«Talento Tech»

Business Intelligence

Clase 04

Buenos Aires Ciudad

Clase N° 04 | Calidad y Seguridad de los Datos

Temario:

- Definición de calidad de datos
- Técnicas para asegurar la calidad de los datos. Normalización.
- Prácticas de seguridad de datos y acceso controlado

Objetivos de la clase:

En esta clase, analizaremos la importancia de la calidad de los datos y su impacto en la toma de decisiones empresariales. Exploraremos técnicas para garantizar datos precisos y confiables, incluyendo estrategias de normalización. Además, abordaremos prácticas de seguridad y control de acceso para proteger la información sensible. Finalmente, destacaremos la necesidad de una formación continua del personal en estos temas, con el fin de minimizar errores y fortalecer la seguridad de los datos en las organizaciones.

Calidad de Datos

La calidad de los datos se refiere a la **adecuación de los datos para cumplir con los propósitos para los cuales fueron recopilados**. Esta calidad se evalúa a través de varias dimensiones, por ejemplo, precisión, consistencia, completitud, actualidad y relevancia. Los **datos de alta calidad** son fundamentales para la toma de decisiones informadas, ya que datos imprecisos o incompletos pueden conducir a conclusiones erróneas y afectar negativamente a las estrategias empresariales.

Por lo tanto, **es esencial que las organizaciones establezcan procesos e indicadores claros para evaluar y controlar la calidad de los datos que manejan**. Un ejemplo práctico de esto es el uso de bases de datos limpias en un sistema de gestión de relaciones con clientes, donde solo la información precisa de los clientes se utiliza para personalizar ofertas y mejorar la satisfacción del cliente.

Revisá el siguiente ejemplo de formulario de Evaluación de Calidad de Datos.pdf

Técnicas para Asegurar la Calidad de los Datos

Para asegurar la calidad de los datos existen diversas **técnicas** que pueden ser implementadas. Una de las más efectivas es la **validación de datos**, que implica establecer reglas y restricciones al momento de la entrada de datos, asegurando que la información cumpla con los criterios necesarios. Otra técnica es el **muestreo y auditoría de datos**, donde se revisan muestras de información para identificar problemas en la base de datos. La **normalización de datos** es otra técnica importante, ya que permite estandarizar la información y reducir duplicidades. Por último, la **formación y concienciación del personal** acerca de la importancia de la calidad de los datos es crucial; cuando los empleados son conscientes de cómo sus acciones impactan la calidad, es más probable que se comprometan con mantener estándares altos. Por ejemplo, una empresa que capacita a sus empleados en la correcta entrada de datos puede reducir significativamente los errores que afectan la calidad de su base de datos.

Normalización de datos

La normalización de datos es un proceso fundamental en el diseño de bases de datos que busca organizar los datos para reducir la redundancia y mejorar la integridad de la información. Este proceso implica dividir grandes tablas en tablas más pequeñas y definir relaciones entre ellas, lo que fomenta un esquema de almacenamiento estructurado y eficiente. **La normalización se logra a través de una serie de reglas estructuradas**, conocidas como "formas normales" (normal forms o NF), que establecen criterios específicos para la organización de los datos.

Hay varias formas normales, desde la primera forma normal (1NF) hasta la tercera forma normal (3NF) y más allá, existen formas normales superiores. A medida que se avanza en estas formas, se tiende a eliminar más redundancias y dependencias anómalas. Por ejemplo, una base de datos en 1NF asegura que todos los atributos en una tabla contengan valores atómicos, mientras que en 2NF se eliminan las dependencias parciales, y en 3NF se eliminan las dependencias transitivas.

A. Primera Forma Normal (1NF)

La **Primera Forma Normal (1NF)** establece que una tabla debe contener solo valores atómicos, es decir, cada columna de la tabla debe almacenar indivisible y exclusivamente un único valor. Además, cada fila de la tabla debe ser única, lo que implica que debe haber una clave primaria que distinga cada registro de manera singular. Para cumplir con la 1NF, se debe evitar la inclusión de columnas que contengan múltiples valores o listas, asegurando que cada elemento de datos esté en su propia celda. Por ejemplo, si una tabla de “Estudiantes” incluye una columna para “Cursos” donde se almacenen múltiples cursos en una sola celda (como "Matemáticas, Ciencias"), esto no es válido en 1NF y debe reorganizarse para que cada curso tenga su propia fila o columna.

