18 September 2023

# Introducción al Big Data

## Sesión 15

José Fuentes

# Objetivo

Aprender sobre Big Data e Ingeniería de los Datos tiene como finalidad adquirir habilidades y conocimientos esenciales para el procesamiento efectivo de grandes volúmenes de datos, permitiendo la toma de decisiones fundamentadas en una variedad de campos.

# Comprender Conceptos Clave de Big Data

- Familiarizarse con conceptos como el almacenamiento distribuido, el procesamiento paralelo, la escalabilidad y la variedad de datos.
- Ejemplo: Entender cómo se almacenan y gestionan grandes cantidades de datos en clústeres de servidores para su posterior análisis.

# Utilizar Herramientas de Análisis de Datos

- Dominar herramientas populares como Hadoop, Spark o Python para el procesamiento y análisis de datos masivos.
- Ejemplo: Realizar un análisis de datos en tiempo real utilizando Apache Kafka y Spark Streaming.

# Aplicar Mejores Prácticas para Sistemas Escalables

- Aprender a diseñar y mantener sistemas que puedan crecer con la demanda de datos.
- Ejemplo: Diseñar una arquitectura de clúster escalable que pueda manejar un aumento repentino en la carga de trabajo.

# Integrar Diferentes Fuentes de Datos

- Saber cómo reunir y consolidar datos de múltiples fuentes para obtener una visión completa.
- Ejemplo: Integrar datos de redes sociales, registros de servidores y transacciones comerciales para analizar el comportamiento del cliente.

# Diseñar Arquitecturas de Datos para Flujos en Tiempo Real

- Crear sistemas capaces de procesar datos en tiempo real y tomar decisiones inmediatas.
- Ejemplo: Desarrollar un sistema de detección de fraudes que analice transacciones bancarias en tiempo real y bloquee actividades sospechosas.

# Desarrollar Habilidades en Visualización y Comunicación de Resultados

- Aprender a presentar de manera efectiva los insights obtenidos a través de visualizaciones y narrativas comprensibles.
- Ejemplo: Crear un tablero interactivo que muestre tendencias y patrones en los datos de ventas a través de gráficos y visualizaciones.

# Recopilación de Datos [0]

- La recopilación de datos implica la obtención de información de diversas fuentes, como sensores, redes sociales, registros empresariales, encuestas, entre otros.
- La recopilación de información es un proceso crítico en la toma de decisiones informadas en diferentes campos y aplicaciones.
- Implica la obtención de datos de diversas fuentes, como encuestas, estudios, experimentos, registros administrativos, redes sociales, entre otros.

# Recopilación de Datos [1]

- El objetivo principal es obtener datos precisos y confiables para análisis posteriores.
- La calidad de los datos es esencial, ya que datos incompletos o inexactos pueden llevar a conclusiones incorrectas y decisiones erróneas.
- Es fundamental utilizar métodos y herramientas efectivas para garantizar la precisión y confiabilidad de los datos.

**Ejemplo:** Una empresa minorista que recopila datos de ventas diarias, datos de inventario y comentarios de clientes en línea.

# Redes Sociales

- Plataformas como Facebook, Twitter e Instagram recopilan datos de sus usuarios, como publicaciones, me gusta, compartidos y comentarios.
- Estos datos se utilizan para analizar tendencias, el comportamiento del usuario y para ofrecer publicidad dirigida.
- **Ejemplo:** Facebook recopila datos de interacción de sus usuarios para personalizar el contenido del feed de noticias y los anuncios que se muestran.

# Internet de las Cosas (IoT)

- Dispositivos como sensores de temperatura, cámaras de seguridad y medidores inteligentes generan grandes volúmenes de datos en tiempo real.
- Estos datos se utilizan para monitorear y controlar sistemas, como edificios inteligentes o redes eléctricas.
- **Ejemplo:** Una empresa de energía utiliza medidores inteligentes para recopilar datos de consumo de energía de sus clientes y ajustar la distribución de energía de manera eficiente.

