Creado por:

Isabel Maniega

1.3. Validación e integridad de datos

1.3.1 Ejecutar y comprender métodos básicos de validación de datos.

En informática, la **validación de datos** es el proceso de garantizar que los datos hayan sido sometidos a una limpieza para garantizar que tengan calidad de datos, es decir, que sean correctos y útiles. Utiliza rutinas, a menudo denominadas "reglas de validación", "restricciones de validación" o "rutinas de verificación", que verifican la corrección, la significatividad y la seguridad de los datos que se ingresan al sistema. Las reglas se pueden implementar a través de las funciones automatizadas de un diccionario de datos o mediante la inclusión de una lógica de validación explícita del programa de aplicación de la computadora y su aplicación.

Esto es distinto de la verificación formal, que intenta probar o refutar la corrección de los algoritmos para implementar una especificación o propiedad.

Descripción general

La validación de datos tiene como objetivo proporcionar ciertas garantías bien definidas de idoneidad y consistencia de los datos en una aplicación o sistema automatizado. Las reglas de validación de datos se pueden definir y diseñar utilizando diversas metodologías y se pueden implementar en diversos contextos. Su implementación puede utilizar reglas de integridad de datos declarativas o reglas de negocio basadas en procedimientos.

Tenga en cuenta que las garantías de validación de datos no incluyen necesariamente la precisión, y es posible que se acepten como válidos errores de entrada de datos, como errores de ortografía. Se pueden aplicar otros controles administrativos o informáticos para reducir la inexactitud dentro de un sistema.

Diferentes tipos

Al evaluar los aspectos básicos de la validación de datos, se pueden hacer generalizaciones con respecto a los diferentes tipos de validación según su alcance, complejidad y propósito.

Por ejemplo:

Validación de tipo de datos; Validación de rango y restricción; Validación de código y referencia cruzada; Validación estructurada; y Validación de consistencia

Verificación de tipo de datos: La validación de tipo de datos se lleva a cabo habitualmente en uno o más campos de datos simples. El tipo más simple de validación de tipo de datos verifica que los caracteres individuales proporcionados a través de la entrada del usuario sean consistentes con los caracteres esperados de uno o más tipos de datos primitivos conocidos, tal como se definen en un lenguaje de programación o en un mecanismo de almacenamiento y recuperación de datos.

Por ejemplo, un campo entero puede requerir que la entrada utilice solo los caracteres del 0 al 9.

Verificación simple de rango y restricción: La validación simple de rango y restricción puede examinar la entrada para comprobar su coherencia con un rango mínimo/máximo, o su coherencia con una prueba para evaluar una secuencia de caracteres, como una o más pruebas con expresiones regulares. Por ejemplo, puede requerirse que el valor de un contador sea un entero no negativo, y puede requerirse que una contraseña cumpla con una longitud mínima y contenga caracteres de varias categorías.

Verificación de código y referencias cruzadas: La validación de código y referencias cruzadas incluye operaciones para verificar que los datos sean consistentes con una o más reglas, requisitos o recopilaciones posiblemente externos relevantes para una organización, contexto o conjunto de suposiciones subyacentes en particular. Estas restricciones de validez adicionales pueden implicar la referencia cruzada de los datos proporcionados con una tabla de búsqueda conocida o un servicio de información de directorio como LDAP.

Por ejemplo, se podría requerir un código de país proporcionado por el usuario para identificar una región geopolítica actual.

1.3.2 – Establecer y mantener la integridad de los datos mediante reglas de validación claras.

Comprender el concepto de integridad de los datos y su importancia para mantener bases de datos confiables y precisas.

Definición de integridad de datos

La integridad de los datos es un concepto y proceso que garantiza la precisión, integridad, consistencia y validez de los datos de una organización. Al seguir el proceso, las organizaciones no solo aseguran la integridad de los datos, sino que también garantizan que tienen datos precisos y correctos en su base de datos.

La importancia de la integridad de los datos aumenta a medida que los volúmenes de datos continúan aumentando exponencialmente. Las grandes organizaciones dependen cada vez más de la integración de datos y de la capacidad de interpretar la información con precisión para predecir el comportamiento de los consumidores, evaluar la actividad del mercado y mitigar los posibles riesgos de seguridad de los datos. Esto es

fundamental para la minería de datos, de modo que los científicos de datos puedan trabajar con la información correcta.

Tipos de integridad de datos

Las organizaciones pueden mantener la integridad de los datos mediante restricciones de integridad, que definen las reglas y los procedimientos en torno a acciones como la eliminación, la inserción y la actualización de información. La definición de integridad de datos se puede aplicar tanto en bases de datos jerárquicas como relacionales, como los sistemas de planificación de recursos empresariales (ERP), gestión de relaciones con los clientes (CRM) y gestión de la cadena de suministro (CRM).

