, datos meteorológicos y datos IoT. Pero si tienes que usar consultas complejas o cambiar tus patrones de consulta con frecuencia, puede que no sea la mejor opción para ti. Las bases de datos de columnas más populares son Cassandra y HBase. Basadas en gráficos: Las bases de datos basadas en gráficos utilizan un modelo gráfico para representar y almacenar datos. Son particularmente útiles para visualizar, analizar y encontrar conexiones entre diferentes fragmentos de datos. Los círculos son nodos, y contienen los datos. Las flechas representan relaciones. Las bases de datos gráficas son una excelente opción para trabajar con datos conectados, que son datos que contienen muchas relaciones interconectadas. Las bases de datos gráficas son excelentes para las redes sociales, las recomendaciones de productos en tiempo real, los diagramas de red, la detección de fraudes y la gestión de acceso. Pero si quieres procesar grandes volúmenes de transacciones, puede que no sea la mejor opción porque las bases de datos de gráficos no están optimizadas para consultas analíticas de gran volumen. Neo4J y CosmosDB son algunas de las bases de datos gráficas más populares. NoSQL se creó en respuesta a las limitaciones de la ACIDtradicional de bases de datos relacionales. La principal ventaja de NoSQL es su capacidad para manejar grandes volúmenes de datos estructurados, semiestructurados y no estructurados. Algunas de sus otras ventajas incluyen: La capacidad de ejecutarse como sistemas distribuidos escalados en múltiples centros de datos, lo que les permite aprovechar la infraestructura de computación en la nube; una arquitectura de escalado eficiente y rentable que proporciona capacidad y rendimiento adicionales con la adición de nuevos nodos; y un diseño más sencillo, un mejor control de la disponibilidad y una mejor escalabilidad que te permite ser más ágil, más flexible e iterar más rápidamente. Para resumir las principales diferencias entre las bases de datos relacionales y no relacionales: Los esquemas RDBMS definen rígidamente la forma en que todos los datos insertados en la base de datos han de ser tipificados y conformados, mientras que las bases de datos NoSQL pueden no tener preferencia por un esquema, lo que permite almacenar y manipular datos no estructurados y semiestructurados. El mantenimiento de sistemas de gestión de bases de datos relacionales comerciales de alta gama es costoso, mientras que las bases de datos NoSQL están diseñadas específicamente para hardware de bajo coste. Las bases de datos relacionales, a diferencia de la mayoría de las NoSQL, soportan el cumplimiento ACID, lo que garantiza la fiabilidad de las transacciones y la recuperación de desastres. RDBMS es una tecnología madura y bien documentada, lo que significa que los riesgos son más o menos perceptibles en comparación con NoSQL, que es una tecnología relativamente más reciente. No obstante, las bases de datos NoSQL han llegado para quedarse, y se utilizan cada vez más para aplicaciones críticas.

# Data Marts, Lago de Datos, ETL y Tuberías de Datos

Anteriormente en el curso, hemos examinado las bases de datos, los data warehouses, y los almacenes de big data. Ahora profundizaremos un poco más en nuestra exploración de data warehouses, data marts y data lakes; y también aprenderemos sobre el proceso ETL y las cadenas de procesado de datos ("data pipelines"). Un data warehouse funciona como un almacenamiento multipropósito para diferentes casos de uso. Cuando los datos llegan al warehouse ("almacén"), ya han sido modelados y estructurados para un propósito específico, lo que significa que están listos para su análisis. Como organización, se optaría por un data warehouse cuando se tienen cantidades masivas de datos de los sistemas operacionales que necesitan estar disponibles para la presentación de informes y su análisis.

Los data warehouses sirven como la única fuente de verdad, almacenando datos actuales e históricos que ya han sido limpiados, conformados y categorizados. Un data warehouse es un instrumento multipropósito de análisis operacional y de rendimiento. Un data mart es una subsección del data warehouse, creado específicamente para una función de negocio concreta, propósito o comunidad de usuarios. La idea es proporcionar a los interesados los datos más pertinentes para ellos, cuando los necesiten. Por ejemplo, los equipos de ventas o financieros que acceden a los datos para elaborar sus informes y proyecciones trimestrales. Dado que un data mart ofrece capacidades analíticas para un área restringida del data warehouse, ofrece seguridad aislada y rendimiento aislado. El papel más importante de un data mart es la generación de informes y análisis específicos de la empresa. Un data lake es un repositorio de almacenamiento que puede albergar grandes cantidades de datos estructurados, semiestructurados y no estructurados en su formato nativo, clasificados y etiquetados con metadatos. Así, mientras que un data warehouse almacena datos procesados para una necesidad específica, un data lake es un conjunto de datos en bruto donde cada elemento de datos recibe un identificador único y es etiquetado con metaetiquetas para su uso posterior. e optaría por un data lake si se generan, o se tiene acceso a, grandes volúmenes de datos de forma continua, pero no se quiere restringir a casos de uso específicos o predefinidos. A diferencia de los data warehouses, un data lake retendría todos los datos de la fuente, sin ninguna exclusión. Y los datos podrían incluir todo tipo de fuentes y tipos de datos. Los data lakes a veces también se utilizan como área para la preparación de un data warehouse. El papel más importante de un data lake es el análisis predictivo y avanzado. Ahora llegamos al proceso que constituye el núcleo de la obtención de valor a partir de los datos: el proceso de Extracción, Transformación y Carga, o ETL. ETL es la forma en que los datos en bruto se convierten en datos listos para el análisis. Es un proceso automatizado en el que se recogen datos en bruto de fuentes identificadas, se extrae la información que se ajusta a las necesidades de información y análisis, se limpia, estandariza y transforma esos datos en un formato que se puede utilizar en el contexto de la organización, y se carga en un repositorio de datos. Mientras que ETL es un proceso genérico, el trabajo real puede ser muy diferente en cuanto a uso, utilidad y complejidad. La Extracción es el paso en el que se recopilan los datos desde las ubicaciones de las fuentes para su transformación. La extracción de datos podría ser a través de: Procesamiento por lotes, lo que significa que los datos de origen se mueven en grandes fragmentos desde la fuente hasta el sistema de destino a intervalos programados. Las herramientas para el procesamiento por lotes incluyen Stitch y Blendo. El procesamiento de flujos, que significa que los datos de la fuente se extraen en tiempo real de la misma y se transforman mientras están en tránsito y antes de que se carguen en el repositorio de datos. Entre las herramientas para el procesamiento de flujos figuran Apache Samza, Apache Storm y Apache Kafka. La transformación implica la ejecución de reglas y funciones que convierten los datos en bruto en datos que pueden utilizarse para el análisis. Por ejemplo, haciendo que los formatos de fecha y las unidades de medida sean consistentes en todos los datos de origen, eliminando los datos duplicados, filtrando los datos que no se necesitan, enriqueciendo los datos, por ejemplo, dividiendo el nombre completo en nombre, segundo nombre y apellidos, estableciendo relaciones clave entre las tablas, aplicando reglas de negocio y validaciones de datos. La carga es el paso en el que los datos procesados se transportan a un sistema de destino o a un repositorio de datos. Podría ser: Carga inicial, es decir, rellenar todos los datos del repositorio, Carga incremental, es decir, aplicar periódicamente actualizaciones y modificaciones continuas según sea necesario; o Refresco completo, es decir, borrar el contenido de una o más tablas y volver a cargarlas con datos actualizados. La verificación de la carga, que incluye la comprobación de los datos en busca de valores faltantes o nulos, el rendimiento del servidor y la supervisión de los fallos de carga, son partes importantes de este paso del proceso. Es vital vigilar los fallos de carga y asegurarse de que se dispone de los mecanismos de recuperación adecuados. ETL se ha utilizado históricamente para cargas de trabajo por lotes a gran escala. Sin embargo, con la aparición de las herramientas ETL de streaming, se están utilizando cada vez más también para flujos en tiempo real para datos de eventos. Es habitual ver que los términos ETL y data pipelines se usan indistintamente. Y aunque ambos mueven datos desde el origen hasta el destino, data pipeline es un término más amplio que abarca todo el viaje de mover datos de un sistema a otro, en el que ETL es un subconjunto. Los data pipelines pueden diseñarse para el procesamiento por lotes, para la transmisión de datos, y una combinación de datos por lotes y en streaming. En el caso de los datos en streaming, el procesamiento o transformación de datos ocurre en un flujo continuo. Esto es particularmente útil para datos que necesitan una actualización constante, como los datos de un sensor que monitorea el tráfico. Un data pipeline es un sistema de alto rendimiento que soporta tanto consultas por lotes de larga duración como consultas interactivas más pequeñas. El destino de un data pipeline suele ser un data lake, aunque los datos también pueden cargarse en diferentes destinos, como en otra aplicación o en una herramienta de visualización. Existen varias soluciones disponibles para data pipelines, siendo las más populares Apache Beam y DataFlow.

# Fundamentos de Big Data

En este mundo digital, todo el mundo deja un rastro. Desde nuestros hábitos de viaje hasta nuestros entrenamientos y entretenimiento, el creciente número de dispositivos conectados a Internet con los que interactuamos diariamente registran grandes cantidades de datos sobre nosotros, incluso hay un nombre para ello: Big Data. Ernst and Young ofrece la siguiente definición: big data se refiere a los volúmenes dinámicos, grandes y dispares de datos creados por personas, herramientas y máquinas. Requiere una tecnología nueva, innovadora y escalable para recopilar, alojar y procesar analíticamente la enorme cantidad de datos recopilados con el fin de obtener percepciones del negocio en tiempo real que se relacionen con consumidores, riesgo, beneficios, rendimiento, gestión de la productividad y un mayor valor para el accionista. No existe una definición única de big data, pero hay ciertos elementos que son comunes a las diferentes definiciones, como velocidad, volumen, variedad, veracidad y valor. Estas son las V del big data. La velocidad es la rapidez con la que se acumulan los datos. Los datos se generan extremadamente rápido en un proceso que nunca se detiene. Las tecnologías de streaming en tiempo real o cercano a él, las locales y las basadas en la nube pueden procesar la información muy rápidamente. El volumen es la escala de los datos o el aumento de la cantidad de datos almacenados. Los motores del volumen son el aumento de los orígenes 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 las bases de datos relacionales, mientras que los datos no estructurados no están organizados de forma predefinida, como los tweets, entradas de blog, imágenes, números y vídeo. La variedad también refleja que los datos provienen de diferentes fuentes; máquinas, personas y procesos, tanto internos como externos a las organizaciones. Los generadores son las tecnologías móviles, redes sociales, tecnologías vestibles ("wearables"), geotecnologías, video, y muchas, muchas más. La veracidad es la calidad y el origen de los datos y su conformidad con los hechos y la exactitud. Los atributos incluyen la consistencia, completitud, integridad y ambigüedad. Los factores determinantes incluyen el coste y la necesidad de trazabilidad. Con la gran cantidad de datos disponibles, el debate se centra en la exactitud de los datos en la era digital. ¿Es la información real o es falsa? Valor es nuestra capacidad y necesidad de convertir los datos en valor. Valor no es sólo ganancias. Puede tener beneficios médicos o sociales, así como satisfacción del cliente, del empleado o personal. La principal razón por la que la gente invierte tiempo en entender el big data es para obtener valor de él. Veamos algunos ejemplos de las Vs en acción. Velocidad. Cada 60 segundos, se suben a YouTube horas de material, que está generando datos. Piensa en la rapidez con la que los datos se acumulan durante horas, días y años. Volumen. La población mundial es de aproximadamente 7.000 millones de personas y la gran mayoría está usando ahora dispositivos digitales. Teléfonos móviles, equipos de escritorio y portátiles, dispositivos vestibles, etc. Todos estos dispositivos generan, capturan y almacenan datos de aproximadamente 2,5 quintillones de bytes cada día. Eso es el equivalente a 10 millones de DVDs Blu-ray. Variedad. Pensemos en los diferentes tipos de datos. Texto, imágenes, películas, sonido, datos de salud de los dispositivos vestibles, y muchos tipos diferentes de datos de los dispositivos conectados a la Internet de las cosas. Veracidad. El ochenta por ciento de los datos se consideran no estructurados y debemos idear formas de producir datos fiables y precisos. Los datos deben clasificarse, analizarse y visualizarse. Los científicos de datos, hoy en día, obtienen conocimientos a partir del big data y se enfrentan a los desafíos que presentan estos masivos conjuntos de datos. La escala de los datos que se están recogiendo significa que no es factible utilizar herramientas convencionales de análisis de datos, sin embargo, sí herramientas alternativas que aprovechen la potencia de computación distribuida pueden superar este problema. Herramientas como Apache Spark, Hadoop y su ecosistema proporcionan formas de extraer, cargar, analizar y procesar los datos mediante recursos de computación distribuidos, proporcionando nuevas percepciones y conocimientos. Esto brinda a las organizaciones más formas de conectar con sus clientes y enriquecer los servicios que ofrecen. Así que la próxima vez que te pongas tu reloj inteligente, desbloquees tu smartphone o hagas un seguimiento de tu entrenamiento, recuerda que tus datos están comenzando un viaje que podría llevarlos alrededor del mundo, pasando por un análisis de big data y volviendo a ti.

# Herramientas de Procesamiento de Big Data

Las tecnologías de procesamiento de Big Data proporcionan formas de trabajar con grandes conjuntos de datos estructurados, semiestructurados y no estructurados de manera que se pueda obtener valor a partir del big data. En algunos de los otros videos, hemos tratado tecnologías de Big Data como las bases de datos NoSQL y los Data Lakes. En este vídeo, vamos a hablar de tres tecnologías de código abierto y el papel que juegan en el análisis de big data: Apache Hadoop, Apache Hive y Apache Spark. Hadoop es una colección de herramientas que proporciona almacenamiento distribuido y procesamiento de big data. Hive es un almacén de datos para la consulta y el análisis de datos creado sobre Hadoop. Spark es un framework de análisis de datos distribuido diseñado para realizar análisis de datos complejos en tiempo real. Hadoop, un framework de código abierto basado en Java, permite el almacenamiento y procesamiento distribuido de grandes conjuntos de datos mediante clusters de computadoras. En el sistema distribuido Hadoop, un nodo es una sola computadora, y un conjunto de nodos forma un clúster. Hadoop puede escalar desde un solo nodo a cualquier número de nodos, cada uno de los cuales ofrece almacenamiento y computación local. Hadoop proporciona una solución fiable, escalable y rentable para el almacenamiento de datos sin requisitos de formato. Usando Hadoop, se puede: Incorporar formatos de datos que van surgiendo, como el audio en streaming, vídeo, sentimiento en redes sociales y datos de clickstream, junto con datos estructurados, semiestructurados y no estructurados que no se utilicen habitualmente en un data warehouse. Proporcionar acceso en tiempo real y con autoservicio a todas las partes interesadas. Optimizar y racionalizar los costes en el data warehouse empresarial consolidando los datos de toda la organización y trasladando los datos "en frío", es decir, los datos que no se utilizan con frecuencia, a un sistema basado en Hadoop. Uno de los cuatro componentes principales de Hadoop es el Sistema de Archivos Distribuído Hadoop, o HDFS, que es un sistema de almacenamiento para big data que se ejecuta sobre múltiples dispositivos de hardware conectados a través de una red. HDFS proporciona almacenamiento de big data escalable y confiable mediante la partición de archivos en múltiples nodos. Divide los archivos grandes entre varias computadoras, permitiendo el acceso a ellos en paralelo. Los cálculos pueden, por lo tanto, ejecutarse en paralelo en cada nodo donde se almacenan los datos. También replica bloques de archivos en diferentes nodos para prevenir la pérdida de datos, haciéndolos tolerantes a fallos. Entendamos esto con un ejemplo. Pongamos que tenemos un archivo que incluye números de teléfono de todas las personas de Estados Unidos; los números de las personas cuyo apellido empieza por A podrían estar almacenados en el servidor 1, B en el servidor 2, y así sucesivamente. Con Hadoop, las partes de esta agenda telefónica se almacenarían en todo el clúster. Para reconstruir toda la agenda telefónica, el programa necesitaría los bloques de cada servidor del clúster. HDFS también replica estas partes más pequeñas en dos servidores adicionales por defecto, asegurando la disponibilidad cuando un servidor falla, Además de una mayor disponibilidad, esto ofrece múltiples beneficios. Permite que el clúster Hadoop divida el trabajo en trozos más pequeños y ejecute esos trabajos en todos los servidores del clúster para una mejor escalabilidad. Por último, se obtiene el beneficio de la localización de los datos, que es el proceso de acercar la computación al nodo en el que residen los datos. Esto es fundamental cuando se trabaja con grandes conjuntos de datos porque minimiza la congestión de la red y aumenta el rendimiento. Algunos otros beneficios que se logran al utilizar el HDFS incluyen: Rápida recuperación de los fallos de hardware, porque HDFS está diseñado para detectar fallos y recuperarse automáticamente. Acceso a flujos de datos, porque HDFS soporta altas tasas de transferencia de datos. Cabida para grandes conjuntos de datos, porque HDFS puede escalar a cientos de nodos, o computadoras, en un solo clúster. Portabilidad, porque el HDFS es portable a 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 gestionar grandes archivos de conjuntos de datos que se almacenan directamente en HDFS o en otros sistemas de almacenamiento de datos como Apache HBase. Hadoop is intended for long sequential scans and, because Hive is based on Hadoop, queries tienen una latencia muy alta, lo que significa que Hive es menos apropiado para aplicaciones que requieren tiempos de respuesta muy rápidos. Además, Hive se basa en la lectura y, por lo tanto, no es adecuado para el procesamiento de transacciones que suelen implicar un alto porcentaje de operaciones de escritura. Hive es más adecuado para tareas de almacenamiento de datos como ETL, elaboración de 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, entre las que se incluyen el Análisis Interactivo, Procesamiento de Streams, Machine Learning, Integración de Datos y ETL. Aprovecha el procesamiento en memoria para aumentar significativamente la velocidad de los cálculos y volcarlos al disco sólo cuando la memoria está limitada. Spark tiene interfaces para los principales lenguajes de programación, incluyendo Java, Scala, Python, R y SQL. Puede funcionar usando su tecnología de clustering independiente, así como sobre otras infraestructuras, como Hadoop. Y puede acceder a datos procedentes de una gran variedad de fuentes de datos, incluyendo HDFS y Hive, lo que la hace muy versátil. La capacidad de procesar rápidamente flujos de datos y realizar complejos análisis en tiempo real es el caso de uso clave para Apache Spark.