# Comercio Electrónico

- Plataformas de comercio electrónico como Amazon y Alibaba recopilan datos de las interacciones de los usuarios, como búsquedas, compras y revisiones de productos.
- Estos datos se utilizan para recomendar productos, mejorar la experiencia del usuario y optimizar la cadena de suministro.
- **Ejemplo:** Amazon utiliza datos de compra y navegación para ofrecer recomendaciones de productos personalizadas a sus usuarios.

# Ciencia Médica

- Los hospitales y las instituciones de investigación médica recopilan datos de pacientes, incluidos registros médicos electrónicos, resultados de pruebas y datos genómicos.
- Estos datos se utilizan para la investigación médica, el diagnóstico y el desarrollo de tratamientos.
- **Ejemplo:** Un hospital recopila datos de pacientes con diabetes para identificar patrones de tratamiento efectivos y mejorar la atención médica.

# Transporte y Logística

- Las empresas de transporte recopilan datos de seguimiento de vehículos, sensores de carga y datos de tráfico en tiempo real.
- Estos datos se utilizan para optimizar rutas, programar mantenimiento preventivo y mejorar la eficiencia de la cadena de suministro.
- **Ejemplo:** Una empresa de entregas utiliza datos de seguimiento de vehículos y datos de tráfico en tiempo real para evitar retrasos en las entregas.

# Medios de Comunicación y Entretenimiento

- Las plataformas de transmisión en línea recopilan datos de visualización de videos, preferencias de contenido y comentarios de los espectadores.
- Estos datos se utilizan para personalizar recomendaciones de contenido y mejorar la experiencia del usuario.
- **Ejemplo: Netflix** recopila datos de visualización para sugerir programas y películas que puedan interesar a sus usuarios.

# Herramientas de Recopilación de Información

Existen diversas herramientas de recopilación de información disponibles en la actualidad que pueden utilizarse para obtener datos de diferentes fuentes. Estas herramientas varían en complejidad y funcionalidad, y se utilizan comúnmente en campos como la investigación científica, los negocios y la política pública.

# Herramientas de Recopilación de Información

Algunas de las herramientas de recopilación de información más comunes incluyen:

- Encuestas en línea
- Entrevistas
- Grupos focales
- Observación
- Análisis de registros administrativos
- Escaneo de documentos
- Minería de datos en redes sociales

# Herramientas de Recopilación de Información

Cada una de estas herramientas tiene ventajas y desventajas, y su elección dependerá del tipo de información que se desea recopilar, la cantidad de datos necesarios y el presupuesto disponible.

Estas herramientas automatizadas pueden ahorrar tiempo y reducir errores humanos, pero también pueden requerir habilidades técnicas avanzadas para su uso efectivo.

# Herramientas de Recopilación de Información

Además, las herramientas de recopilación de información también pueden ser automatizadas. Algunos ejemplos de herramientas automatizadas son:

- Chatbots que recopilan información en línea
- Automatización de procesos robóticos
- Otras herramientas que permiten la recopilación de información a gran escala y en tiempo real

# Encuestas en línea

Herramientas como SurveyMonkey, Google Forms y Qualtrics permiten crear encuestas en línea y recopilar datos de un gran número de encuestados en diferentes ubicaciones geográficas.

Ejemplo: Una empresa de marketing utiliza encuestas en línea para recopilar datos sobre la satisfacción del cliente y las preferencias de compra.

# Minería de Datos en Redes Sociales

Herramientas de análisis de redes sociales como Brandwatch, Hootsuite Insights y Mention recopilan datos de redes sociales para analizar tendencias, sentimiento y participación de la audiencia.

Ejemplo: Una empresa de medios utiliza herramientas de minería de datos en redes sociales para evaluar la reacción del público a sus programas de televisión y ajustar su estrategia de contenido en consecuencia.

# Chatbots

Los chatbots automatizados, como Chatfuel y Dialogflow, recopilan información en línea a través de conversaciones con usuarios en sitios web o aplicaciones.

Ejemplo: Un banco utiliza un chatbot para recopilar información sobre las necesidades de sus clientes y proporcionar respuestas automáticas a preguntas frecuentes.