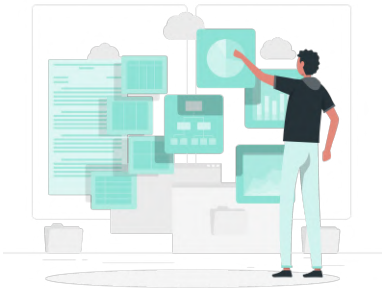
B. Segunda Forma Normal (2NF)

La **Segunda Forma Normal (2NF)** se aplica a tablas que ya cumplen con la 1NF y se centra en eliminar las dependencias parciales, donde un atributo no clave depende solo de una parte de una clave primaria compuesta. En otras palabras, todos los atributos de la tabla deben depender completamente de la clave primaria y no de solo una parte de ella. Para lograr esto, se deben separar los datos en diferentes tablas donde las dependencias se mantengan completas. Por ejemplo, si tenemos una tabla "Pedido" que incluye tanto el número de pedido como la información del cliente, y la información del cliente depende solo de la parte de la clave primaria relacionada con el cliente, deberíamos mover los datos del cliente a una tabla separada, asegurando que la información del pedido y del cliente estén completamente normalizadas.

C. Tercera Forma Normal (3NF)

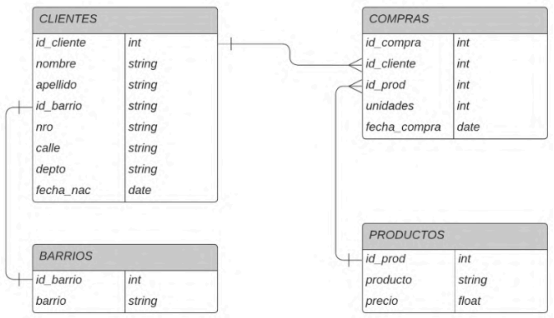
La **Tercera Forma Normal (3NF)** se centra en eliminar las dependencias transitivas, asegurando que los atributos de una tabla no dependan de otros atributos no clave. Es decir, si un atributo A determina a otro atributo B, este último debe ser dependiente únicamente de la clave primaria y no de otro atributo. Para cumplir con la 3NF, se deben descomponer las tablas de manera que cada atributo sea directamente dependiente solo de la clave primaria. Por ejemplo, si en una tabla de "Estudiantes" se incluye tanto el "ID de Estudiante" como el "Nombre del Curso" y el "Instructor del Curso", y el "Instructor del Curso" depende del "Nombre del Curso", estaríamos ante una dependencia transitiva. Para normalizar, es necesario separar esta información en dos tablas: una para "Cursos" (que incluye el "Nombre del Curso" y el "Instructor") y otra para "Estudiantes". Esto asegura que todos los atributos sean directamente dependientes de la clave primaria.

Modelo Entidad-Relación y su Relación con Bases de Datos Relacionales



El **modelo entidad-relación (ER)** es una herramienta conceptual utilizada en la etapa de **diseño de bases de datos**. Representa las entidades (es decir, objetos o conceptos relevantes en el dominio del problema) y las relaciones entre ellas. En el contexto de la programación orientada a objetos, las "entidades" se pueden mapear directamente a "clases", y las "relaciones" se asemejan a las asociaciones entre estas clases.

El **Lenguaje de Modelado Unificado** (UML) es una forma estandarizada de visualizar el diseño de sistemas. Los diagramas de clases en UML contienen conceptos similares a los diagramas ER, donde las clases se representan como rectángulos y las relaciones como líneas que conectan estas clases. Esta representación facilita la comprensión y la comunicación de la estructura de datos en un sistema.



Relación con Bases de Datos Relacionales

Las **bases de datos relacionales** utilizan el modelo relacional para almacenar los datos en tablas (o relaciones). Cada tabla representa una entidad y cada fila en la tabla representa una instancia de esa entidad. Las columnas representan los atributos de esa entidad. La normalización es, por lo tanto, esencial para garantizar que estas tablas se diseñen de manera eficiente, con el fin de evitar problemas como la duplicación de datos y las anomalías de actualización.

Por ejemplo, si tenemos una tabla "Clientes" y otra "Órdenes", en una base de datos correctamente normalizada, no se almacenará la información del cliente en cada fila de la tabla de órdenes; en cambio, habría una relación entre ambas tablas, a menudo implementada mediante claves primarias (PK) y claves foráneas (FK). Este enfoque asegura que, si un cliente cambia su información, solo se necesite actualizarla en un lugar.