Resumen y Aspectos Resaltantes

En esta lección, has aprendido la siguiente información:

Un Repositorio de Datos es un término general que se refiere a los datos que han sido recopilados, organizados y aislados para que puedan ser utilizados para la presentación de informes, análisis y también para fines de archivo.

Los diferentes tipos de Repositorios de Datos incluyen:

● Bases de Datos, que pueden ser relacionales o no relacionales, cada una de las cuales sigue un conjunto de principios organizativos, los tipos de datos que pueden almacenar y las herramientas que pueden utilizarse para consultar, organizar y recuperar datos.

● Almacenes de Datos, que consolidan los datos entrantes en un almacén integral.

● Data Marts, que son esencialmente subsecciones de un almacén de datos, construidas para aislar los datos para una función comercial o un caso de uso particular.

● Lago de Datos, que sirven como repositorios de almacenamiento para grandes cantidades de datos estructurados, semi-estructurados y no estructurados en su formato nativo.

● Almacenes de Grandes Volúmenes de Datos, que proporcionan una infraestructura distribuida de computación y almacenamiento para almacenar, escalar y procesar conjuntos de datos muy grandes.

ETL o Proceso de Extracción, Transformación y Carga, es un proceso automatizado que convierte los datos en bruto en datos listos para el análisis por:

● Extracción de datos de las fuentes.

● Transformar los datos en bruto mediante su limpieza, enriquecimiento, estandarización y validación.

● Cargar los datos procesados en un sistema de destino o en repositorio de datos.

La Tubería de Datos, que en ocasiones se utiliza de manera intercambiable con el ETL, abarca todo el viaje de traslado de los datos desde la fuente hasta el lago de datos o aplicación de destino, utilizando el proceso ETL.

Los Grandes Volúmenes de Datos se refieren a las grandes cantidades de datos que se están produciendo en cada momento de cada día, por personas, herramientas y maquinas. Solo la velocidad, el volumen y la variedad de los datos desafían las herramientas y los sistemas utilizados para los datos convencionales. Estos desafíos llevaron al surgimiento de herramientas y plataformas de procesamiento diseñadas específicamente para Los Grandes Volúmenes de Datos, como Apache Hadoop, Apache Hive y Apache Spark.

# Identificando Datos para el Análisis

En este punto, ya tienes un entendimiento del problema y del resultado deseado, sabes "Dónde estás" y "Dónde quieres estar". También tienes una métrica bien definida, sabes "Qué se medirá" y "Cómo se medirá". El siguiente paso es identificar los datos que necesitas para tu caso de uso. El proceso de identificar los datos comienza con determinar la información que quieres recopilar. En este paso se toman decisiones sobre: a) la información específica que se necesita; y b) los posibles orígenes de esos datos. Tus objetivos determinan las respuestas a estas preguntas. Tomemos el ejemplo de una empresa de productos que desea crear campañas de marketing dirigidas a un grupo de edad que es el que más compra sus productos. Su objetivo es diseñar las acciones que más atraigan a este segmento y les animen a seguir influyendo en sus amigos y compañeros para que compren estos productos. En base a este caso de uso, algunos de los datos obvios que identificarás incluyen el perfil del cliente, historial de compras, ubicación, edad, educación, profesión, ingresos y estado civil, por ejemplo. Para asegurarse de obtener una mayor comprensión de este segmento, también puedes decidir recopilar los datos de las quejas de los clientes de este segmento para comprender el tipo de problemas a los que se enfrentan, ya que esto podría desanimarles a recomendar tus productos. Para conocer el grado de satisfacción de la resolución de sus problemas, se pueden recoger las valoraciones de las encuestas del servicio de atención al cliente. Llevando esto un paso adelante, puede que quieras entender cómo estos clientes hablan de tus productos en las redes sociales y cuántos de sus contactos se involucran con ellos en estas discusiones, por ejemplo, los likes, compartidos y comentarios que reciben sus posts. El siguiente paso en el proceso es definir un plan para la recopilación de datos. Es necesario establecer un plazo para la recopilación de los datos que se han identificado. Algunos de los datos que se necesitan pueden ser necesarios de manera continua y otros a lo largo de un período de tiempo definido. Para recopilar datos de los visitantes de un sitio web, por ejemplo, puede ser necesario actualizar los números en tiempo real. Pero si estás controlando los datos de un evento específico, tienes una fecha de inicio y fin definida para la recopilación de los datos. En este paso, también puedes definir cuántos datos serían suficientes para que pudieras llegar a un análisis creíble. Es el volumen definido por el segmento, por ejemplo, todos los clientes dentro del rango de edad de 21 a 30 años; o un conjunto de datos de cien mil clientes dentro del rango de edad de 21 a 30 años. También se puede utilizar este paso para definir las dependencias, riesgos, plan de mitigación y varios otros factores de este tipo que son relevantes para tu campaña. El propósito del plan debe ser establecer la claridad que necesitas para su ejecución. El tercer paso del proceso es determinar los métodos de recopilación de datos. En este paso, determinarás los métodos de recopilación de datos que necesitas. Definirás cómo vas a recabar los datos de las fuentes de datos que has identificado, tales como sistemas internos, sitios de redes sociales o proveedores de datos de terceros. Los métodos dependerán del tipo de datos, el plazo en el que se necesitan los datos y el volumen de los mismos. Una vez que el plan y los métodos de recopilación de datos estén ultimados, podrás poner en práctica la estrategia de recopilación de datos y comenzar a obtenerlos. Se irán haciendo actualizaciones del plan sobre la marcha, ya que las condiciones evolucionan a medida que se aplica el plan sobre el terreno. Los datos que identifiques, la fuente de esos datos y las prácticas que emplees para recopilarlos tienen repercusiones en la calidad, la seguridad y la privacidad. Ninguna de estas consideraciones es de carácter puntual, sino que son pertinentes a lo largo del ciclo de vida del proceso de análisis de datos. Trabajar con datos de fuentes dispares sin tener en cuenta lo bien que cuadran respecto a la métrica de calidad puede conducir al fracaso. Para ser fiables, los datos deben estar libres de errores, ser precisos, completos, pertinentes y accesibles. Es necesario definir los rasgos de calidad, la métrica y los puntos de control para asegurar que el análisis se base en datos de calidad. También hay que estar atento a las cuestiones relativas a la gobernanza de los datos, como la seguridad, reglamentación y cumplimiento. Las políticas y procedimientos de Gobernanza de Datos se relacionan con la usabilidad, integridad y disponibilidad de los datos. Las sanciones por incumplimiento pueden ascender a millones de dólares y pueden dañar la credibilidad no sólo de tus conclusiones, sino también de tu organización. Otra consideración importante es la privacidad de los datos. Los datos que se recopilan deben cumplir los requisitos de confidencialidad, licencia de uso y cumplimiento de las normas establecidas. Es necesario planificar las verificaciones, validaciones y un seguimiento auditable. La pérdida de confianza en los datos utilizados para el análisis puede comprometer el proceso, dar lugar a hallazgos sospechosos e invitar a sanciones. La identificación de los datos correctos es un paso muy importante del proceso de análisis de datos. Si se hace correctamente, te estarás asegurando de que se pueda examinar un problema desde múltiples perspectivas y que los hallazgos sean creíbles y fiables.

# Fuentes de Datos

Las fuentes de datos pueden ser internas o externas a la organización, y pueden ser fuentes primarias, secundarias o de terceros. Veamos un par de ejemplos para analizar lo que entendemos por fuentes de datos primarias, secundarias y de terceros. El término datos primarios se refiere a la información obtenida directamente por uno mismo a partir de la fuente. Puede ser de fuentes internas como datos de la organización, CRM, RRHH o aplicaciones de flujo de trabajo. También podría incluir datos que obtienes directamente por medio de encuestas, entrevistas, charlas, observaciones y grupos de debate. Los datos secundarios se refieren a la información obtenida de fuentes existentes, como bases de datos externas, artículos de investigación, publicaciones, material de formación y búsquedas en Internet, o registros financieros disponibles como datos públicos. También podría incluir los datos reunidos externamente mediante encuestas, entrevistas, coloquios, observaciones y grupos de debate. Los datos de terceros son los que se adquieren de agregadores que reúnen datos de diversas fuentes y los combinan en conjuntos de datos completos con el único fin de venderlos. A continuación, se examinarán algunas de las diferentes fuentes de las que se podrían estar reuniendo datos. Las bases de datos pueden ser una fuente de datos primarios, secundarios y de terceros. La mayoría de las organizaciones tienen aplicaciones internas para gestionar sus procesos, flujos de trabajo y clientes. Las bases de datos externas están disponibles por suscripción o para su compra. Un número significativo de empresas han pasado o se están pasando a la nube, que se está convirtiendo cada vez más en una fuente de acceso a la información en tiempo real y a percepciones bajo demanda. La web es una fuente de datos de acceso público que está a disposición de las empresas y de los individuos para uso gratuito o comercial. La Web es una fuente rica de datos disponibles en dominio público. Pueden ser libros de texto, registros gubernamentales, documentos y artículos e son para consumo público, sitios de redes sociales y plataformas interactivas como Facebook, Twitter, Google, YouTube e Instagram. Se utilizan cada vez más para obtener datos y opiniones de los usuarios. Las empresas están utilizando estas fuentes de datos para obtener conocimientos cuantitativos y cualitativos. Clientes actuales y potenciales. Datos de sensores generados por dispositivos vestibles, edificios inteligentes, ciudades inteligentes, teléfonos inteligentes, dispositivos médicos, incluso los electrodomésticos son una fuente de datos muy utilizada. El intercambio de datos es una fuente de datos de terceros que implica el intercambio voluntario de datos entre los proveedores de datos y los consumidores de datos. Los individuos, las organizaciones y los gobiernos pueden ser a la vez proveedores y consumidores de datos. Los datos que se intercambian podrían incluir datos procedentes de aplicaciones para negocios, dispositivos de sensores, actividad de redes sociales, datos de localización o datos de comportamiento de los consumidores. Las encuestas recogen información mediante cuestionarios distribuidos a un grupo selecto de personas. Por ejemplo, midiendo el interés de los clientes existentes en gastar en una versión actualizada de un producto. Las encuestas pueden realizarse en la web o en papel. Los datos del censo también son una fuente comúnmente utilizada para recopilar tos de los hogares, como por ejemplo, datos de riqueza e ingresos, o datos de población. Las entrevistas son una fuente para reunir datos cualitativos, como las opiniones y experiencias de los participantes. Por ejemplo, una entrevista realizada para comprender los retos cotidianos a los que se enfrenta un ejecutivo de atención al cliente. Las entrevistas pueden ser telefónicas, a través de la web o cara a cara. Los estudios incluyen la vigilancia de los participantes en un entorno específico o mientras realizan una tarea concreta. Por ejemplo, observar a los usuarios que navegan por un sitio de comercio electrónico para evaluar la facilidad con que pueden encontrar productos y hacer una compra. Datos de encuestas, entrevistas y observación. Los estudios podrían estar disponibles como datos primarios, secundarios y de terceros. Las fuentes de datos nunca han sido tan dinámicas y diversas como lo son hoy en día. También están evolucionando continuamente. Complementar los datos primarios con fuentes de datos secundarios y de terceros puede ayudar a explorar problemas y soluciones en formas nuevas y significativas.

# Cómo Recopilar e Importar Datos

En este video, vamos a aprender sobre los diferentes métodos y herramientas disponibles para recopilar datos de las fuentes de datos tratadas anteriormente en el curso, tales como bases de datos, la web, datos de sensores, intercambios de datos, y otras diversas fuentes que se aprovechan para cubrir necesidades específicas de datos. También aprenderemos sobre la importación de datos a diferentes tipos de repositorios de datos. SQL, o Lenguaje de Consulta Estructurado, es un lenguaje de consulta usado para extraer información de bases de datos relacionales. SQL ofrece comandos sencillos para especificar lo que se debe recuperar de la base de datos, la tabla de la que se debe extraer, agrupar los registros con valores coincidentes, dictar la secuencia en la que se muestran los resultados de la consulta y limitar el número de resultados que puede devolver la consulta, entre otras muchas características y funcionalidades. Las bases de datos no relacionales pueden consultarse utilizando SQL o herramientas de consulta similares a SQL. Algunas bases de datos no relacionales vienen con sus propias herramientas de consulta como CQL para Cassandra y GraphQL para Neo4J. Las Interfaces de Programación de Aplicaciones (o APIs) también se utilizan comúnmente para extraer datos de diversas fuentes de datos. Las APIs se invocan desde las aplicaciones que requieren los datos y acceden a un punto final que contiene los datos. Los puntos finales pueden incluir bases de datos, servicios web y mercados de datos. Las APIs también se utilizan para la validación de datos. Por ejemplo, un analista de datos puede utilizar una API para validar direcciones postales y códigos postales. El web scraping, también conocido como screen scraping o web harvesting, se utiliza para descargar datos específicos de páginas web en función de parámetros definidos. Entre otras cosas, el web scraping se utiliza para extraer datos como texto, información de contacto, imágenes, vídeos, podcasts y artículos de productos a partir de un sitio web. Las fuentes RSS son otro recurso que suele utilizarse para recabar datos actualizados de foros online y sitios de noticias en los que se actualizan los datos de forma continua. Los flujos de datos son una fuente popular para agregar flujos constantes de datos que provienen de fuentes como instrumentos, dispositivos y aplicaciones IoT y datos de GPS de automóviles. Los flujos y las fuentes RSS también se utilizan para extraer datos de los sitios de redes sociales y las plataformas interactivas. Las plataformas de Intercambio de Datos permiten el intercambio de datos entre los proveedores de datos y los consumidores de datos. Los intercambios de datos tienen un conjunto de normas, protocolos y formatos de intercambio bien definidos y pertinentes para el intercambio de datos. Esas plataformas no sólo facilitan el intercambio de datos, sino que también garantizan el mantenimiento de la seguridad y la gobernanza. Proporcionan flujos de trabajo para la concesión de licencias de datos, anonimización y protección de la información personal, marcos jurídicos y un entorno de análisis protegido. Entre los ejemplos de plataformas de intercambio de datos populares se encuentran AWS Data Exchange, Crunchbase, Lotame y Snowflake. Se puede recurrir a muchas otras fuentes de datos para cubrir necesidades de datos específicas. Para las tendencias de marketing y el gasto en publicidad, por ejemplo, se sabe que empresas de investigación como Forrester y Business Insider proporcionan datos fiables. Las empresas de investigación y asesoramiento como Gartner y Forrester son fuentes de amplia confianza para la orientación estratégica y operacional. Del mismo modo, hay muchos nombres de confianza en el área de los datos sobre el comportamiento de los usuarios, el uso del móvil y de la web, las encuestas de mercado y los estudios demográficos. Los datos que se han identificado y recopilado de las diversas fuentes de datos necesitan ahora que se carguen o importen a un repositorio de datos antes de que puedan prepararse, extraerse y analizarse. El proceso de importación implica la combinación de datos de diferentes fuentes para proporcionar una vista combinada y una interfaz única con la que se puedan consultar y manipular los datos. Según el tipo de datos, el volumen de los mismos y el tipo de repositorio de destino, es posible que se necesiten diversas herramientas y métodos. Los repositorios de datos específicos están optimizados para determinados tipos de datos. Las bases de datos relacionales almacenan datos estructurados con un esquema bien definido. Si utilizas una base de datos relacional como sistema de destino, sólo podrás almacenar datos estructurados, como datos de sistemas OLTP, hojas de cálculo, formularios online, sensores, registros de red y web. Los datos estructurados también se pueden almacenar en NoSQL. Los datos semiestructurados son datos que tienen algunas propiedades de organización pero no un esquema rígido, como por ejemplo, datos de correos electrónicos, XML, archivos comprimidos, ejecutables binarios y protocolos TCP/IP. Los datos semi-estructurados pueden almacenarse en clústeres NoSQL. XML y JSON se utilizan habitualmente para almacenar e intercambiar datos semiestructurados. JSON es también el tipo de datos preferido para los servicios web. Los datos no estructurados son los que no tienen una estructura y no pueden organizarse en un esquema, como los datos de las páginas web, fuentes de redes sociales, imágenes, vídeos, documentos, registros de medios y encuestas. Las bases de datos NoSQL y los Data Lakes constituyen una buena opción para almacenar y manipular grandes volúmenes de datos no estructurados. Los data lakes pueden albergar todos los tipos de datos y esquemas. Las herramientas ETL y los data lakes proporcionan funciones automatizadas que facilitan el proceso de importación de datos. Herramientas como Talend e Informatica, y lenguajes de programación como Python y R, y sus bibliotecas, se utilizan ampliamente para importar datos.