# Automatización de Procesos Robóticos (RPA)

Las herramientas de RPA, como UiPath y Automation Anywhere, automatizan tareas repetitivas y recopilan datos de sistemas y aplicaciones empresariales.

Ejemplo: Una empresa de recursos humanos utiliza RPA para recopilar y procesar automáticamente currículos de candidatos de correo electrónico y aplicaciones en línea.

# Análisis de Registros Administrativos

En el ámbito gubernamental, se utilizan sistemas de registros administrativos para recopilar datos sobre impuestos, salud pública, educación y más.

Ejemplo: Un gobierno municipal recopila datos de registros administrativos para analizar patrones de tráfico y tomar decisiones informadas sobre la planificación de infraestructuras de transporte.

# Escaneo de Documentos

Herramientas de escaneo y reconocimiento óptico de caracteres (OCR), como Adobe Acrobat y ABBYY FineReader, digitalizan y recopilan datos a partir de documentos impresos.

Ejemplo: Una biblioteca universitaria utiliza OCR para recopilar datos de libros y artículos académicos y crear una base de datos digital.

# Observación

La observación directa de eventos o comportamientos es una herramienta de recopilación de datos en investigación científica y estudios de mercado.

Ejemplo: Un investigador observa y registra el comportamiento de las aves migratorias para estudiar sus patrones de migración.

# Entrevistas y Grupos Focales

Se utilizan en investigación cualitativa para recopilar datos a través de conversaciones estructuradas o discusiones grupales.

Ejemplo: Un equipo de investigación de mercado realiza entrevistas en profundidad y grupos focales con consumidores para comprender sus preferencias de productos.

# Limpieza de los Datos

Es un proceso crítico en el análisis de datos, que implica la eliminación de errores, duplicados, inconsistencias y valores atípicos que puedan afectar la calidad y la precisión de los datos.

La limpieza de datos es esencial para garantizar que los análisis y modelos posteriores se basen en datos precisos y confiables.

# Proceso de limpieza de datos

El proceso de limpieza de datos puede ser manual o automatizado, y generalmente implica varias etapas, como la detección de valores atípicos y la eliminación de duplicados.

La limpieza de datos también puede requerir la imputación de valores faltantes o la corrección de errores de registro, que pueden afectar la calidad y la precisión de los datos.

# Importancia de la limpieza de los datos

Es importante destacar que la limpieza de los datos es un proceso continuo y debe realizarse en cada etapa del análisis de datos.
La calidad de los datos puede afectar significativamente los resultados del análisis, y la falta de limpieza de los datos puede llevar a conclusiones erróneas y decisiones incorrectas.

# Eliminación de Valores Atípicos

Ejemplo: En un conjunto de datos de registros de transacciones financieras, se detectan valores atípicos que podrían ser errores de entrada. Estos valores deben identificarse y eliminarse antes de realizar análisis financieros.

Ejemplo: En un conjunto de datos de registros de transacciones financieras, se detectan valores atípicos que podrían ser errores de entrada. Estos valores deben identificarse y eliminarse antes de realizar análisis financieros.

# Deduplicación de Datos

Ejemplo: En una base de datos de clientes de una empresa, se encuentran registros duplicados debido a errores en la entrada de datos. La limpieza implica identificar y fusionar estos registros duplicados en un solo registro por cliente.

Ejemplo: En una base de datos de clientes de una empresa, se encuentran registros duplicados debido a errores en la entrada de datos. La limpieza implica identificar y fusionar estos registros duplicados en un solo registro por cliente.

# Imputación de Valores Faltantes

Ejemplo: En un conjunto de datos de encuestas en línea, algunos encuestados pueden no haber respondido todas las preguntas. Se debe realizar una imputación de valores para estimar respuestas faltantes y garantizar que los análisis posteriores sean representativos.

# Normalización de Datos

Ejemplo: En una base de datos de direcciones postales, se pueden encontrar variaciones en la forma en que se ingresan las direcciones (por ejemplo, "Calle", "C.", "Cll"). La limpieza implica normalizar estas variaciones para garantizar la coherencia de los datos.