La normalización de datos es un componente esencial del diseño de bases de datos que directa o indirectamente se relaciona con conceptos de programación orientada a objetos y modelos de diseño como UML. A través de un enfoque sistemático hacia la organización de datos, se maximiza la integridad y la eficiencia de las bases de datos relacionales, lo que a su vez contribuye significativamente a la calidad de los análisis de datos.

Materiales y recursos adicionales

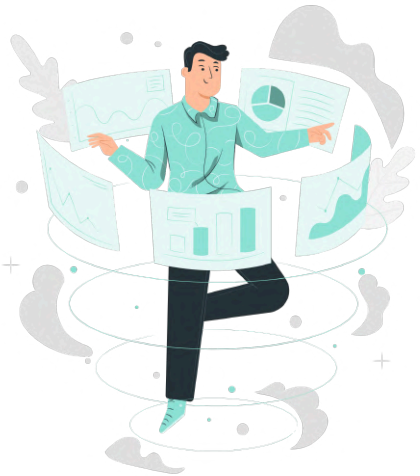
- Ejemplo de normalización de una tabla de Google Sheets
- Ejemplo de formulario de Evaluación de Calidad de Datos.pdf

Prácticas de Seguridad de Datos y Acceso Controlado

Las prácticas de seguridad de datos son fundamentales para **proteger la información sensible de las organizaciones**. La información debe ser visible solamente para los involucrados, evitándose lo más posible que los datos sean expuestos. Las prácticas de seguridad incluyen el uso de filtros, cifrado (que transforma los datos en un formato ilegible para proteger información confidencial), anonimización de datos, así como la implementación de controles de acceso estrictos, asegurando que solo el personal autorizado pueda acceder a ciertos datos. Además, es importante establecer **políticas de gestión de datos**, que definan cómo se deben almacenar, transferir y destruir los datos. Un ejemplo ilustrativo de buenas prácticas de seguridad es el uso de contraseñas robustas y la autenticación multifactor, que añaden capas adicionales de seguridad. También es esencial la **formación continua en materia de seguridad** para todos los empleados, ya que la mayoría de las brechas de seguridad son el resultado de errores humanos. Adicionalmente, llevar a cabo **auditorías regulares de seguridad** puede ayudar a identificar y solucionar vulnerabilidades antes de que sean explotadas.

Reflexión final

La **calidad de los datos** y la **seguridad** son pilares fundamentales en el ámbito de la inteligencia empresarial. A medida que las organizaciones se vuelven cada vez más dependientes de los datos para tomar decisiones estratégicas, se necesita no sólo contar con información precisa y actualizada, sino también garantizar que esta esté protegida contra accesos no autorizados y brechas de seguridad. Implementar **técnicas efectivas** para asegurar la calidad de los datos, junto con **prácticas robustas de seguridad y acceso controlado**, no solo mejora la confianza en la información disponible, sino que también fortalece la integridad y la reputación de la organización ante sus clientes y socios.



Además, la formación continua del personal es esencial, ya que es el primer escudo contra errores y vulnerabilidades. En conjunto, estos elementos crean un entorno en el que los datos, como activos valiosos, pueden ser utilizados de manera responsable y eficiente, impulsando a las empresas hacia el éxito en un entorno cada vez más competitivo.

Próximos pasos

- Introducción a la preparación de datos
- Limpieza de datos en Google Sheets

Ejercicios prácticos:

DATAWISE CONSULTING
INSIGHTS ESTRATÉGICOS

Actividad 1: Normalización de bases de datos

Contexto

Llegás a la oficina de DataWise Consulting. Silvia, la Project Manager y especialista en datos, te está guiando en tu primera asignación de la semana. Se trata de normalización de datos, un paso crucial para garantizar que la información sea precisa y útil en el análisis. Es la base sobre la cual se construirán informes y decisiones estratégicas.

Objetivos

El objetivo de esta actividad es normalizar los datos de una planilla de Google Sheets. Aprenderás a estructurar y organizar la información de manera que sea más fácil de analizar y manipular.

Ejercicio Práctico

Recibís una planilla de Google Sheets con datos sin formatear de diferentes fuentes. Tu tarea es normalizar estos datos: asegurarte de que los formatos sean

consistentes y eliminar cualquier duplicado o error. Esto podría incluir estandarizar nombres, corregir errores tipográficos o definir formatos para fechas y números.

Sets de Datos

- Actividad 1 - normalización Planilla de Google Sheets con datos sin normalizar.