# Resumen y Aspectos Resultantes

En esta lección, has aprendido:

● El proceso de identificación de los datos comienza por determinar la información que se debe reunir, que a su vez está determinada por el objetivo que se pretende alcanzar.

● Una vez identificados los datos, tu siguiente paso es identificar las fuentes de las que extraerás los datos necesarios y definir un plan de recopilación de datos. En esta etapa también pesan las decisiones relativas al plazo en que se necesita el conjunto de datos, y la cantidad de datos que bastaría para llegar a un análisis creíble.

● Las Fuentes de Datos pueden ser internas o externas a la organización y pueden ser primarias, secundarias o terciarias, dependiendo de si se obtienen los datos directamente de la fuente original, se recuperan de fuentes de datos externas disponibles o se compran a agregadores de datos.

● Algunas de las fuentes de datos de las que podrías estar reuniendo datos son las bases de datos, la web, las redes sociales, las plataformas interactivas, los dispositivos de sensores, los intercambios de datos, las encuestas y los estudios de observación.

● Los datos identificados y reunidos de las diversas fuentes de datos se combinan utilizando diversos instrumentos y métodos para proporcionar una única interfaz que permita consultar y manipular los datos.

● Los datos que se identifican, la fuente de ellos y las prácticas que se emplean para reunirlos, tiene repercusiones en la calidad, la seguridad y la privacidad, que deben considerarse en esta etapa

# ¿Qué es Data Wrangling?

Data wrangling, también conocido como data munging, es un proceso iterativo que implica la exploración, transformación y validación de los datos, y hacerlos disponibles para un análisis creíble y significativo. Comprende una serie de tareas relacionadas con la preparación de los datos en bruto para un propósito claramente definido, en que los datos en bruto en esta etapa son datos que se han compilado desde diversas fuentes de datos en un repositorio de datos. Data wrangling comprende una gama de tareas relacionadas con la preparación de los datos para su análisis. Por lo general, se trata de un proceso de cuatro pasos que implica: Descubrimiento, Transformación, Validación y Publicación. La fase de Descubrimiento, también conocida como la fase de Exploración, consiste en comprender mejor los datos con respecto a nuestro caso de uso. El objetivo es averiguar específicamente cómo se pueden limpiar, estructurar, organizar y mapear mejor los datos que tenemos para nuestro caso de uso. La siguiente fase, que es la fase de Transformación, forma el grueso del proceso de data wrangling. Involucra las tareas que se realizan para transformar los datos, tales como estructurar, normalizar, desnormalizar, limpiar y enriquecer los datos. Comencemos con la primera tarea de la transformación: la Estructuración. Esta tarea incluye acciones que cambian la forma y el esquema de los datos. Los datos de entrada pueden estar en diversos formatos. Por ejemplo, puede haber algunos datos procedentes de una base de datos relacional y algunos datos obtenidos de APIs de la Web. Para fusionarlos, deberás cambiar el formato o el esquema de los datos. Este cambio puede ser tan simple como modificar el orden de los campos dentro de un registro o conjunto de datos, o tan complejo como combinar los campos en estructuras complejas. Join y Union son las transformaciones estructurales más comunes utilizadas para combinar datos de una o más tablas. La forma en que combinan los datos es diferente. Los Join combinan columnas. Cuando dos tablas se unen con join, las columnas de la primera tabla fuente se combinan con las columnas de la segunda tabla fuente en la misma fila. Por lo tanto, cada fila de la tabla resultante contiene columnas de ambas tablas. Union combina filas. Las filas de datos de la primera tabla de origen se combinan con las filas de datos de la segunda tabla de origen en una sola tabla. Cada fila de la tabla resultante es de una tabla de origen u otra. La transformación también puede incluir la normalización y la desnormalización de los datos. La normalización se centra en la limpieza de la base de datos de los datos no utilizados y en la reducción de la redundancia y la inconsistencia. Los datos procedentes de sistemas transaccionales, por ejemplo, en los que se realizan continuamente varias operaciones de inserción, actualización y borrado, están muy normalizados. La desnormalización se utiliza para combinar los datos de varias tablas en una sola tabla, de modo que puedan consultarse más rápidamente. Por ejemplo, los datos normalizados procedentes de sistemas transaccionales se suelen desnormalizar antes de ejecutar las consultas para la presentación de informes y análisis. Otro tipo de transformación es la Limpieza. Las tareas de limpieza son acciones que corrigen las irregularidades en los datos para producir un análisis creíble y preciso. Los datos que son inexactos, faltantes o incompletos pueden sesgar los resultados del análisis y deben ser tenidos en cuenta. También puede ser que los datos estén sesgados, o que tengan valores nulos en los campos relevantes, o que tengan valores atípicos. Por ejemplo, tal vez se quiera averiguar la información demográfica sobre la venta de un determinado producto, pero los datos que se han recibido no reflejan el género. Es necesario obtener este punto de datos y fusionarlo con el conjunto de datos existente, o tal vez sea necesario eliminarlo y no tener en cuenta los registros en los que falta este campo. Exploraremos muchos más ejemplos de limpieza de datos más adelante en el curso. Enriquecer los datos es el cuarto tipo de transformación. Cuando valoras los datos que tienes, para mirar puntos de datos adicionales que podrían hacer tu análisis más significativo, estás buscando enriquecer tus datos. Por ejemplo, en una gran organización con información fragmentada en varios sistemas, puede ser necesario enriquecer el conjunto de datos proporcionado por un sistema con información disponible en otros sistemas, o incluso con conjuntos de datos públicos. Consideremos un escenario en el que vendes periféricos de TI a empresas y quieres analizar los patrones de compra de tus clientes en los últimos cinco años. Dispones de las tablas maestras de clientes y de transacciones desde donde has capturado la información de los clientes y el historial de compras. Complementar tu conjunto de datos con los datos de rendimiento de estas empresas, posiblemente disponibles como un conjunto de datos públicos, podría ser valioso para que comprendas los factores que influyen en sus decisiones de compra. La inserción de metadatos también enriquece los datos. Por ejemplo, calcular una puntuación de sentimiento a partir de un registro de comentarios de clientes, recopilar datos meteorológicos basados en la geografía de una ubicación de centros turísticos para analizar las tendencias de ocupación, o capturar la hora y las etiquetas publicadas para una entrada de blog. Después de la transformación, la siguiente fase del Data Wrangling es la Validación. Aquí es donde se comprueba la calidad de la estructuración, normalización, limpieza y enriquecimiento de la entrada de datos. Las reglas de validación se refieren a los pasos de programación repetitivos utilizados para verificar la consistencia, calidad y seguridad de los datos que tienes. Esto nos lleva a la Publicación - la cuarta fase del proceso de data wrangling. La publicación implica la entrega del resultado de los datos preparados para las necesidades del proyecto. Lo que se publica es la versión transformada y validada del conjunto de datos de entrada junto con los metadatos referentes al conjunto de datos. Por último, es preciso señalar la importancia crítica de documentar los pasos y valoraciones que se han hecho para convertir los datos en bruto en datos listos para el análisis. Todas las fases del data wrangling son de naturaleza iterativa. A fin de replicar los pasos y volver a examinar las decisiones tomadas para llevar a cabo estos pasos, es vital que se documenten todas esas decisiones y acciones.

# Herramientas para Data Wrangling

In this video, we will look at some of the popularly used data wrangling software and tools, such as: Excel Power Query / Spreadsheets, OpenRefine, Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python and R. Let’s begin with the most basic software used for manual wrangling—Spreadsheets. Spreadsheets such as Microsoft Excel and Google Sheets have a host of features and in-built formulae that can help you identify issues, clean, and transform data. Add-ins are available that allow you to import data from several different types of sources and clean and transform data as needed—such as Microsoft Power Query for Excel and Google Sheets Query function for Google Sheets. OpenRefine is an open-source tool that allows you to import and export data in a wide variety of formats, such as TSV, CSV, XLS, XML, and JSON. Using OpenRefine, you can clean data, transform it from one format to another, and extend data with web services and external data. OpenRefine is easy to learn and easy to use. It offers menu-based operations, which means you don’t need to memorize commands or syntax. Google DataPrep is an intelligent cloud data service that allows you to visually explore, clean, and prepare both structured and unstructured data for analysis. It is a fully managed service, which means you don’t need to install or manage the software or the infrastructure. DataPrep is extremely easy to use. With every action that you take, you get suggestions on what your ideal next step should be. DataPrep can automatically detect schemas, data types, and anomalies. Watson Studio Refinery, available via IBM Watson Studio, allows you to discover, cleanse, and transform data with built-in operations. It transforms large amounts of raw data into consumable, quality information that’s ready for analytics. Data Refinery offers the flexibility of exploring data residing in a spectrum of data sources. It detects data types and classifications automatically and also enforces applicable data governance policies automatically. Trifacta Wrangler is an interactive cloud-based service for cleaning and transforming data. It takes messy, real-world data and cleans and rearranges it into data tables, which can then be exported to Excel, Tableau, and R. It is known for its collaboration features, allowing multiple team members to work simultaneously. Python has a huge library and set of packages that offer powerful data manipulation capabilities. Let’s look at a few of these libraries and packages. Jupyter Notebook is an open-source web application widely used for data cleaning and transformation, statistical modeling, also data visualization. Numpy, or Numerical Python, is the most basic package that Python offers. It is fast, versatile, interoperable, and easy to use. It provides support for large, multi-dimensional arrays and matrices, and high-level mathematical functions to operate on these arrays. Pandas is designed for fast and easy data analysis operations. It allows complex operations such as merging, joining, and transforming huge chunks of data, performed using simple, single-line commands. Using Pandas, you can prevent common errors that result from misaligned data coming in from different sources. R, also offers a series of libraries and packages that are explicitly created for wrangling messy data—such as Dplyr, Data.table, and Jsonlite. Using these libraries, you can investigate, manipulate, and analyze data. Dplyr is a powerful library for data wrangling. It has a precise and straightforward syntax. Data.table helps to aggregate large data sets quickly. Jsonlite is a robust JSON parsing tool, great for interacting with web APIs. Tools for data wrangling come with varying capabilities and dimensions. Your decision regarding the best tool for your needs will depend on factors that are specific to your use case, infrastructure, and teams—such as supported data size, data structures, cleaning and transformation capabilities, infrastructure needs, ease of use, and learnability.

# Limpieza de Datos

Según un informe de Gartner sobre la calidad de los datos, la mala calidad de los datos debilita la posición competitiva de una organización y socava los objetivos críticos de la empresa. Datos incompletos, inconsistentes o incorrectos pueden llevar a conclusiones falsas y por lo tanto a decisiones ineficaces. Y en el mundo de los negocios, eso puede ser costoso. Los conjuntos de datos recogidos de fuentes dispares pueden tener varios problemas, como valores faltantes, inexactitudes, duplicados, delimitadores incorrectos o ausentes, registros incoherentes y parámetros insuficientes. En algunos casos, los datos pueden corregirse manual o automáticamente con la ayuda de herramientas y scripts de data wrangling, pero si no pueden repararse, deben eliminarse del conjunto de datos. Aunque los términos Data Cleaning ("limpieza de datos") y Data Wrangling se utilizan a veces indistintamente, es importante tener en cuenta que la limpieza de datos es sólo un subconjunto de todo el proceso de Data Wrangling. La limpieza de datos forma una parte muy significativa e integral de la fase de Transformación en un flujo de trabajo de Data Wrangling. Un flujo de trabajo típico de limpieza de datos incluye: Inspección, Limpieza y Verificación. El primer paso en el flujo de trabajo de limpieza de datos es detectar los diferentes tipos de problemas y errores que puede tener el conjunto de datos. Se pueden utilizar scripts y herramientas que permiten definir reglas y restricciones específicas, y validar los datos frente a estas reglas y restricciones. Para la inspección, también se pueden utilizar herramientas de perfilado y visualización de datos. La creación de perfiles de datos ayuda a inspeccionar los datos de origen para comprender la estructura, contenido e interrelaciones de los datos. Descubre anomalías y problemas en la calidad de los datos. Por ejemplo, valores en blanco o null, datos duplicados, o si el valor de un campo está dentro del rango esperado. La visualización de datos mediante métodos estadísticos puede ayudar a detectar valores atípicos. Por ejemplo, trazar el promedio de ingresos en un conjunto de datos demográficos puede ayudarte a detectar los valores atípicos. Esto nos lleva a la limpieza real de los datos. Las técnicas que apliques para limpiar el conjunto de datos dependerán de tu caso de uso y del tipo de problemas que encuentres. Veamos algunos de los problemas más comunes de los datos. Empecemos con los valores que faltan. Es muy importante tratar los valores ausentes, ya que pueden causar resultados inesperados o sesgados. Puedes optar por dejar fuera los registros con valores perdidos o encontrar una forma de obtener esa información en caso de que sea intrínseca a tu caso de uso. Por ejemplo, los datos de edad que falten en un estudio demográfico. Una tercera opción es un método conocido como imputación, que calcula el valor faltante basándose en valores estadísticos. La decisión sobre el curso de acción que elijas debe basarse en lo que sea mejor para tu caso de uso. También podrías encontrarte con datos duplicados, puntos de datos que se repiten en el conjunto de datos. Estos se han de eliminar. Otro tipo de problema que puedes encontrar es el de los datos irrelevantes. Los datos que no encajan en el contexto de tu caso de uso pueden considerarse datos irrelevantes. Por ejemplo, si estás analizando datos sobre la salud general de un segmento de la población, sus números de contacto pueden no ser relevantes para ti. La limpieza también puede implicar la conversión del tipo de datos. Esto es necesario para garantizar que los valores de un campo se almacenen como el tipo de datos de ese campo, por ejemplo, los números almacenados como tipo de datos numéricos o la fecha almacenada como tipo de datos de fecha. También puede ser necesario limpiar los datos para estandarizarlos. Por ejemplo, en el caso de las cadenas de texto, es posible que quieras que todos los valores estén en minúsculas. Del mismo modo, es necesario normalizar los formatos de fecha y las unidades de medida. También hay errores de sintaxis. Por ejemplo, los espacios en blanco o los espacios adicionales al principio o al final de una cadena son un error de sintaxis que debe rectificarse. Esto también puede incluir la corrección de errores tipográficos o de formato, por ejemplo, el nombre del estado que se introduce en algunos registros como una forma completa, como Nueva York frente a una forma abreviada como NY. Los datos también pueden tener valores atípicos, o valores que son muy diferentes de otras observaciones en el conjunto de datos. Los valores atípicos pueden o no ser incorrectos. Por ejemplo, cuando un campo de edad en una base de datos de votantes tiene el valor 5, sabes que son datos incorrectos y deben corregirse. Ahora consideremos un grupo de personas cuyo ingreso anual está en el rango de cien mil a doscientos mil dólares, excepto por una única persona que gana un millón de dólares al año. Si bien este punto de datos no es incorrecto, es atípico y debe examinarse. Dependiendo de tu caso de uso, puede que tengas que decidir si la inclusión de estos datos distorsionará los resultados de manera que no sirva a tu caso de uso. Esto nos lleva al siguiente paso en el flujo de trabajo de limpieza de datos: la Verificación. En este paso, inspeccionas los resultados para determinar la eficacia y la precisión alcanzadas como resultado de la operación de limpieza de datos. Es necesario volver a inspeccionar los datos para asegurarse de que las reglas y restricciones aplicables a los datos siguen vigentes después de las correcciones realizadas. Y, al final, es importante señalar que todos los cambios realizados como parte de la operación de limpieza de datos deben documentarse. No sólo los cambios, sino también las razones para hacer esos cambios, y la calidad de los datos almacenados actualmente. Informar del grado de salud de los datos es un paso muy crucial.