# Corrección de Errores de Formato

Ejemplo: En un conjunto de datos de números de teléfono, algunos números pueden estar en formatos inconsistentes (por ejemplo, "123-456-7890" vs. "(123) 456-7890"). La limpieza implica estandarizar el formato para facilitar su análisis.

# Validación de Datos

Ejemplo: En un conjunto de datos de registros médicos, se pueden encontrar fechas de nacimiento que son inválidas (por ejemplo, fechas futuras o incorrectas). La limpieza implicaverificar la validez de las fechas y corregirlas si es necesario.

# Eliminación de Datos Irrelevantes

Ejemplo: En un conjunto de datos de registro de actividad del sitio web, puede haber información redundante o no relevante para el análisis. La limpieza implica eliminar estos datos para reducir la carga de procesamiento.

# Consistencia de Categorías

Ejemplo: En un conjunto de datos de productos de comercio electrónico, las categorías de productos pueden tener nombres similares pero ligeramente diferentes. La limpieza implica agrupar categorías similares bajo una etiqueta común.

# Detección y Tratamiento de Datos Inconsistentes

Ejemplo: En una base de datos de registros de pacientes, puede haber inconsistencias en la información demográfica, como la edad declarada que no coincide con la fecha de nacimiento. La limpieza implica resolver estas inconsistencias.

# Eliminación de Datos Sensibles o Privados

Ejemplo: En el contexto de la privacidad de los datos, es necesario eliminar o anonimizar información personalmente identificable (PII) de los conjuntos de datos antes de realizar análisis de Big Data para cumplir con las regulaciones de privacidad.

# Procesamiento y Análisis de Datos

- El procesamiento y análisis de datos se refiere a la transformación de los datos crudos en información significativa y útil.
- Ejemplo: Un banco que analiza los patrones de gastos de sus clientes para identificar transacciones inusuales que puedan ser indicativas de fraude.

# Procesamiento de la Información

El procesamiento de la información implica la transformación de los datos crudos en una forma utilizable para el análisis posterior. Algunas tareas comunes en el procesamiento de la información incluyen:

- Agregación de datos.
- Normalización de variables.
- Eliminación de valores atípicos.
- Imputación de valores faltantes.

Estas etapas de procesamiento son fundamentales para garantizar la calidad y la consistencia de los datos antes de realizar cualquier análisis adicional.

# Exploración de la Información

La exploración de la información implica el uso de herramientas y técnicas estadísticas para comprender y visualizar los datos. Algunas actividades comunes en la exploración de la información son:

- Utilizar herramientas de visualización de datos, como gráficos y tablas.
- Realizar análisis de correlación y regresión.
- Aplicar técnicas de agrupamiento y clasificación.
- Realizar análisis de componentes principales.

La exploración de la información nos permite identificar patrones, relaciones y tendencias relevantes en los datos, y generar hipótesis y preguntas adicionales para el análisis posterior.

# Iteratividad y Continuidad

Es importante destacar que el procesamiento y la exploración de la información son procesos iterativos y continuos. Esto significa que pueden requerir ajustes y modificaciones en función de los resultados y descubrimientos obtenidos en el análisis de datos posterior.

La calidad y la precisión de los resultados del análisis de datos dependen en gran medida de la calidad y la precisión del procesamiento y la exploración de la información. Por lo tanto, es fundamental dedicar tiempo y esfuerzo a estas etapas del proceso de análisis de datos.

# Crecimiento Exponencial de Datos

- La cantidad de datos disponibles está aumentando rápidamente debido a la digitalización y la proliferación de dispositivos conectados.
- Ejemplo: La cantidad de datos generados por dispositivos IoT (Internet de las cosas) como sensores de temperatura, cámaras de seguridad y dispositivos de seguimiento de fitness.

# Desafíos y Oportunidades

- El crecimiento de datos crea desafíos, como la gestión de la privacidad y la seguridad de los datos, pero también oportunidades para obtener información valiosa.
- Ejemplo: Una empresa de medios sociales que enfrenta desafíos para proteger la privacidad de los usuarios mientras analiza datos para mejorar la experiencia del usuario y orientar la publicidad de manera efectiva.