¿Por qué importa esto en DataWise Consulting?

La normalización es clave en DataWise Consulting, ya que asegura que todos los datos sean coherentes y precisos. Esto permite que Luis y Matías trabajen con información confiable, lo que se traduce en mejores decisiones y soluciones más efectivas para los clientes.

Actividad 2: Criterios de Calidad de los Datos

Contexto

Luego de aprender sobre normalización, es el momento de profundizar en la calidad de los datos. Sabrina, consultora estratégica de la startup, se acerca a vos para que la ayudes en la evaluación de un nuevo conjunto de datos. La calidad de la información es fundamental para que las decisiones que tomen sean efectivas y alineadas con las necesidades del cliente.

Objetivos

El objetivo de esta actividad es importar un conjunto de datos a Google Sheets y aplicar criterios de calidad para asegurarte de que los datos sean confiables. Además, crearás un informe que resuma tus hallazgos.

Ejercicio Práctico

1. Importá un conjunto de datos en Google Sheets.
2. Creá un checklist con criterios de calidad de datos. Aplicá ese checklist al conjunto de datos y registrá los hallazgos.
3. Agregá validación de datos a una de las columnas.
4. Creá una vista controlada de la tabla.
5. Presentá un breve informe de calidad, detallando los criterios y los cambios aplicados.
6. **Recordá que siempre debés preservar intacta una copia de la fuente de datos original.**

Sets de Datos

- Conjunto de datos para importar. (Kaggle) Cafe Sales - Dirty Data for Cleaning Training
- Formulario con checklist de criterios de calidad. (Podés crear uno parecido al del ejemplo Evaluación de Calidad de Datos.pdf)

¿Por qué importa esto en DataWise Consulting?
La calidad de los datos es vital para el éxito de DataWise Consulting. Cada decisión estratégica que toman Silvia, Luis, Matías y Sabrina se basa en datos de calidad. Al garantizar que los datos sean precisos y fiables, contribuyes a que la empresa pueda ofrecer soluciones más efectivas y personalizadas a sus clientes.

Actividad 3: Formularios de Google y Datos Sintéticos

Contexto

Con la experiencia adquirida en normalización y criterios de calidad, es hora de aplicar tus habilidades en una actividad más creativa y técnica. Matías, el desarrollador de BI, te propone crear un formulario de Google que recopile información sobre cursos que ofrece la empresa. Esta será una excelente oportunidad para combinar tus conocimientos técnicos con la recolección de datos valiosos.

Objetivos

El objetivo de esta actividad es diseñar un formulario en Google que permita recopilar información detallada sobre los cursos, y luego generar un set de datos sintéticos en formato CSV para simular cómo se cargarían 600 registros.

Texto del Ejercicio Práctico

Crea un formulario de Google que incluya las siguientes preguntas:

1. Curso (opciones: Python, PHP, Java, Diseño Frontend, SQL, Testing & QA).
2. Sede (opciones: Central, Norte, Sur, Este, Oeste, Noreste, Noroeste, Sureste, Suroeste).
3. Comisión (valores enteros entre 24001 y 24600).
4. Nombre y apellido del instructor.
5. Cantidad de inscriptos (valores enteros entre 20 y 50).
6. Promedio de calificación del instructor (valores entre 1 y 10, que admitan hasta 2 decimales y un 10% de valores en blanco).

Luego, genera un set de datos en formato CSV que simule la carga de 600 registros en el formulario. Asegúrate de que el promedio de calificación del instructor cumpla con los criterios especificados y que el campo de marca temporal no sea relevante.

Sets de Datos

- Formulario de Google para la recolección de datos.
- Set de datos en formato CSV.

¿Por qué importa esto en DataWise Consulting?

Crear y gestionar formularios efectivos es fundamental para DataWise Consulting. Esta actividad te permitirá entender cómo se recopila información valiosa que alimenta los análisis y reportes, facilitando así la personalización de cursos y mejorando la satisfacción de los clientes. Las decisiones basadas en datos precisos conducen a un mejor rendimiento en el mercado.

⛔ **Estos ejercicios son una simulación de cómo se podría resolver el problema en este contexto específico. Las soluciones encontradas no aplican de ninguna manera a todos los casos.**
Recuerda que las soluciones dependen de los sets de datos, el contexto y los requerimientos específicos de los stakeholders y las organizaciones.


Buenos Aires
aprende
Agencia de Habilidades para el Futuro
Buenos
Aires
Ciudad