# Data Cleaning

En este segmento, los profesionales de los datos comparten qué parte de su trabajo implica la recopilación, la limpieza y la preparación de los datos para el análisis. Diría que una proporción relativamente grande de mi trabajo involucra la recopilación, preparación y limpieza de datos para el análisis. Trabajo en una empresa con un equipo de ingeniería de datos realmente grande. Así que no tengo que hacer este tipo de trabajo tanto como otros científicos de datos. Pero aún así, cualquier persona que trabaje estrechamente con los datos, ya sea un científico de datos, un analista de datos, un ingeniero de machine learning, realmente necesita obtener una buena comprensión de dónde provienen los datos. Inevitablemente, ningún conjunto de datos es perfecto. Siempre habrá compromisos o pequeños errores. Así que es realmente importante pasar una porción significativa del tiempo entendiendo los datos subyacentes que se usaron para generar el conjunto de datos y qué problemas potenciales podrían surgir con esos datos. Mi trabajo como contable público implica mucho análisis. Extractos financieros, actividad de cuentas, procesos de evaluación y controles. La recopilación puede ser bastante simple siempre y cuando, la información contable resida en un sistema de libro mayor o en un repositorio central donde los datos sean fáciles de recopilar. Probablemente, alrededor del 30 por ciento del trabajo consiste en preparar todo. Así que cuando te metes en el análisis, puedes profundizar directamente en el tema. Por ello, hay que rastrear los datos, asegurarse de que son precisos, asegurarse de que las cosas cuadran. Asegúrate de que tienes toda la información. Así que, por ejemplo, en cuanto a los extractos financieros, tengo que asegurarme de que la gente me ha dado 12 meses de extractos financieros, que no me falta ningún dato y que, si es así, que tengo suficiente información para poder proyectar o pronosticar o incluso mirar hacia atrás para estimar lo que se hizo durante el año en base a lo que tengo. Eso es definitivamente útil. En este segmento, los profesionales de los datos hablan de los pasos que dan para asegurar que los datos sean fiables. Uno de los pasos esenciales para asegurarse de que los datos son fiables es elaborar estadísticas de sumario en las columnas individuales de los datos y asegurarse de que son coherentes con la realidad. Por ejemplo, si tienes una columna en algún lugar que registra las visitas por mes a un sitio web y realizas estadísticas de sumario en esa columna, obtienes el mínimo, la media, la mediana, el máximo, y ves algo raro como, un mes hay visitas negativas o algo así. Ya sabes, esos datos no son fiables. La información financiera en particular debe ser fiable. No debe tener sesgos. Debe estar libre de errores. Esos son sólo algunos de los muchos atributos necesarios para que los datos sean confiables. Es decir, hacer lo que yo llamo una comprobación lógica antes de entrar en los detalles de una transacción. ¿Tiene sentido a alto nivel? Si esperabas que los ingresos de alto nivel aumentaran, pero ves que han disminuido drásticamente, entonces calcula esa parte primero. ¿Es mi fuente correcta? ¿Estoy haciendo una consulta en el período correcto? ¿Estoy sacando la cuenta correcta del libro mayor? Así que empieza por ahí, asegúrate de que se hayan abordado primero las preguntas básicas sobre la integridad de los datos. Una vez que sepamos que los datos son fiables, entonces podemos empezar a profundizar en las revisiones y formar conclusiones sobre el rendimiento financiero basado en nuestro análisis de los datos.

Resumen y Aspectos Resultantes

En esta lección, has aprendido la siguiente información:

Una vez que los datos que identificaste son reunidos e importados, tu próximo paso es hacer que estén listos para el análisis. Aquí es donde el proceso de Transformación de Datos Brutos (Data Wrangling o Data Munging), entra en juego. La Transformación de Datos Brutos es un proceso iterativo que implica la exploración, transformación y validación de datos.

La transformación de los datos brutos incluye las tareas a las que te comprometes:

● Manipular y combinar estructuralmente los datos utilizando Joins y Unions.

● Normalizar los datos, es decir, limpiar la base de datos de los datos no utilizados y redundantes.

● Desnormalizar los datos, es decir, combinar los datos de múltiples tablas en una sola tabla para que puedan ser consultados más rápidamente.

● Limpiar los datos, lo que implica perfilar los datos para descubrir problemas de calidad, visualizar los datos para detectar valores atípicos y arreglar problemas como valores perdidos, datos duplicados, datos irrelevantes, formatos inconsistentes, errores de sintaxis y valores atípicos.

● Enriquecer los datos, lo que implica considerar puntos de datos adicionales que podrían añadir valor al conjunto de datos existentes y conducir a un análisis más significativo.

Se dispone de una variedad de software y herramientas para el proceso de Transformación de Datos Brutos. Algunos de los más utilizados son Excel Power Query, Hojas de Cálculo, OpenRefine, Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python y R, cada uno con su propio conjunto de características, fortalezas, limitaciones y aplicaciones.

# Visión General del Análisis Estadístico

Antes de entender el Análisis Estadístico, su relación con el Análisis de Datos, y específicamente la extracción de datos ("data mining"), examinemos primero qué es la Estadística. La Estadística es una rama de las matemáticas que se ocupa de la recogida, análisis, interpretación y presentación de datos numéricos o cuantitativos. Está a nuestro alrededor en nuestra vida diaria. Ya sea que hablemos de ingresos promedio, edad promedio, o profesiones mejor pagadas, todo es estadística. Hoy en día, la estadística se aplica en todas las industrias para la toma de decisiones basadas en datos. Por ejemplo, los investigadores utilizan las estadísticas para analizar los datos de la producción de vacunas para garantizar la seguridad y la eficacia, o las empresas utilizan las estadísticas para reducir la pérdida de clientes al obtener una mayor comprensión de las necesidades de los clientes. Ahora veamos qué es el análisis estadístico. El Análisis Estadístico es la aplicación de métodos estadísticos a una muestra de datos para desarrollar una comprensión de lo que esos datos representan. Incluye la recopilación y el escrutinio de cada muestra de datos en un conjunto de elementos de los que se pueden extraer muestras. Una muestra, en Estadística, es una selección representativa extraída de una población total, en la que la población es un grupo discreto de personas o cosas que pueden identificarse por lo menos por una característica común a efectos de recopilación y análisis de datos. Por ejemplo, en un determinado caso de uso, la población puede ser todas las personas de un estado que tengan permiso de conducir, y una muestra de esta población que sea parte, o subconjunto, de la población podría ser de hombres conductores mayores de 50 años. Los métodos estadísticos son principalmente útiles para asegurar que los datos se interpreten correctamente, y que las relaciones aparentes sean significativas y no sucedan por casualidad. Siempre que recogemos datos de una muestra, hay dos tipos diferentes de estadísticas que podemos realizar. Estadísticas descriptivas para resumir la información sobre la muestra; y Estadísticas inferenciales para hacer inferencias o generalizaciones sobre la población en general. La Estadística Descriptiva permite presentar los datos de una manera significativa permitiendo una interpretación más simple de los mismos. Los datos se describen utilizando diagramas, tablas y gráficos de resumen sin ningún intento de sacar conclusiones sobre la población de la que se toma la muestra. El objetivo es facilitar la comprensión y la visualización de los datos en bruto sin sacar conclusiones sobre ninguna hipótesis que se haya formulado. Por ejemplo, queremos describir los resultados de la prueba de inglés en una clase específica de 25 estudiantes. Registramos los resultados de todos los estudiantes, calculamos las estadísticas de resumen y generamos un gráfico. Algunas de las métricas comunes del Análisis Estadístico Descriptivo incluyen la Tendencia Central, la Dispersión y la Asimetría: Tendencia Central, o localizar el centro de una muestra de datos. Algunas de las métricas comunes de la tendencia central incluyen la media, la mediana y la moda. Estas métricas te dicen dónde caen la mayoría de los valores de tu conjunto de datos. Así, en el ejemplo anterior, la puntuación media, o la media matemática, de la clase de 25 estudiantes sería la suma total de las puntuaciones de todos los 25 estudiantes, dividida por 25, es decir, el número de estudiantes. Si ordena el conjunto de datos anterior desde el valor de puntuación más pequeño al valor de puntuación más alto de los 25 estudiantes y elige el valor medio, es decir, el valor con 12 valores a la izquierda y 12 valores a la derecha de un valor de puntuación, ese valor de puntuación sería la mediana de este conjunto de datos. Si 12 estudiantes han obtenido una puntuación inferior al 75% y 12 estudiantes han obtenido una puntuación superior al 75%, entonces la mediana es 75. La mediana es única para cada conjunto de datos y no se ve afectada por los valores atípicos. La moda es el valor que ocurre con más frecuencia en un conjunto de observaciones. Por ejemplo, si la puntuación más común en este grupo de 25 estudiantes es 72%, entonces esa es la moda de este conjunto de datos. Por tanto, puedes ver que observar el conjunto de datos a través de estos valores puede ayudarte a comprender mejor el conjunto de datos. La dispersión es la medida de la variabilidad en un conjunto de datos. Las medidas comunes de dispersión estadística son la Varianza, la Desviación Estándar y el Rango. La varianza define lo lejos que están los puntos de datos del centro, es decir, la distribución de los valores. Cuando una distribución tiene menor variabilidad, los valores de un conjunto de datos son más consistentes. Sin embargo, cuando la variabilidad es mayor, los puntos de datos son más dispares, y tener valores extremos se hace más probable. Comprender la variabilidad puede ayudar a comprender la probabilidad de que ocurra un evento. La desviación estándar indica lo estrechamente que los datos están agrupados en torno a la media. Y el Rango te da la distancia entre los valores más pequeños y más grandes de los conjuntos de datos. Asimetría es la medida de si la distribución de los valores es simétrica alrededor de un valor central o está sesgada a la izquierda o a la derecha. Los datos sesgados pueden afectar a los tipos de análisis que se pueden realizar con validez. Estos son algunos de los instrumentos estadísticos descriptivos básicos y más utilizados, pero también hay otros instrumentos, por ejemplo, el uso de la correlación y los gráficos de dispersión para evaluar las relaciones de los datos emparejados. El segundo tipo de análisis estadístico es la Estadística Inferencial. La estadística inferencial toma datos de una muestra para hacer deducciones sobre la población más amplia de la que se extrajo la muestra. Utilizando métodos de estadística inferencial se pueden extraer generalizaciones que aplican los resultados de la muestra a la población en su conjunto. Algunas metodologías comunes de la Estadística inferencial son la Prueba de Hipótesis, los Intervalos de Confianza y el Análisis de Regresión: Prueba de Hipótesis - Por ejemplo, para estudiar la eficacia de una vacuna comparando los resultados en un grupo de control, las pruebas de hipótesis pueden indicar si es probable que la eficacia de una vacuna observada en un grupo de control exista también en la población. Los Intervalos de Confianza incorporan la incertidumbre y el error de la muestra para crear un rango de valores en el que se encuentra el valor real de la población. El Análisis de Regresión incorpora pruebas de hipótesis que ayudan a determinar si las relaciones observadas en los datos de la muestra existen realmente en la población y no sólo en la muestra. Existen varios paquetes de software para realizar el análisis de datos estadísticos, como el Sistema de Análisis Estadístico (o SAS), el Paquete Estadístico para las Ciencias Sociales (o SPSS) y Stat Soft. La estadística constituye el núcleo de la extracción de datos por: Proporcionar las métricas y metodologías necesarias para la minería de datos; e Identificar patrones que ayuden a descubrir las diferencias entre el ruido aleatorio y los hallazgos significativos. Tanto la minería de datos, sobre la que estudiaremos más en este curso, como la Estadística, como técnicas de análisis de datos, ayudan a tomar mejores decisiones.

# ¿Qué es Data Mining?

Data mining ("minería de datos"), o el proceso de extracción de conocimientos de los datos, es el corazón del proceso de análisis de datos. Es un campo interdisciplinario que implica el uso de tecnologías de reconocimiento de patrones, análisis estadístico y técnicas matemáticas. Su objetivo es identificar correlaciones en los datos, encontrar patrones y variaciones. Comprender las tendencias y predecir las probabilidades. Con frecuencia se oye hablar de patrones y tendencias en el contexto del análisis de datos, así que primero entendamos estos conceptos. El reconocimiento de patrones es el descubrimiento de la regularidad o lo común en los datos.

Reproduce el video desde ::51 y sigue la transcripción0:51

Consideremos los datos de registro de los inicios de sesión de una aplicación en una organización. Contiene información como nombre de usuario, fecha y hora de inicio de sesión, tiempo empleado en cada sesión de inicio de sesión y actividades realizadas. Cuando analizamos estos datos para obtener información sobre los hábitos o comportamientos de los usuarios, por ejemplo, el momento del día en que el número máximo de usuarios tiende a iniciar sesión o los roles de usuario que típicamente pasan el máximo de horas conectados a la aplicación o los módulos del flujo de trabajo de la aplicación que se están utilizando, ya sea examinando los datos manualmente o mediante herramientas para descubrir patrones ocultos en los datos. Una tendencia, por otra parte, es la evolución general de un conjunto de datos para cambiar con el paso del tiempo. Por ejemplo, el calentamiento global a corto plazo, como las temperaturas base anuales, que pueden permanecer iguales o subir o bajar unos pocos grados, pero las temperaturas globales en general siguen aumentando con el tiempo, lo que hace del calentamiento global una tendencia.

Reproduce el video desde :1:58 y sigue la transcripción1:58

La minería de datos tiene aplicaciones en todos los sectores y disciplinas. Por ejemplo, la elaboración de perfiles del comportamiento sobre las necesidades de los clientes y los ingresos disponibles a fin de ofrecer campañas dirigidas, instituciones financieras, rastrear las transacciones de los clientes en busca de comportamientos inusuales y marcar las transacciones fraudulentas mediante modelos de minería de datos. El uso de modelos estadísticos para predecir la probabilidad de un paciente de padecer determinados trastornos de salud y dar prioridad al tratamiento.

Reproduce el video desde :2:29 y sigue la transcripción2:29

El acceso a los datos de rendimiento de los estudiantes para predecir los niveles de rendimiento y hacer un esfuerzo concentrado para proporcionar apoyo donde sea necesario. Ayudar a las agencias de investigación a desplegar la fuerza policial en los lugares donde la probabilidad de que se cometa un delito es mayor, y alinear la oferta y la logística con las previsiones de la demanda. Existen varias técnicas que se pueden utilizar para detectar patrones y crear modelos precisos para el descubrimiento, ya sea descriptivo, de diagnóstico, predictivo o prescriptivo. Entendamos algunas de las técnicas más utilizadas. La clasificación es una técnica que clasifica los atributos en categorías de objetivos, por ejemplo, clasificando a los clientes en gastadores bajos, medios o altos en función de lo que ganan.

Reproduce el video desde :3:18 y sigue la transcripción3:18

La agrupación es similar a la clasificación, pero implica agrupar los datos en clústeres para que puedan tratarse como grupos. Por ejemplo, la agrupación de clientes basada en la detección de anomalías de regiones geográficas o de valores atípicos es una técnica que ayuda a encontrar patrones y datos que no son normales o son inesperados. Por ejemplo, picos en el uso de una tarjeta de crédito que pueden indicar un posible uso indebido.

Reproduce el video desde :3:45 y sigue la transcripción3:45

La minería de reglas de asociación es una técnica que ayuda a establecer nuestra relación entre dos eventos de datos. Por ejemplo, la compra de un portátil que frecuentemente va acompañada de la compra de un panel de refrigeración. Patrones secuenciales es la técnica que rastrea una serie de eventos que tienen lugar en una uencia. Por ejemplo, seguir el rastro de las compras de un cliente desde el momento en que entra en una tienda online hasta el momento en que sale de ella.