# Procesos Clave

- La gestión de datos implica una serie de procesos clave, desde la adquisición y limpieza de datos hasta la visualización de resultados.
- Ejemplo: Un equipo de científicos de datos que trabaja en un proyecto de investigación que implica la limpieza y la normalización de datos de encuestas antes de realizar análisis estadísticos.

# Herramientas y Técnicas Avanzadas

- Para abordar estos desafíos y oportunidades, se requiere el uso de herramientas y técnicas avanzadas, como algoritmos de aprendizaje automático, análisis de big data y visualización de datos.
- Ejemplo: Una empresa de análisis financiero que utiliza algoritmos de aprendizaje automático para predecir el rendimiento del mercado de valores.

# Visualización de Datos

La visualización de los datos implica la creación de gráficos, tablas y otros elementos visuales para representar los patrones y tendencias identificados en los datos. Esto ayuda a los analistas a identificar patrones y relaciones que no son evidentes en los datos crudos. Además, facilita la comprensión de los resultados del análisis por parte de los usuarios.

# Comunicación de Resultados

La comunicación de los resultados implica presentar los hallazgos del análisis de manera clara y accesible a una audiencia específica. Esto puede incluir la creación de informes y presentaciones que resuman los resultados en términos de hallazgos clave, conclusiones y recomendaciones. La adaptación del mensaje a la audiencia es crucial para garantizar una comunicación efectiva.

# Iteración y Ajustes

Es importante destacar que la visualización y la comunicación son etapas iterativas y continuas en el análisis de datos. Los comentarios y preguntas de la audiencia pueden requerir ajustes y modificaciones en la presentación de los resultados. La visualización y la comunicación efectivas son fundamentales para asegurar que los resultados del análisis sean comprensibles y accionables.

# Herramientas de Visualización

Las herramientas de visualización son conjuntos de herramientas y técnicas utilizadas para representar gráficamente los datos de manera clara y comprensible. Estas herramientas permiten a los analistas y usuarios explorar y comprender los datos de manera más efectiva, lo que ayuda en la identificación de patrones y tendencias importantes.

# Tipos de Herramientas

Existen diversas herramientas de visualización de datos disponibles, desde hojas de cálculo y software de gráficos básicos hasta herramientas especializadas en visualización interactiva, como Tableau, Power BI y QlikView. Estas herramientas permiten a los usuarios crear gráficos, mapas y otros elementos visuales para explorar y comprender mejor los datos.

# Interactividad de las Herramientas

Las herramientas de visualización también permiten a los usuarios interactuar con los datos de manera más efectiva, mediante la creación de gráficos y paneles interactivos que les permiten explorar diferentes aspectos de los datos y ver cómo los cambios en los datos afectan los resultados.

# Selección de Herramientas

Es importante destacar que la selección de las herramientas de visualización adecuadas depende de los requisitos específicos de análisis de datos y de las habilidades y experiencia del usuario. Los usuarios deben elegir herramientas que se adapten a sus necesidades de análisis de datos y que puedan ser utilizadas de manera efectiva y eficiente.

# Monitoreo y Evaluación

- Seguimiento del rendimiento del sistema: El monitoreo y evaluación en la ingeniería de datos implica supervisar de cerca el rendimiento del sistema de gestión de datos. Esto implica el seguimiento de métricas y parámetros clave, como el tiempo de respuesta, el rendimiento del almacenamiento, el rendimiento de consultas y la disponibilidad del sistema. El monitoreo continuo ayuda a identificar posibles cuellos de botella, problemas de rendimiento o fallas en el sistema.

# Monitoreo y Evaluación

- Detección y solución de problemas: El monitoreo constante permite detectar y solucionar problemas en el sistema de ingeniería de datos de manera temprana. Si se presentan errores, retrasos o anomalías, el monitoreo puede alertar a los ingenieros de datos para que tomen medidas correctivas rápidas y minimicen los impactos negativos en la integridad y disponibilidad de los datos.