Reproduce el video desde :4:13 y sigue la transcripción4:13

La agrupación por afinidad es una técnica utilizada para descubrir la coocurrencia en las relaciones. Esta técnica se utiliza ampliamente en las tiendas online para realizar ventas cruzadas y vender algo mejor que lo que inicialmente se deseaba, recomendando productos a las personas basándose en el historial de compras de otras personas que han comprado el mismo artículo.

Reproduce el video desde :4:32 y sigue la transcripción4:32

Los árboles de decisión ayudan a construir modelos de clasificación en forma de una estructura de árbol con múltiples ramas, en la que cada rama representa una ocurrencia probable. Esta técnica ayuda a comprender claramente la relación entre la entrada y la salida.

Reproduce el video desde :4:49 y sigue la transcripción4:49

La regresión es una técnica que ayuda a identificar la naturaleza de la relación entre dos variables, que puede ser causal o correlativa. Por ejemplo, basándose en factores como la ubicación y la superficie ocupada, se podría utilizar un modelo de regresión para predecir el valor de una casa.

Reproduce el video desde :5:8 y sigue la transcripción5:08

La minería de datos ayuda esencialmente a separar el ruido de la información real y ayuda a que las empresas centren sus energías sólo en lo que es relevante.

# Herramientas para Data Mining

In this video, we will learn about some of the commonly used software and tools for data mining, such as: Spreadsheets, R-Language, Python, IBM SPSS Statistics, IBM Watson Studio; and SAS. Spreadsheets, such as Microsoft Excel and Google Sheets, are commonly used for performing basic data mining tasks. Spreadsheets can be used to host data that has been exported from other systems in an easily accessible and easy-to-read format. You can pivot tables to showcase specific aspects of your data, which is vital when you have huge amounts of data to sort through and analyze. They also make it relatively easier to make comparisons between different sets of data. Add-ins available for Excel, such as the Data Mining Client for Excel, XLMiner, and KnowledgeMiner for Excel, allow you to perform common mining tasks, such as classification, regression, association rules, clustering, and model building. GoogleSheets also has an array of add-ons that can be used for analysis and mining, such as Text Analysis, Text Mining, Google Analytics.

Reproduce el video desde :1:27 y sigue la transcripción1:27

R is one of the most widely used languages for performing statistical modeling and computations by statisticians and data miners. R is packaged with hundreds of libraries explicitly built for data mining operations such as regression, classification, data clustering, association rule mining, text mining, outlier detection, and social network analysis. Some of the popular R packages include tm and twitteR. tm, a framework for text mining applications within R, provides functions for text mining. twitteR provides a framework for mining tweets. R Studio is a popularly used open-source Integrated Development Envionrment (or IDE) for working with the R programming language. Python libraries like Pandas and NumPy are commonly used for Data Mining. Pandas is an open-source module for working with data structures and analysis. It is possibly one of the most popular libraries for data analysis in Python. It allows you to upload data in any format and provides a simple platform to organize, sort, and manipulate that data. Using Pandas, you can: perform basic numerical computations such as mean, median, mode, and range; calculate statistics and answer questions regarding correlation between data and distribution of data; explore data visually and quantitatively; visualize data with help from other Python libraries. NumPy is a tool for mathematical computing and data preparation in Python. NumPy offers a host of built-in functions and capabilities for data mining. Jupyter Notebooks have become the tool of choice for Data Scientists and Data Analysts when working with Python to perform data mining and statistical analysis. SPSS stands for Statistical Process for Social Sciences. While the name suggests its original usage in the field of Social Sciences, it is popularly used for advanced analytics, text analytics, trend analysis, validation of assumptions, and translation of business problems into data science solutions. SPSS is closed source and requires a license for use. SPSS has an easy to use interface that requires minimal coding for complex tasks. It comprises of efficient data management tools and is popular because of its in-depth analysis capabilities and accurate data results. IBM Watson Studio, included in the IBM Cloud Pak for Data, leverages a collection of open source tools such as Jupyter notebooks, and extends them with closed source IBM tools that make it a powerful environment for data analysis and data science. It is available through a web browser on the public cloud, private cloud, and as a desktop app. Watson Studio enables team members to collaborate on projects, that can range from simple exploratory analysis to building machine learning and AI models. It also includes SPSS Modeller flows that enable you to quickly develop predictive models for your business data. SAS Enterprise Miner is a comprehensive, graphical workbench for data mining. It provides powerful capabilities for interactive data exploration, which enables users to identify relationships within data. SAS can manage information from various sources, mine and transform data, and analyze statistics. It offers a graphical user interface for non-technical users. With SAS, you can: identify patterns in the data using a range of available modeling techniques; explore relationships and anomalies in data; analyze big data; validate the reliability of findings from the data analysis process. SAS is very easy to use because of its syntax and is also easy to debug. It has the ability to handle large databases and offers high security to its users. In this video, we have learned about just a few of the data mining tools available today. Your decision regarding the best tool for your needs will be driven by the data size and structures the tool supports, the features it offers, its data visualization capabilities, infrastructure needs, ease of use, and learnability. It’s fairly common to use a combination of data mining tools to meet all your needs.

# Resumen y Aspectos Resultantes

En esta lección, has aprendido la siguiente información:

La estadística es una rama de las matemáticas que se ocupa de la recopilación, análisis, interpretación y representación de datos numéricos o cuantitativos.

El análisis estadístico comprende el uso de métodos estadísticos para desarrollar una comprensión de los que representan los datos.

El Análisis Estadístico puede ser:

●      Descriptivo; es aquel que proporciona un resumen de lo que representan los datos. Las medidas comunes incluyen la Tendencia Central, la Dispersión y la Asimetría.

●      Inferencial; es la que implica hacer inferencia o generalizaciones, sobre los datos. Las medidas comunes incluyen la Prueba de Hipótesis, los Intervalos de Confianza y el Análisis de Regresión.

La Minería de Datos, simplemente, es el proceso de extraer conocimiento de los datos. Implica el uso de tecnologías de reconocimiento de patrones, análisis estadísticos y técnicas matemáticas, a fin de identificar correlaciones, patrones, variaciones y tendencias en los datos.

Hay varias técnicas que pueden ayudar a minar datos, como la clasificación de los atributos de los datos, la agrupación de los datos en grupos, el establecimiento de las relaciones entre los eventos, las variables, las entradas y las salidas.

Existen una diversidad de programas informáticos y herramientas para analizar y extraer datos. Algunos de los más utilizados son Hojas de Cálculo, R-Language, Python, IBM SPSS Statistics, IBM Watson Studio y SAS, cada uno con su propio conjunto de características, fortalezas, limitaciones y aplicaciones.

MODACONJUNTO DE DAT

# Visión General de la Comunicación y el Intercambio de los Resultados del Análisis de Datos

El proceso de análisis de datos comienza con la comprensión del problema que hay que resolver y el resultado deseado al que hay que llegar. Y termina con la comunicación de los hallazgos de una forma que impacte en la toma de decisiones. Los proyectos de datos son el resultado de un esfuerzo de colaboración que se extiende a todas las funciones de la empresa en las que participan personas con habilidades multidisciplinarias, y los resultados se incorporan a una iniciativa empresarial más amplia. El éxito de lo que comuniques depende de lo bien que los demás puedan entender y confiar en tus percepciones para emprender acciones posteriores. Por lo tanto, como analistas de datos, es necesario contar la historia con los datos, visualizando las percepciones con claridad y creando una narrativa estructurada dirigida explícitamente a tu audiencia. Antes de empezar a crear la comunicación, necesitas reconectar con tu audiencia. Empieza por hacerte estas preguntas: ¿Quién es mi público? ¿Qué es importante para ellos? ¿Qué les ayudará a confiar en mí? Tu público va a ser en su mayoría un grupo diverso, en términos de las funciones empresariales que representan, si desempeñan un papel operativo o estratégico en la organización, cómo les afecta el problema y otros factores similares. La presentación debe enmarcarse en el nivel de información que ya tiene tu público. Sobre la base de tu comprensión del público, decidirás qué información es esencial, y en qué medida, para permitir una mejor comprensión de tus conclusiones. Es tentador sacar a la luz todos los datos con los que se ha estado trabajando, pero hay que considerar qué elementos son más importantes para el público que otros. Una presentación no es un volcado de datos. Los hechos y las cifras por sí solos no influyen en las decisiones ni llevan a la gente a la acción. Tienes que contar una historia convincente. Incluye sólo la información necesaria para abordar el problema de la empresa. Demasiada información hará que tu audiencia se esfuerce por entender el punto que estás planteando. Empieza tu presentación demostrando a tu audiencia que entiendes el problema de la empresa. Es fácil caer en la suposición de que todos sabemos para qué estamos aquí, pero mostrar que se entiende el problema que hay que resolver y el resultado que hay que lograr es un gran primer paso para ganar su atención y comenzar con confianza. Hablar en el lenguaje del ámbito de la empresa es otro factor importante para crear una conexión con el público. El siguiente paso en el diseño de la comunicación es estructurar y organizar la presentación para lograr el máximo impacto. Consulta los datos que has recopilado. Recuerda que los datos, la base misma de todo lo que estás comunicando, es como una caja negra para la audiencia. Si no se puede comprobar la credibilidad de los datos, la gente no sabrá si puede confiar en los resultados. Comparte tus fuentes de datos, hipótesis y validaciones. Trata de establecer la credibilidad de tus conclusiones sobre la marcha - no pases por alto ninguna de las suposiciones clave hechas durante el análisis. Organiza la información en categorías lógicas basadas en la información que tienes, ¿tienes información tanto cualitativa como cuantitativa, por ejemplo? Intenta adoptar un enfoque de arriba a abajo o de abajo a arriba en tu narración. Ambos pueden ser eficaces, según el público y el caso de uso. Sé coherente en tu enfoque. Es importante determinar qué formatos de comunicación serán más útiles para tu público. ¿Necesitan llevarse un resumen ejecutivo, una hoja de hechos o un informe? Lo que debe determinar los formatos que elijas es la forma en que tu público va a utilizar la información que has presentado. Las percepciones deben explicarse de manera que inspiren a la acción. Si tu público no comprende el significado de tus percepciones o no están convencidos de su utilidad, esas percepciones no tendrán ningún valor. Un ensayo de mil palabras no tendrá el mismo impacto que un contenido visual en la creación de una imagen mental clara en las mentes de su audiencia. Una visualización potente cuenta una historia a través de la representación gráfica de hechos y cifras. Las visualizaciones de datos - gráficos, tablas, diagramas - son una gran manera de dar vida a los datos. Ya sea que muestres una comparación, relación, distribución o composición, tienes herramientas que pueden ayudarte a mostrar patrones y conclusiones acerca de las hipótesis. Los datos tienen valor en las historias que cuentan. Tu público debe ser capaz de confiar en ti, entenderte y sentirse identificado con tus hallazgos y percepciones. Al determinar la credibilidad de los hallazgos, presentar los datos dentro de una narración y apoyarlos con expresiones visuales, se puede ayudar a que el público obtenga percepciones valiosas.

# Puntos de Vistas: Narración de historias de Análisis de Datos

En este vídeo, vamos a escuchar a profesionales de los datos hablando sobre el papel que juega la narración de historias y la vida de un analista de datos. El papel de la narración de historias en la vida de un analista de datos no puede subestimarse. Es absolutamente necesario ser muy bueno narrando historias con datos. Creo que los humanos entienden de forma natural el mundo a través de las historias. Si intentas convencer a alguien de que haga algo con los datos, lo primero que tienes que hacer es contar una historia clara, concisa y convincente. También creo que puede ser muy útil para el analista de datos desarrollar una historia en cualquier momento que esté trabajando con un conjunto de datos, para ayudarse a sí mismo a entender mejor el conjunto de datos subyacente y lo que está haciendo. Siempre habrá un equilibrio entre contar una historia clara, coherente y sencilla, y asegurarse de que se transmiten todas las complejidades que se pueden encontrar en los datos. Creo que encontrar ese equilibrio puede ser muy difícil, pero es realmente crucial. El arte de la narración de historias es significativo en la vida de un analista de datos. No importa cuánta o lo maravillosa que sea la información que hayas encontrado. Si no puedes hallar una forma de comunicar eso a tu audiencia, ya sea el consumidor o un director o una persona de nivel ejecutivo, entonces no sirve de nada. Tienes que encontrar una forma de comunicar eso y normalmente es mejor hacerlo de forma visual o mediante el relato de una historia, para que entiendan cómo esa información puede ser útil. Tengo que decir que contar historias es una habilidad esencial. Es como la última milla en un reparto. Mucha gente puede controlar el lado técnico tras un corto período de entrenamiento. Sin embargo, la capacidad de extraer valor de los datos y comunicarlo escasea. Si piensas en una carrera a largo plazo, creo que es fundamental saber cómo contar una historia convincente con datos. Contar historias es absolutamente crucial para los datos y el análisis. Así es como realmente transmites tu mensaje. Todo el mundo puede mostrar números, pero si no tienes una historia que contar, si no tienes una razón convincente para actuar, entonces, en última instancia, lo que estás presentando no va a calar en tu audiencia. Llevaron a cabo un estudio en Stanford donde hicieron que la gente presentara sus argumentos y en esa presentación tenían simplemente KPIs (indicadores clave), estadísticas numéricas, pero también contaban la historia. A los miembros de la audiencia se les preguntó después qué es lo que recordaban de cada una de esas presentaciones, y fueron esas historias las que se les quedaron grabadas. Sí, en la historia había hechos y cifras, pero así es como lo haces llegar. Tener esa conexión emocional con la historia, con la comprensión, con los datos, es realmente como vas a conseguir que la gente lleve a cabo la acción que quieres y necesitas que emprendan.

# Introducción a la Visualización de Datos

La visualización de datos es la disciplina de comunicar información mediante el uso de elementos visuales como gráficos, diagramas y mapas. Su objetivo es hacer que la información sea fácil de comprender, interpretar y retener. Imagina tener que mirar entre miles de filas de datos para sacar interpretaciones y compararlas con una representación visual de esos mismos datos que resuma los resultados. Utilizando la visualización de datos, se puede proporcionar un resumen de las relaciones, tendencias y patrones ocultos en los datos, que, aunque no imposible, sería muy difícil de descifrar a partir de un volcado de datos. Para que la visualización de datos resulte de valor, hay que elegir la visualización que más eficazmente haga llegar los resultados a la audiencia. Y para ello, tienes que empezar por hacerte algunas preguntas. ¿Cuál es la relación que estoy tratando de establecer? ¿Quiero comparar la proporción relativa de las subpartes de un conjunto, por ejemplo, la contribución de las diferentes líneas de productos en los ingresos totales de la empresa? ¿Quiero comparar valores múltiples, como el número de productos vendidos y los ingresos generados en los últimos tres años? O bien, ¿quiero analizar un único valor a lo largo del tiempo?, lo que en este ejemplo podría significar cómo ha cambiado la venta de un producto específico en los últimos tres años. ¿Necesito que mi público vea la correlación entre dos variables? La correlación entre las condiciones meteorológicas y las reservas en una estación de esquí, por ejemplo. ¿Quiero detectar anomalías en los datos, por ejemplo, encontrar valores en los datos que podrían potencialmente desviar los resultados? La pregunta que estoy tratando de responder no es sólo una pregunta general sobre el diseño y el proceso de visualización de datos, sino que debes ser capaz de responder a esta pregunta para tu público con cada conjunto de datos e información que visualices. También debes considerar si la visualización debe ser estática o interactiva. Una visualización interactiva, por ejemplo, puede permitirte cambiar los valores y ver los efectos en una variable relacionada en tiempo real. Por lo tanto, has de pensar en la conclusión principal que tu público va a obtener, anticipar sus necesidades de información y las preguntas que puedan tener, y luego planear la visualización que transmita tu mensaje de forma clara e impactante. Veamos algunos ejemplos básicos de los tipos de gráficos que se pueden crear para visualizar los datos. Las gráficas de barras son excelentes para comparar conjuntos de datos relacionados o partes de un todo. Por ejemplo, en esta gráfica de barras, puedes ver los números de población de 10 países diferentes y cómo se comparan entre sí. Las Gráficas de Columna comparan los valores uno al lado del otro. Se pueden utilizar con bastante eficacia para mostrar cambios a lo largo del tiempo. Por ejemplo, mostrando cómo las visitas a las páginas y el tiempo de las sesiones de usuario de tu sitio web está cambiando mes a mes. Aunque son similares, excepto por la orientación, las gráficas de barras y las gráficas de columnas no siempre se pueden utilizar indistintamente. Por ejemplo, una gráfica de columnas puede ser más adecuada para mostrar valores negativos y positivos. Las gráficas de tarta muestran el desglose de una entidad en sus subpartes y la proporción de las subpartes en relación con cada una de ellas. Cada porción de la tarta representa un valor estático o una categoría, y la suma de todas las categorías es igual al cien por cien. En este ejemplo, en una campaña de marketing con cuatro canales de comercialización -redes sociales, publicidad nativa, influencers pagados y eventos en directo- se puede ver el número total de clientes potenciales que genera cada canal. Las Gráficas de Líneas muestran tendencias. Son excelentes para mostrar cómo cambia el valor de un dato en relación con una variable continua. Por ejemplo, cómo ha cambiado la venta de un producto, o de varios productos, a lo largo del tiempo, donde el tiempo es la variable continua. Las gráficas de líneas pueden utilizarse para comprender las tendencias, patrones y variaciones en los datos; también, para comparar conjuntos de datos diferentes pero relacionados con diversas series. La visualización de datos también se puede utilizar para construir cuadros de mando. Los cuadros de mando organizan y muestran informes y visualizaciones procedentes de múltiples fuentes de datos en una única interfaz gráfica. Los cuadros de mando pueden utilizarse para supervisar el progreso diario o la salud general de una función empresarial o incluso de un proceso específico. Los cuadros de mando pueden presentar tanto datos operativos como analíticos. Por ejemplo, podrías tener un cuadro de mandos de marketing con el que supervisar tu actual campaña de marketing en relación a los contactos directos, consultas generadas y conversiones de ventas, en tiempo real. Como parte del mismo cuadro de mando, también podrías ver cómo la tasa de conversión de esta campaña se compara con la tasa de conversión de otras campañas ejecutadas con éxito en el pasado. Los cuadros de mando son una gran herramienta para presentar una vista panorámica de la imagen completa, a la vez que permite profundizar en el siguiente nivel de información para cada parámetro. Los cuadros de mando: son fáciles de comprender para un usuario medio, facilitan la colaboración entre equipos y permiten generar informes sobre la marcha. Usando cuadros de mando, puedes ver el resultado de las variaciones en los datos y métricas casi instantáneamente, y esto puede ayudarte a evaluar una situación desde múltiples perspectivas, sobre la marcha, sin tener que volver al tablero de diseño.

# Introducción al Software de Visualización y Cuadros de Mando

En este video, vamos a ver algunos de los programas y herramientas de visualización de datos más utilizados. Entre ellas se incluyen: Hojas de cálculo, Cuadernos Jupyter y librerías Python, R-Studio y R-Shiny, IBM Cognos Analytics, Tableau y Microsoft Power BI. Algunas de ellas son soluciones de análisis de datos de principio a fin, mientras que otras son específicamente para la visualización de datos, desde herramientas gratuitas de código abierto hasta soluciones disponibles en el mercado. Las hojas de cálculo, como Microsoft Excel y Google Sheets, son posiblemente el software más utilizado para hacer representaciones gráficas de conjuntos de datos. Las hojas de cálculo son fáciles de aprender y tienen una gran cantidad de documentación y tutoriales en vídeo disponibles en línea para una fácil consulta. Excel proporciona varios tipos de gráficas que van desde las gráficas básicas de barras, líneas, tarta y dinámicas, hasta las opciones más avanzadas como las gráficas de dispersión, líneas de tendencia, gráficas de Gantt, gráficas de cascada y gráficas de combinación (con las que se pueden combinar más de un tipo de gráficas). Excel también proporciona recomendaciones sobre la mejor representación visual para tu conjunto de datos. Para que los gráficos sean más presentables, puedes añadir un título al gráfico, cambiar los colores de los elementos y añadir etiquetas a los datos. Google Sheets también ofrece tipos de gráficos similares para la visualización, aunque Excel tiene más opciones basadas en fórmulas integradas que Google Sheets. Al igual que Excel, Google Sheets puede ayudarte a elegir la visualización adecuada. Todo lo que tienes que hacer es resaltar los datos que deseas visualizar y hacer clic en el botón de la gráfica, y obtendrás una lista de gráficas sugeridas que se adaptan mejor a tus datos. Las gráficas y los informes se actualizan automáticamente, tanto en Excel como en Google Sheets, a medida que se cambian los datos subyacentes. Cuando es necesario que varios usuarios colaboren, es preferible Google Sheets a Excel. Jupyter Notebook es una aplicación web de código abierto que proporciona una excelente manera de explorar datos y crear visualizaciones. No tienes que ser un experto en Python para usar Jupyter Notebook. Python proporciona una gran cantidad de bibliotecas que se utilizan para la visualización de datos. Veamos algunas de esas bibliotecas. Matplotlib es una biblioteca muy utilizada para la visualización de datos en Python. Proporciona diferentes tipos de gráficos en 2D y 3D, y la flexibilidad de crear gráficos de varias maneras diferentes. Usando Matplotlib, se pueden crear gráficas interactivas de alta calidad con sólo unas pocas líneas de código. Tiene un gran respaldo de la comunidad y soporte multi-plataforma, ya que es una herramienta de código abierto. Bokeh proporciona gráficos y diagramas interactivos y es conocido por ofrecer una interactividad de alto rendimiento sobre grandes conjuntos de datos o en streaming. Bokeh ofrece flexibilidad para aplicar a la visualización interacción, diseños y diferentes opciones de estilo. También puede transformar visualizaciones creadas en otras bibliotecas de Python, como Matplotlib, Seaborn y Ggplot. Dash es un framework de Python para crear visualizaciones interactivas basadas en la web. Usando Dash, puedes crear aplicaciones web altamente interactivas usando código Python. Aunque es útil tener conocimientos de HTML y javascript, no es obligatorio. Dash es de fácil mantenimiento, multiplataforma, y listo para su uso en móviles. Usando R-Studio, puedes crear visualizaciones básicas como histogramas, gráficas de barras, de líneas, de cajas y de dispersión; y visualizaciones avanzadas, como mapas de calor, mapas de mosaico, gráficas 3D y correlogramas. Shiny es un paquete R que ayuda a crear aplicaciones web interactivas que se pueden alojar como aplicaciones independientes en una página web. Estas aplicaciones web muestran de manera transparente objetos R, como gráficos y tablas, y pueden ponerse a disposición de cualquiera en vivo. También puedes crear cuadros de mando usando Shiny. Lo fácil que es trabajar con Shiny es lo que lo hizo popular entre los profesionales de los datos. IBM Cognos Analytics es una solución analítica integral. Algunas de las características de visualización que proporciona Cognos incluyen: Importación de visualizaciones personalizadas; Una función de previsión que proporciona modelado de datos de series temporales y previsiones basadas en los datos presentados en las visualizaciones correspondientes; Recomendación de visualizaciones basadas en tus datos; Formato condicional que te permite ver la distribución de los datos y resaltar puntos de datos excepcionales, por ejemplo, resaltando las cifras de ventas altas y bajas por encima de un determinado umbral; Cognos es conocido por sus excelentes visualizaciones y la superposición de datos en el mundo físico mediante sus capacidades geoespaciales. Tableau es una empresa de software que elabora productos de visualización interactiva de datos. Utilizando los productos de Tableau, se pueden crear gráficos y tablas interactivas en forma de cuadros de mando y hojas de trabajo, usando arrastrar y soltar. Tableau también ofrece la opción de publicar los resultados en forma de historias. Puedes importar scripts R y Python en Tableau y aprovechar sus características de visualización, que son muy superiores a las de otros lenguajes. Las capacidades de visualización de Tableau son fáciles e intuitivas de usar. Tableau es compatible con archivos de Excel, archivos de texto, bases de datos relacionales y fuentes de bases de datos en la nube como Google Analytics y Amazon Redshift. Power BI es un servicio de análisis de negocios basado en la nube de Microsoft que permite crear informes y cuadros de mando. Es una herramienta potente y flexible conocida por su velocidad y eficiencia, y una interfaz de arrastrar y soltar fácil de usar. Power BI es compatible con múltiples fuentes, incluyendo Excel, SQL Server y repositorios de datos basados en la nube, lo que lo convierte en una excelente opción para los profesionales de los datos. Power BI proporciona la capacidad de colaborar y compartir paneles personalizados e informes interactivos de forma segura, incluso en los móviles. El cuadro de mando de Power BI se compone de muchas visualizaciones en una sola página que ayudan a contar la historia. Estas visualizaciones, llamadas "mosaicos", están fijadas al cuadro de mando. El panel es interactivo, lo que significa que un cambio en un mosaico afecta al otro. Al decidir qué herramientas usar, es necesario tener en cuenta la facilidad de uso y el propósito de la visualización. En términos de las herramientas que están disponibles y las capacidades de visualización que ofrecen - si lo puedes visualizar, lo puedes crear.

# Puntos de vista: Herramientas de Visualización

En este vídeo, los profesionales de los datos hablan de las herramientas de visualización en las que más confían y por qué.

Reproduce el video desde ::16 y sigue la transcripción0:16

>> La herramienta de visualización en la que confío más a diario es Cognos Analytics. Hay varias razones para esto. Una, me permite tomar muy rápidamente una hoja de cálculo importante, conectarla a una base de datos y visualizar mis datos, ya sea yo mismo tratando de comprender lo que quiero ver y arrastrando los campos, o usando nuestro asistente de IA para presentar los datos y ayudarme a comprender y explorar lo que podría ser interesante, si es un nuevo conjunto de datos, con el que nunca he trabajado antes.

Reproduce el video desde ::41 y sigue la transcripción0:41

Además de eso, también puedo seguir adelante y empezar a hacer algunas cosas más complejas, o incluso sólo algunos análisis más robustos con nuestra herramienta de informes que me permite elaborar y programar los informes para su entrega. Si quisiera que mis ventas tuvieran un informe de conductos o un informe de oportunidades de ventas todos los lunes por la mañana, lo configuro una vez, uso lo que llamamos bursky, y luego lo envío automáticamente todos los domingos por la noche, por lo que les estará esperando por la mañana. Además de esto puedo buscar y combinar múltiples fuentes de datos, y el sistema me ayuda a crear las uniones de esos datos. Y luego puedo visualizarlos todos en un simple cuadro de mandos de símbolos que es altamente interactivo, permitiendo filtrar y ordenar dinámicamente, así como compartir eso con el resto de una organización, para que no todos los usuarios tengan que pasar por la misma experiencia. Hemos creado un panel una vez, y todos pueden tener acceso a él. >> En cuanto a las herramientas de visualización, me apoyo más en Looker, que es una herramienta de visualización de datos que se encuentra en la base de datos interna de mi empresa. Es similar a Tableau, que también he usado en el pasado y me parece bastante fácil de usar. Y lo mejor de estas herramientas de visualización de datos como Looker y Tableau es que permiten que todos en la organización, independientemente de si son o no profesionales de los datos, puedan ver fácilmente sus datos y hacer una agrupación u ordenación básica de los mismos. Una herramienta de visualización de datos en la que realmente confío para el análisis exploratorio de datos es R. En los últimos años me he convertido en un gran partidario de la efectividad de hacer análisis básicos de datos y visualización de datos en R, particularmente usando Tidyverse, que es una colección de paquetes que ayudan a cargar fácilmente los datos, agruparlos a diferentes niveles y también visualizarlos rápida y fácilmente. >> Tableau y Power BI son una decisión obvia, son fáciles de entender y muy útiles para presentar los datos. Y a medida que más y más empresas y personas han empezado a utilizarlos, hay más y más plantillas integradas en sus bibliotecas. >> Yo diría que lo visual, probablemente sería Excel y Word, sólo la suite de Microsoft, y sólo mirando y utilizando sumatorios y macros para asegurarme de que los datos, antes incluso de sumergirme en lo que viene, tienen sentido y están preparados para lo que necesitamos que sean.

Resumen y Aspectos Resultantes

En esta lección, has aprendido la siguiente información:

Los datos tienen valor a través de las historias que cuentan. Para poder comunicar tus hallazgos de manera impactante, es necesario:

● Asegurar de que tu público pueda confiar en ti, comprender y relacionar tus hallazgos y percepciones.

● Establecer la credibilidad de los resultados.

● Presentar los datos dentro de una narrativa estructurada.

● Apoyar la comunicación con visualizaciones sólidas para que el mensaje sea claro y conciso y que también impulse a la audiencia a tomar medidas.

La visualización de datos es la disciplina de comunicar información a través del uso de elementos visuales tales como gráficos, tablas y mapas. El objetivo de la visualización de datos es hacer que la información sea fácil de comprender, interpretar y retener.

Para que la visualización de datos tenga valor, es necesario:

● Pensar en el tema principal que puede aprender tu público.

● Anticipar sus necesidades de información y preguntas, para luego planear la visualización, que entregues tu mensaje de manera clara e impactante.

Hay varios tipos de gráficos y cuadros disponibles para que tú puedas trazar cualquier tipo de datos, como gráficos de barras, de columnas, circulares y de líneas.

También puedes utilizar la visualización de datos para construir paneles de datos. Estos paneles organizan y muestran informes y visualizaciones procedentes de múltiples fuentes de datos en una única interfaz gráfica. Son fáciles de comprender y permiten generar informes sobre la marcha.

Al decidir qué herramientas utilizar para la visualización de datos, es necesario tener en cuenta la facilidad de uso y el propósito de la visualización. Algunas de las herramientas más utilizadas son las hojas de cálculo, Jupyter Notebook, Bibliotecas de Python, R-Studio y R-Shiny, IBM Cognos Analytics, Tableau y Power BI.

# Oportunidades Profesionales en Análisis de Datos

Existen vacantes para analistas de datos en la industria, el gobierno y el mundo académico. Cada industria, ya sea banca y finanzas, seguros, salud, comercio minorista o tecnología de la información tiene espacio para analistas de datos cualificados. Estos puestos son muy solicitados en las grandes empresas, así como en startups y nuevas empresas. Según Forbes, se espera que el gran mercado mundial de análisis de datos, que se situó en 37.340 millones de dólares estadounidenses en 2018, crezca a una tasa anual compuesta del 12,3% de 2019 a 2027 para alcanzar los 105.080 millones de dólares estadounidenses en el año 2027. En la actualidad, la demanda de analistas de datos cualificados supera con creces la oferta, lo que significa que las empresas están dispuestas a pagar un plus por contratar a analistas de datos cualificados.

Reproduce el video desde :1:6 y sigue la transcripción1:06

Hay una amplia variedad de puestos de trabajo disponibles para los analistas de datos. Las opciones para una carrera profesional están abiertas, clasificaremos los roles en analistas de datos, roles especializados y roles de especialistas en determinados dominios. Los puestos de especialista en análisis de datos son para los analistas de datos que quieren mantenerse centrados y crecer en los aspectos técnicos y funcionales de su rol. En esta trayectoria. Podrías comenzar tu carrera como analista de datos asociado o junior y avanzar por los roles de analista, analista senior, analista responsable y analista principal. Los límites entre estos roles, los años de experiencia que te capacitan para el siguiente nivel y la naturaleza de la experiencia que necesitas obtener para ascender podrían variar dependiendo del sector, el tamaño de la organización y el tamaño de tu equipo. En equipos más pequeños, por ejemplo, podrías adquirir experiencia en todas las facetas del análisis de datos, desde la recopilación de datos hasta la visualización y presentación de las conclusiones a los interesados, y esto puede ocurrir en un breve lapso de tiempo en equipos y organizaciones más grandes, los roles se pueden bifurcar normalmente en función de la actividad, lo que significa que podrías estar adquiriendo experiencia en una fase específica del proceso antes de pasar a la siguiente. Esto te ayuda a perfeccionar tus habilidades en una parte del proceso antes de pasar a la siguiente. En tu trayectoria desde analista de datos asociado a analista de datos responsable o principal, estarás continuamente mejorando tus habilidades técnicas, estadísticas y analíticas desde un nivel básico a un nivel experto. Estarás demostrando tu capacidad para trabajar con una amplia gama de herramientas y ataformas. Diferentes aspectos del proceso de análisis de datos y una amplia variedad de casos de uso en términos de habilidades técnicas, puedes empezar conociendo sólo una herramienta de consulta y un lenguaje de programación. Cualquier tipo de repositorio de datos o un conjunto limitado de herramientas de visualización. A medida que vayas adquiriendo más experiencia, se espera que aprendas y demuestres tu capacidad para trabajar con más y más herramientas, lenguajes, datos, repositorios y nuevas tecnologías, tus habilidades de comunicación, de presentación, de gestión de las partes interesadas y de gestión de proyectos, todas ellas deben perfeccionarse y adquirirse progresivamente. Como analista responsable o principal, también puedes ser el responsable de establecer procesos en tu equipo, haciendo recomendaciones sobre software y herramientas. El equipo debe trabajar en el perfeccionamiento del equipo y en su ampliación para incluir más perfiles. En algunas organizaciones, estas responsabilidades podrían alinearse con la persona de nivel directivo que ha subido de rango para dirigir un equipo de analistas de datos. Los especialistas de un sector, también conocidos como analistas funcionales, son analistas que requieren especialización en un sector específico y son considerados una autoridad en su sector, como en el sector de la salud, las ventas, las finanzas, las redes sociales o el marketing digital. Puede que no sean las personas más capacitadas técnicamente. Estos roles llevan títulos como analista, analista de marketing, analista de ventas, analista de atención sanitaria o analista de redes sociales.