# Monitoreo y Evaluación

- Optimización del rendimiento: El monitoreo y evaluación también brinda información valiosa para optimizar el rendimiento del sistema de ingeniería de datos. Al analizar los datos recopilados durante el monitoreo, los ingenieros pueden identificar áreas de mejora y tomar decisiones informadas para optimizar el rendimiento del sistema. Esto puede incluir ajustar la configuración del hardware, mejorar el diseño de bases de datos, optimizar consultas y mejorar los procesos de extracción, transformación y carga (ETL).

# Monitoreo y Evaluación

- Evaluación de resultados y cumplimiento de objetivos: El monitoreo y evaluación también se utiliza para evaluar si el sistema de ingeniería de datos está cumpliendo con los objetivos establecidos. Se pueden establecer métricas de rendimiento y objetivos específicos relacionados con la calidad de los datos, el tiempo de procesamiento, la eficiencia o la precisión del análisis. El monitoreo continuo permite evaluar si se están alcanzando estos objetivos y, si es necesario, realizar ajustes para mejorar el rendimiento del sistema.

# Monitoreo y Evaluación

- Toma de decisiones basada en datos: El monitoreo y evaluación proporciona información en tiempo real sobre el estado y el rendimiento del sistema de ingeniería de datos. Esta información es fundamental para la toma de decisiones basada en datos en la organización. Los informes y análisis generados a partir del monitoreo pueden proporcionar información valiosa para la toma de decisiones estratégicas, como la asignación de recursos, la planificación de proyectos futuros o la identificación de áreas de mejora.

# Monitoreo y Evaluación

En resumen, el proceso de monitoreo y evaluación en la ingeniería de datos sirve para supervisar, detectar y solucionar problemas, optimizar el rendimiento del sistema, evaluar el cumplimiento de objetivos y proporcionar información para la toma de decisiones basada en datos.

# Herramientas para el Monitoreo y Evaluación

- Software de seguimiento de proyectos: Estas herramientas permiten recopilar y analizar datos de manera eficiente y efectiva para el monitoreo y evaluación de proyectos. Proporcionan funcionalidades para el registro de actividades, seguimiento de hitos, asignación de recursos y generación de informes de progreso.

# Herramientas para el Monitoreo y Evaluación

- Herramientas de encuestas en línea: Estas herramientas facilitan la recopilación de datos a través de encuestas en línea. Permiten diseñar cuestionarios personalizados, enviar invitaciones a los participantes y recopilar respuestas de manera automatizada. Además, ofrecen funciones de análisis y generación de informes.

# Herramientas para el Monitoreo y Evaluación

- Sistemas de información geográfica (SIG): Estas herramientas permiten visualizar y analizar datos geográficos para el monitoreo y evaluación. Proporcionan funcionalidades para la captura, almacenamiento, gestión, análisis y visualización de datos espaciales. Son útiles para proyectos que requieren análisis basados en la ubicación geográfica.

# Herramientas para el Monitoreo y Evaluación

- Herramientas de análisis de datos: Estas herramientas son utilizadas para procesar y explorar datos en el monitoreo y evaluación. Proporcionan funcionalidades para la limpieza, transformación y análisis de datos. Permiten realizar cálculos, aplicar algoritmos y generar visualizaciones para obtener conclusiones informadas.

# Herramientas para el Monitoreo y Evaluación

- Herramientas de visualización y comunicación de datos: Estas herramientas son importantes para asegurar que los resultados del monitoreo y evaluación sean comprensibles para diferentes audiencias. Permiten crear gráficos, tablas, mapas y otros elementos visuales para presentar la información de manera clara y concisa.

# Herramientas para el Monitoreo y Evaluación

En resumen, existen diversas herramientas disponibles para el monitoreo y evaluación, como software de seguimiento de proyectos, herramientas de encuestas en línea, sistemas de información geográfica, herramientas de análisis de datos y herramientas de visualización y comunicación de datos. Estas herramientas permiten recopilar y analizar datos de manera eficiente y efectiva, y aseguran el éxito continuo de un proyecto o programa.