Reproduce el video desde :4:31 y sigue la transcripción4:31

Y luego están los roles de trabajo orientados a la analítica. Estos incluyen roles como directores de proyecto, directores de marketing y directores de recursos humanos. Estos son trabajos en los que las habilidades analíticas conducen a una mayor eficiencia y eficacia. Una buena parte de los puestos de trabajo de analista de datos están orientados al análisis, ya que cada vez más organizaciones dependen de los datos para la toma de decisiones.

Reproduce el video desde :4:56 y sigue la transcripción4:56

Como analista de datos, también tienes opciones para explorar y aprender nuevas habilidades para entrar en otras profesiones relacionadas con los datos, como la ingeniería de datos o la ciencia de los datos. Por ejemplo, si estás empezando como analista de datos junior y te gusta mucho trabajar con data lakes y repositorios de big data, puedes adquirir más experiencia en estas tecnologías y hacer evolucionar tu carrera hasta convertirte en un gran ingeniero de datos. Si el lado empresarial de las cosas te entusiasma más, podrías explorar de forma similar las habilidades necesarias para hacer un movimiento lateral hacia el análisis de negocios o el análisis de inteligencia de negocios.

Reproduce el video desde :5:35 y sigue la transcripción5:35

Aunque el panorama de la carrera de analista de datos es muy vasto, lo bueno es que tienes una plétora de recursos disponibles para ayudarte a crecer para tener éxito en tu viaje como analista de datos, todo lo que necesitas hacer es aprovechar las oportunidades que quieres perseguir, o las que se te presenten y aprender a lo largo del camino.

# Puntos de vista: Entrar en una Profesión dedicada a los Datos

En este video, vamos a escuchar a profesionales de los datos hablando de cómo entraron en esta profesión. Mi papel actual como profesional de los datos no existía antes de que yo asumiera el puesto. Me di cuenta de que había una necesidad en nuestra empresa de proporcionar datos de una manera más rápida y eficiente, que ir al departamento de SI que tendría una reunión para tratar la reunión, para tener los requisitos, y entonces tendrían un producto final con el que la gente no estaría satisfecha. Pero tenías que ponerte al final de la fila y pasar por todo el proceso de nuevo, para conseguir lo que buscabas. Al cubrir la necesidad de la empresa para proporcionar informes en dos semanas, preparé una base de datos de la empresa que tenía acceso a más información. Tenemos analistas que ahora son capaces de satisfacer esa demanda desatendida en la empresa. Me metí en el rol de profesional de datos por casualidad. En realidad estaba trabajando en mi doctorado en Economía en la Universidad de Illinois, Urbana-Champaign, cuando un colega mío sugirió que un máster en estadística también sería un excelente valor añadido. Así es como entré en el programa de estadística, también en Illinois. Pero una vez que empecé con eso, me enganché bastante y no hubo vuelta atrás, por así decirlo. En otras palabras, mi objetivo original de convertirme en economista se convirtió en una carrera llena de datos, modelado, análisis, recopilación de información, comunicación, visualización y, por supuesto, en la base de la resolución de problemas basada en datos. Me metí en un puesto de analista de datos en una empresa de datos financieros, en realidad por accidente. En aquel entonces, mi compañía comenzó a contratar analistas de datos de valores en China, y tuve mucha suerte de unirme al equipo, porque buscaban a alguien que tuviera habilidades de análisis financiero, que yo podía aportar. Después de eso, mi equipo comenzó a contratar gente con habilidades técnicas como Python, R y Sickle. Siempre me han gustado los números. Una de las cosas que pasa es que cuando trabajas tanto con números, éstos empiezan a contar una historia, y la capacidad de mirar esos números y contar esa historia es lo que me motiva. Siempre he tenido ese interés por los números, ya sea porque siempre me atrajo el análisis de datos y las hojas de cálculo de Excel, o porque fuera QuickBooks, o cualquier tipo de conjunto de datos que pueda ayudarnos a obtener la información que buscamos, especialmente en el sector financiero, donde nos fijamos en los beneficios, las pérdidas y el balance general, y en lo que sucede cuando una empresa compra a otra. Siempre estamos mirando esos datos para hablar, y hablar de la historia de la empresa, y del futuro. Conseguí mi puesto actual como científico de datos directamente desde mi programa de graduación, que fue un Master en Ciencias de los Datos. Antes de mi graduación, trabajé como analista de datos y director de análisis.

# Puntos de vista: Qué buscan los Contrastadores en un Analista de Datos

En este video, vamos a escuchar a los profesionales de los datos hablar de lo que los empleadores buscan en un Analista de Datos. Los empleadores buscan Analistas de Datos con integridad. Durante el proceso de contratación, les preguntaré, si tuvieran que elegir sólo uno, ¿preferirían cumplir con un plazo o conseguir una respuesta correcta? Siempre busco a alguien que diga, quiero asegurarme de que la información es correcta. El incumplimiento de una fecha límite no es tan perjudicial para una empresa como una decisión multimillonaria sobre una información errónea, o como que alguien pierda su trabajo porque no se obtuvo o no se informó correctamente. Es mucho más importante tener integridad. Creo que lo que más buscan los empleadores en los analistas de datos es alguien que pueda comunicarse con claridad. Si haces el análisis más brillante del mundo, pero no puedes comunicarlo a las partes interesadas externas, entonces no vale nada. Creo que esa es una habilidad muy demandada. Creo que otra cosa que las empresas obviamente esperan cuando buscan un Analista de Datos es fluidez con los números, habilidad para entender análisis complejos, habilidad para entender las pruebas AB y lo que dicen los resultados de las pruebas AB, y la implicación de esos resultados. También creo que, cada vez más, las empresas buscan Analistas de Datos con conocimientos de SQL muy fuertes. Otra cosa que los empleadores buscan en los analistas de datos es una mentalidad de crecimiento y la voluntad de aprender, porque la industria está cambiando a un ritmo realmente rápido. Creo que están buscando conocimientos de programación, incluyendo Python, R, SQL. Al mismo tiempo, están buscando ciertas personalidades. Ya sea que estés interesado en los detalles, que te guste trabajar con datos, y que seas un solucionador de problemas, etcétera, etcétera. Como empleadora, contrato gente todo el tiempo. ¿Qué es lo que estoy buscando? Buscamos gente que esté centrada en los detalles y que haga más de lo esperado. No sólo quieren hacer lo que tienen delante, sino que quieren ir más allá. Buscamos personas que tengan aspiraciones más altas, y que también sean capaces de pensar de forma innovadora. Si digo, haz ABC, no sólo van a hacer eso, van a hacer más, haz más y dame algunas alternativas. Gente que sea capaz de resolver problemas. Si algo sale mal, no van a parar y decir, Dios mío, tengo que ir a hablar con mi supervisor. Van a decir: "Aquí hay un problema, esto es lo que pienso. Aquí hay dos posibles soluciones para resolver esto, para que el trabajo y la empresa puedan seguir adelante." Eso es lo que se quiere. No sólo centrado en los detalles y no sólo bueno con los números. También tienes que ser alguien que pueda pensar de forma innovadora, y ser capaz de resolver problemas, y de solucionarlos. Eso es lo que los empleadores van a estar buscando ahora más que nunca. Buscan la capacidad de conocer los datos, y por conocer los datos nos referimos a varias cosas. Estar cómodo con ellos en varios formatos, ser capaz de pensar en eso. Con eso queremos decir, saber qué datos quieres para resolver los problemas que se presentan. La capacidad de conocer los datos es muy importante. La resolución de problemas es otra habilidad muy importante. Es decir, si se presenta un problema a un Analista de Datos, éste debe ser capaz de saber cómo abordar ese problema utilizando los datos en cualquier formato en que se encuentren, y ser capaz de analizarlos y presentar las ideas que luego resolverán el problema. También necesitan ser muy dinámicos, en el sentido de que si se les presenta un conjunto de datos muy diferente de repente, que no se parece en nada a lo que era antes, deben ser capaces de adaptarse a ese cambio. Por eso la cualidad de ser dinámico y adaptable también es importante. También han de ser capaces de adquirir conocimientos técnicos rápidamente. Con esto queremos decir que si hay un tal DIAdem SQL que se utiliza en un entorno, tienen que ser capaces de trabajar bajo un paradigma diferente. Si hay un lugar que está usando RStudio, pero conocen Python, deben ser capaces de manejar RStudio rápidamente, y esas cosas. Ser capaz de aprender rápido, ser dinámico, y conocer los datos, eso es lo poco que buscan los empleadores en un buen Analista de Datos.

# Los Múltiples Caminos en Análisis de Datos

There are various paths you can take for gaining entry into the data analyst field. While some employers may ask for an academic degree as a pre-requisite, even if you don’t have a degree, you still have several options available to you that can help you gain an entry, or even make a lateral move, into the field of data analysis. Let’s start with the most obvious path. An academic degree in Data Analytics, Statistics, Computer Science, Management Information Systems, or Information Technology Management can start you off with a strong advantage. You could alternately enroll in online training programs that can equip you with the required knowledge. Comprehensive online programs for data analysis are multi-course specializations offered by learning platforms such as Coursera, edX, and Udacity. These courses are designed and delivered by some of the world’s best domain experts. Since you have a fair idea, by now, of the technical, functional, and soft skills you need in order to be a data analyst, choosing the right learning path should be fairly straightforward. As you gather more work experience, you can keep advancing your knowledge and skills in specific areas, for example, Statistics, Spreadsheets, SQL, Python, Data Visualization, Problem-Solving, Storytelling, or making impactful presentations. These courses also give you hands-on assignments and projects which give you a feel for the real-world application of your knowledge and skills. You can even add these projects to your portfolio. So, if you don’t have an academic qualification, these courses can help you gain opportunities at an entry-level and work your way up as your experience grows. Now let’s look at a scenario where you have a couple of years of experience in a different line of work and want to make a switch into the data analysis field. There’s a very good chance that you can do that successfully if you plan well. Since data analysis is a vast field, it would be useful for you to first research the knowledge and skills you need, the various job opportunities that are available, and the growth opportunities available on the path you may be considering. You can tap into online resources, forums, and your network of friends and colleagues to connect with people in this field and gain insights into real-world scenarios. If you’re currently working in a non-technical role, you may consider exploring the Domain Specialist, or Functional Analyst path. If you’re in Sales, you could consider starting your journey by positioning and skilling yourself for a Sales Analyst position. You begin with the advantage of industry experience and skill yourself in other areas such as Statistics and programming, for example. If you’re currently working in a technical role, you have the ability to quickly pick up the tools and software you need for the data analyst role. You’re also probably stepping in with the advantage of having a good understanding of the domain or industry you’re from. For some of the other skills, such as problem-solving, project management, communication, and storytelling—you may already be using these in some capacity in your existing job. You can always enhance these skills through trainings, online courses, communities of practice, and forums. Data Analysis is a fast-moving field. If you’re curious, open to learning new things, and excited about the field, you will be able to forge a path forward, regardless of the formal qualifications you think you may be missing.

# Puntos de vista: Opciones de Carrera para Profesionales de los Datos

En este video, vamos a escuchar a profesionales de los datos en ejercicio hablando sobre las diversas opciones de carrera disponibles en este sector. El conjunto de la profesión relacionada con los datos hoy en día también se ha vuelto muy colorida, muy dinámica, evolucionando todo el tiempo, y también presenta un gran abanico de opciones a cualquiera que quiera entrar en el sector para ser un profesional de los datos. Parte desde, si pensaras en diferentes círculos como opciones, comenzando con Analista de Datos. A partir de ahí, puedes ampliar mucho más la escala y convertirte en un científico de datos. También puedes convertirte en un estadístico, que es lo que yo era cuando empecé. Luego puedes especializarte más aún en una dirección específica para convertirte en ingeniero de datos. O puedes empezar siendo un analista de BI o un especialista y luego no ir a convertirte en ingeniero de datos. En otras palabras, puedes hacer una carrera de Analista de Datos y Científico de Datos, o puedes hacer una carrera de Analista de BI y de Ingeniero de Datos. Esas son trayectorias paralelas dentro de la profesión de los datos. También puedes ir al otro extremo, donde puedes convertirte en un Ingeniero de Machine Learning, un Ingeniero de IA y así sucesivamente. Hay muchos roles de este tipo que cualquier persona interesada en el campo de los datos puede realmente asumir. Algunas de las opciones de carrera más comunes disponibles para un Analista de Datos consisten en profundizar en la materia con Machine Learning e Ingeniería, y convertirse en Científico de Datos o en Ingeniero de Machine Learning, que se centran más en modelar Machine Learning. Otra opción de carrera disponible para el Analista de Datos es sumergirse más profundamente en el negocio en el que ya estás y aportar a la estrategia de la empresa a alto nivel. Creo que ese papel es realmente importante e interesante y ha evolucionado mucho en los últimos años. Otra ruta para un analista de datos es comenzar a convertirse en gestor de personas y dirigir a otros Analistas de Datos y trabajar para seleccionar en qué se trabaja. Porque siempre habrá más preguntas en la organización que se puedan responder con datos que personas que las puedan contestar. El rol de Director de Datos puede ser realmente interesante y crítico en términos de asegurar de que se trabaje realmente en las partes más importantes. Puedes ser Administrativo. Puedes ser Contable. Puedes ser Contable Público. Puedes ser Corredor de Bolsa o Analista Financiero para el gobierno o para muchas grandes empresas. Puedes ser un Agente Inmobiliario. Mucha gente es buena como Analista de Datos, pero para hacer eso tienen que gustarte mucho los números y tienes que estar muy interesado en los detalles. Si no eres así y los números no te llaman la atención, Analista de Datos podría no ser lo correcto para ti.

# Puntos de vista: Consejos para aspirantes a Analistas de Datos

En este video, vamos a escuchar a profesionales de los datos dando consejos a aspirantes a analistas de datos. Un consejo que le daría a los aspirantes a analistas de datos es que sigan aprendiendo y no se desanimen. Se ha escrito más sobre análisis de lo que podrías aprender en toda tu vida. No intentes aprender todo de una vez, pero tómate tu tiempo y asegúrate de que cada semana, cada mes, cada año estés constantemente aprendiendo algo nuevo. Creo que eso te servirá bien. Un consejo que me han ado en mi carrera y que me ha sido muy útil es considerar tu carrera como una T mayúscula, y debes tener amplios conocimientos. La parte superior de la T representa que debes tener un amplio conocimiento en un número de áreas diferentes. Aunque no tiene que ser profundo, deberías saber un poco, al menos sobre las pruebas A/B, sobre machine learning, sobre visualización de datos, sobre SQL, sobre Python, sobre R. La parte inferior de la T es que deberías profundizar en al menos un área. Debería haber un área entre las que acabo de mencionar, n la que se tenga un entendimiento realmente profundo y riguroso de la misma. Utiliza todos los trabajos que tengas a tu favor, lo que significa que siempre hay en todo algo que se puede sacar. Ya sea mirando el presupuesto de tus padres o preguntando a tus padres si puedes ver los talonarios de cheques o si trabajas en un restaurante de comida rápida, mirando los números. ¿Cuántas personas van a venir? ¿Cuántos dólares se están entregando? Habla con el gerente sobre lo siguiente, lo que significan los números en realidad. ando estés hablando con empleadores potenciales, ten listos tus ejemplos. No tiene que ser necesariamente sólo una experiencia de palabra, sino tu experiencia en la vida, cómo usas los análisis, cómo has usado los análisis incluso en tu vida personal. Si puedes contarme y hablarme sobre lo que has hecho, personal o profesionalmente, y cómo se relaciona con lo que estamos haciendo. Eso te hará avanzar mucho. Un consejo que daría a los aspirantes a científicos de datos es que creen un portafolio profesional que muestre sus conocimientos en ciencia o análisis de datos. Puedes hacer esto buscando conjuntos de datos divertidos online y analizando esos conjuntos de datos. También puedes hacerlo en tu trabajo. Incluso si tu trabajo actual no es ser analista de datos, busca oportunidades en las que puedas hacer cálculos, y eso te llevará naturalmente a tener un buen portafolio o a obtener buenos resultados en términos de proyectos de análisis de datos. Mi consejo para un aspirante a analista de datos es que siga su pasión. Encuentra un trabajo que satisfaga tus necesidades y te aporte felicidad al hacerlo. No hay nada peor que levantarse cada mañana y odiar ir a tu puesto de trabajo. Hay tantos trabajos de analista de datos en diversas industrias, departamentos. Hay tantas opciones que no hay necesidad de aceptar un trabajo, sólo por tener un trabajo. Encuentra algo que realmente alimente tu pasión y te dé algo por lo que levantarte cada mañana.

# Puntos de vista: Mujeres en Profesiones de Datos

En este video, vamos a escuchar a mujeres que comparten su experiencia de ser profesionales de los datos, y su consejo a las mujeres que aspiran a entrar en este campo. Como mujer en Ciencia de los Datos, todavía me encuentro con el estereotipo de que este es un trabajo de hombres. He entrado en reuniones y he visto gente decepcionada o confundida. Lo tomo como una oportunidad para demostrarles que están equivocados. Este no es un trabajo sólo para hombres, es para una persona que tiene la perspicacia, la habilidad y la motivación para hacer el trabajo. Mientras posea esas habilidades, no hay razón para que nadie no pueda hacer nada que se proponga. Ya seas hombre o mujer, blanco o negro, tienes la oportunidad de demostrar que la gente se equivoca gracias al trabajo que realizas. Diría que puede ser difícil, pero tienes que encontrar tu voz y no tener miedo de usarla. Muchas veces, como mujeres, no hemos sido capaces de expresarnos o decir lo que pensamos, o tenemos miedo de cómo nos tratará la gente si decimos lo que pensamos. Pero es más importante que se te escuche y se te vea, no sólo hablando en voz alta, sino que si tienes los datos para respaldarte, si tienes buenos contenidos y cosas que quieres decir, no tengas miedo de levantar la mano y hacer saber a la gente que eres una pensadora y que puedes conseguirlo, porque eso va a ser importante a medida que avances. La única forma real de avanzar es impulsarse, y la gente no sabe que tienes impulso si eres demasiado callada. Si estás trabajando tranquilamente en un rincón, muchas veces la gente no puede verlo. Habla, asegúrate de que tu voz se oiga, asegúrate de que te vean como una mujer que sabe cómo crecer y cómo ayudar en el campo de la ciencia de los datos. Cuando empecé, eran sobre todo hombres en mi clase, especialmente en la escuela de posgrado. Pero ahora, veo que los equipos de datos, tanto de ciencia de datos como de ingeniería de datos, y también están llenos de mujeres. Aconsejaría a las mujeres que continúen mejorando sus habilidades.

Reproduce el video desde :2:17 y sigue la transcripción2:17

Si les gusta y si les gusta una carrera repleta de programación, datos y resolución de problemas, entonces deberían continuar desarrollando sus habilidades técnicas, para poder representarse a sí mismas en el panorama de los profesionales de los datos con la mayor fuerza posible. No permitas que tu género sea un obstáculo. Sigue esforzándote, ponte a trabajar y muestra al mundo tus increíbles talentos. No hay roles que se reserven para géneros específicos. Si tienes la suerte de trabajar en una profesión que disfrutas, ve a por ello.

# Resumen y Aspectos Resaltantes

En esta lección, has aprendido la siguiente información:

Los puestos de trabajo de los Analistas de Datos son buscados en todas las industrias, ya sea en la Banca y Finanzas, los Seguros, la Salud, el Comercio Minorista o la Tecnología de la Información.

Actualmente, la demanda de analistas de datos capacitados supera ampliamente la oferta, lo que significa que las empresas están dispuestas a pagar una suma adicional para contratar a los analistas de datos capacitados.

Los puestos de trabajo de los Analistas de Datos pueden clasificarse en general de la siguiente manera:

●      Puestos de Analista de Datos Especializado – En esta ruta, se comienza como Analista de Datos Junior y se asciende al nivel de Analista Principal, avanzando continuamente en tus habilidades técnicas, estadísticas y analíticas desde un nivel básico hasta el nivel experto.

●      Puestos de Especialista en Dominios – Estos puestos son para ti si has adquirido la especialización en un dominio específico y quieres ascender para ser visto como una autoridad en tu dominio.

●      Puestos de Trabajo con Capacidad Analítica – Estos puestos incluyen trabajos en los que tener habilidades analíticas puede elevar tu rendimiento y diferenciarte de tus compañeros.

●      Otras Profesiones de Datos – Hay muchos otros puestos en un ecosistema de datos moderno, tales como los Ingenieros de Datos, Ingeniero de Grandes Volúmenes de Datos, Científico de Datos, Analista de Negocios o Analista de Inteligencia de Negocios. Si te capacitas en base a las habilidades requeridas, puedes hacer la transición a estos puestos de trabajo.

Existen varias rutas que puedes considerar para entrar en el campo de Analista de Datos. Estas incluyen:

●      Un título académico en Análisis de Datos o disciplinas como Estadística y Ciencias de la Computación.

●      Especializaciones multi-curso en línea ofrecidas por las plataformas de aprendizaje como Coursera, edX y Udacity.

●      Transición a mitad de carrera en Análisis de Datos mediante la mejora de sus habilidades. Si tienes una formación técnica, por ejemplo, puedes centrarte en el desarrollo de las habilidades técnicas específicas del Análisis de Datos. Si no tienes una formación técnica, puedes planear capacitarte en algunas tecnologías básicas y luego trabajar para ascender desde un puesto de nivel inicial.

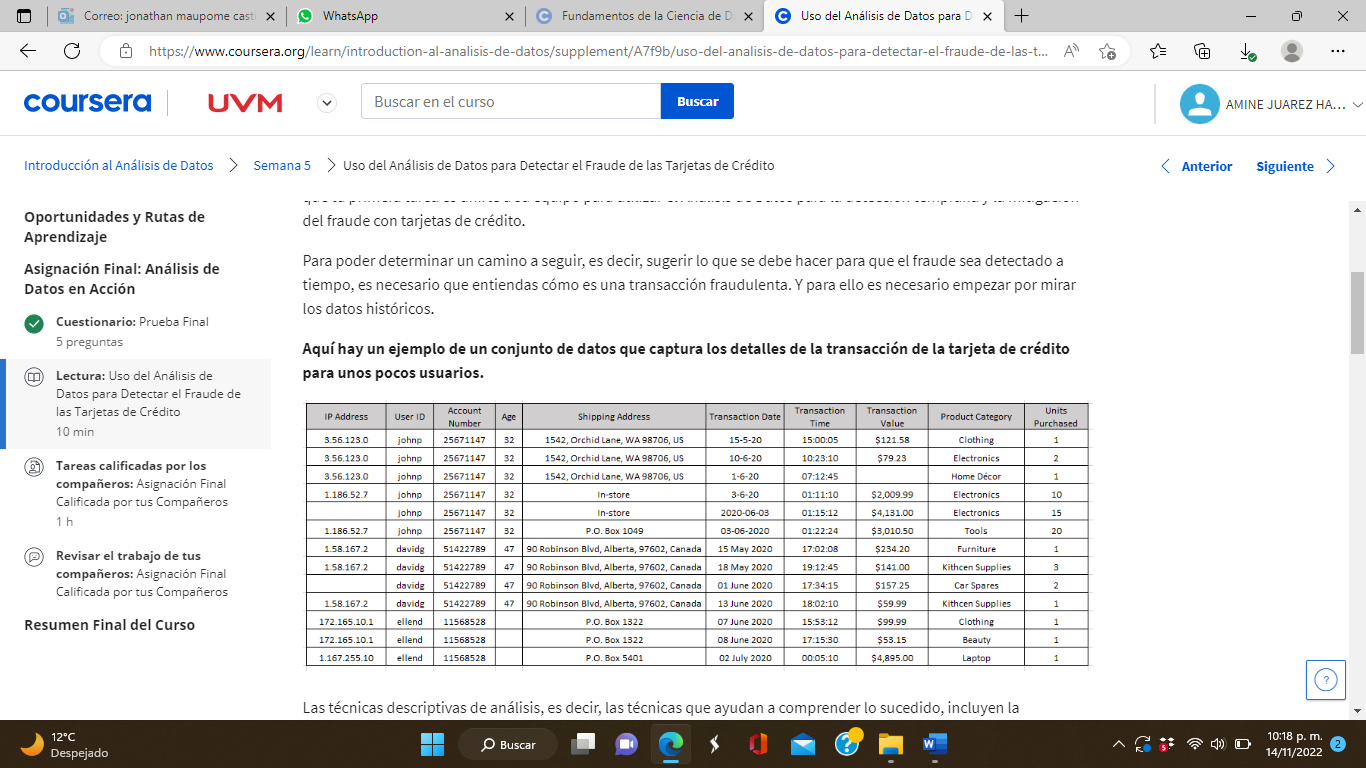
# Uso del Análisis de Datos para Detectar el Fraude de las Tarjetas de Crédito

Las compañías emplean hoy en día técnicas analíticas para la detección temprana de fraudes con tarjetas de crédito, un factor clave para mitigar los daños causados por el fraude. El tipo más común de fraude con tarjetas de crédito no consiste en el robo físico de la tarjeta, sino en el de las credenciales de la tarjeta de crédito, que luego se utilizan para compras en línea.

Imagina que fuiste contratado como Analista de Datos para trabajar en la División de Tarjetas de Crédito de un banco. Y que tu primera tarea es unirte a su equipo para utilizar el Análisis de Datos para la detección temprana y la mitigación del fraude con tarjetas de crédito.

Para poder determinar un camino a seguir, es decir, sugerir lo que se debe hacer para que el fraude sea detectado a tiempo, es necesario que entiendas cómo es una transacción fraudulenta. Y para ello es necesario empezar por mirar los datos históricos.

Aquí hay un ejemplo de un conjunto de datos que captura los detalles de la transacción de la tarjeta de crédito para unos pocos usuarios.



Las técnicas descriptivas de análisis, es decir, las técnicas que ayudan a comprender lo sucedido, incluyen la identificación de patrones y anomalías en los datos. Las anomalías significan una variación en un patrón que parece poco característico o fuera de lo normal. Las anomalías pueden ocurrir por razones perfectamente válidas y genuinas, pero justifican una evaluación porque pueden ser un signo de actividad fraudulenta.

En estudios anteriores se ha sugerido que algunos de los acontecimientos comunes a los que hay que estar atento incluyen:

● Un cambio en la frecuencia de los pedidos realizados, por ejemplo, un cliente que suele hacer un par de pedidos al mes, realiza repentinamente números transacciones en un corto lapso de tiempo, a veces a pocos minutos del pedido anterior.

● Los pedidos que son significativamente más altos que la transacción promedio de un usuario.

● Pedidos al por mayor del mismo artículo con ligeras variaciones como el color o la talla—especialmente si es atípico del historial de transacciones del usuario.

● Un cambio repentino en la preferencia de entrega, por ejemplo, un cambio de la dirección de entrega de la casa u oficina a la entrega en la tienda, el almacén o un apartado postal.

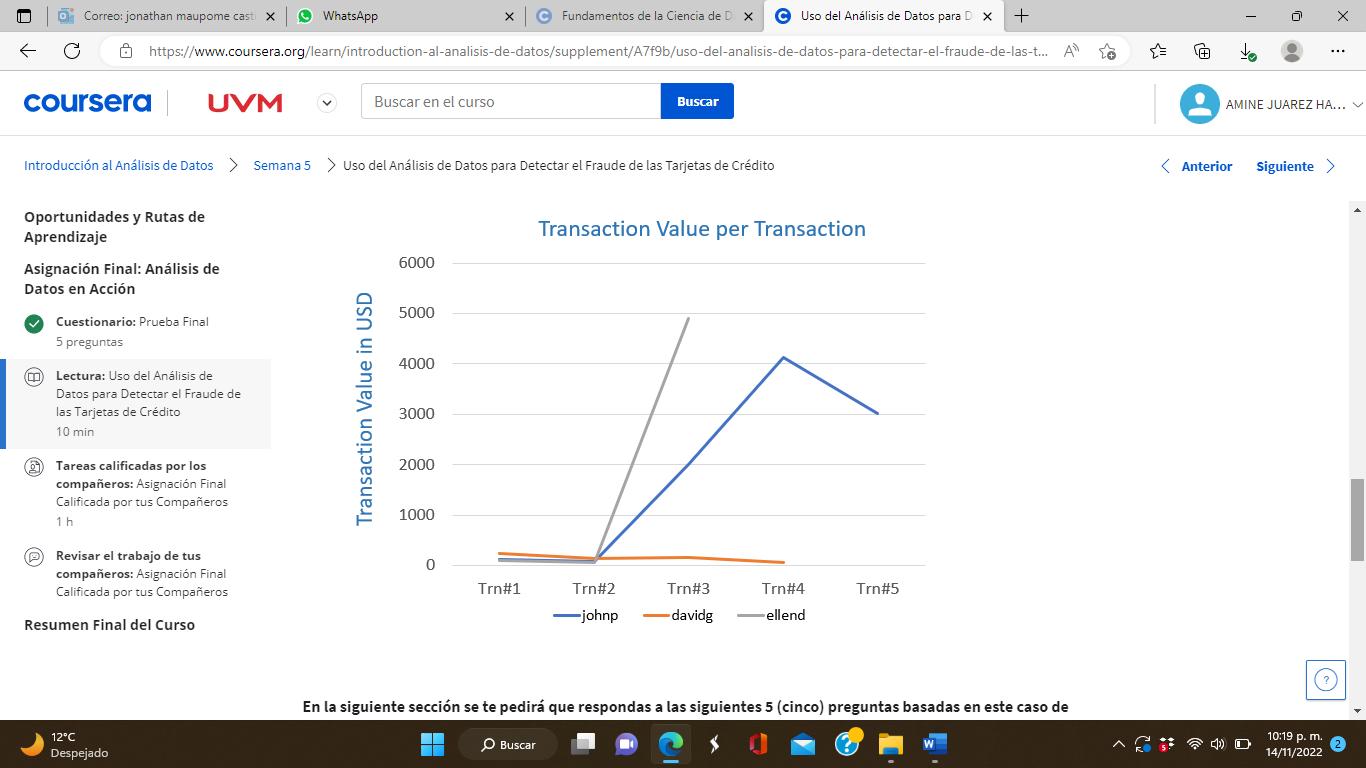
● Una dirección IP que no coincide o una dirección IP que no es del lugar o área de la dirección de facturación.

Antes de poder analizar los datos para detectar patrones y anomalías, es necesario:

● Identificar y reunir todos los puntos de datos que pueden ser relevantes para tu caso de uso. Por ejemplo, los datos del titular de la tarjeta, los detalles de la transacción, los detalles de la entrega, la ubicación y la red son algunos de los puntos de datos que podrían explorarse.

● Limpiar los Datos. Es necesario identificar y corregir los problemas en los datos que pueden conducir a resultados falsos o incompletos, tales como valores de datos faltantes y datos incorrectos. También puede ser necesario normalizar los formatos de los datos en algunos casos, por ejemplo, los campos de fecha.

Finalmente, cuando se llegue a los resultados, se crearán visualizaciones apropiadas que comuniquen los hallazgos a tu público. En el gráfico que figura a continuación se muestra una de esas visualizaciones que se utilizaría para captar una tendencia oculta en el conjunto de datos de muestra compartido anteriormente en el caso de estudio.



En la siguiente sección se te pedirá que respondas a las siguientes 5 (cinco) preguntas basadas en este caso de estudio:

Enumera al menos 5 (cinco) puntos de datos que se requieren para el análisis y la detección de un fraude con tarjetas de crédito. (3 puntos)

Identifica 3 (tres) errores/problemas que podrían afectar a la exactitud de tus hallazgos, basándote en una tabla de datos proporcionada. (3 puntos)

Identifica 2 (dos) anomalías o comportamientos inesperados, que te hagan creer que la transacción puede ser sospechosa, basándote en una tabla de datos proporcionada. (2 puntos)

Explica de forma breve tu clave de partida de la tabla de visualización de datos proporcionada. (1 punto)

Identifica el tipo de análisis que estás realizando cuando analizas los datos históricos de las tarjetas de crédito para comprender cómo es una transacción fraudulenta [Pista: Los cuatro tipos de análisis incluyen: Descriptivo, Diagnóstico, Predictivo, Prescriptivo] (1 punto)