Las organizaciones pueden lograr la integridad de los datos mediante lo siguiente:

Integridad física

La integridad física significa proteger la precisión, exactitud y totalidad de los datos cuando se almacenan y recuperan. Esto suele verse comprometido por problemas como cortes de energía, erosión del almacenamiento, piratas informáticos que atacan las funciones de la base de datos y desastres naturales, que impiden el almacenamiento y la recuperación precisos de los datos.

Integridad lógica

La integridad lógica garantiza que los datos permanezcan inalterados mientras se utilizan de diferentes maneras a través de bases de datos relacionales. Este enfoque también tiene como objetivo proteger los datos de problemas de piratería o errores humanos, pero lo hace de manera diferente a la integridad física.

La integridad lógica se presenta en cuatro formatos diferentes:

1. **Integridad de entidad** La integridad de entidad es una característica de los sistemas de relaciones que almacenan datos dentro de tablas, que se pueden usar y vincular de varias maneras. Se basa en claves primarias y valores únicos que se crean para identificar un dato. Esto garantiza que los datos no se puedan enumerar varias veces y que los campos de una tabla no puedan ser nulos.

2. **Integridad referencial** La integridad referencial es una serie de procesos que garantizan que los datos permanezcan almacenados y se utilicen de manera uniforme. Las estructuras de bases de datos están integradas con reglas que definen cómo se utilizan las claves externas, lo que garantiza que solo se puedan realizar modificaciones, modificaciones y eliminaciones de datos adecuadas. Esto puede evitar la duplicación de datos y garantizar la precisión de los datos.

3. **Integridad del dominio** La integridad del dominio es una serie de procesos que garantizan la precisión de los datos dentro de un dominio. Un dominio se clasifica por un conjunto de valores que las columnas de una tabla pueden contener, junto con restricciones y medidas que limitan la cantidad, el formato y el tipo de datos que se pueden ingresar.

4. **Integridad definida por el usuario** La integridad definida por el usuario significa que los usuarios crean reglas y restricciones en torno a los datos para alinearlos con sus requisitos específicos. Esto generalmente se usa cuando otros procesos de integridad no salvaguardan los datos de una organización, lo que permite la creación de reglas que incorporan las medidas de integridad de datos de una organización.

Integridad de datos vs. calidad de datos La calidad de los datos es una pieza crucial del rompecabezas de la integridad de los datos. Permite a las organizaciones cumplir con sus estándares de datos y garantizar que la información se ajuste a sus requisitos con una variedad de procesos que miden la antigüedad, precisión, integridad, relevancia y confiabilidad de los datos. La calidad de los datos va un paso más allá al implementar procesos y reglas que rigen la entrada, el almacenamiento y la transformación de los datos.

Integridad de datos vs. seguridad de datos La seguridad de los datos implica proteger los datos del acceso no autorizado y evitar que se corrompan o roben. La integridad de los datos suele ser un beneficio de la seguridad de los datos, pero solo se refiere a la precisión y validez de los datos, en lugar de a su protección.

Integridad de datos y cumplimiento del RGPD La integridad de los datos es un proceso clave para ayudar a las organizaciones a cumplir con las regulaciones de privacidad y protección de datos, como el Reglamento General de Protección de Datos (RGPD) de la Unión Europea.

¿Cuáles son algunos de los riesgos de integridad de datos? Las amenazas clave para las organizaciones que garantizan la integridad de los datos incluyen:

- El **error humano** ofrece un importante riesgo de integridad de datos para las organizaciones. Esto suele deberse a que los usuarios introducen datos duplicados o incorrectos, eliminan datos, no siguen los protocolos o cometen errores con los procedimientos establecidos para proteger la información.

- **Errores y virus** Los piratas informáticos amenazan la integridad de los datos de las organizaciones al utilizar software, como malware, spyware y virus, para atacar las computadoras en un intento de robar, modificar o eliminar datos de los usuarios.

- **Errores de transferencia** Si no se pueden transferir datos entre ubicaciones de bases de datos, significa que se ha producido un error de transferencia. Estos se producen cuando los datos se encuentran en la tabla de destino pero no en la tabla de origen de una base de datos relacional.

- **El hardware comprometido** puede provocar fallos en el dispositivo o el servidor y otras fallas y mal funcionamiento de la computadora. En consecuencia, los datos pueden procesarse de forma incompleta o incorrecta, el acceso a los datos puede eliminarse o limitarse, o los datos pueden volverse difíciles de manejar para los usuarios.

¿Cómo garantizar la integridad de los datos? Para evitar los problemas y riesgos mencionados anteriormente, es necesario preservar la integridad de los datos mediante procesos como:

- **Validar la entrada** La entrada de datos debe validarse y verificarse para garantizar su precisión. La validación de la entrada es importante cuando los datos provienen de fuentes conocidas y desconocidas, como aplicaciones, usuarios finales y usuarios malintencionados.
- **Eliminar datos duplicados** Es importante asegurarse de que los datos confidenciales almacenados en bases de datos seguras no se puedan duplicar en documentos, correos electrónicos, carpetas u hojas de cálculo disponibles públicamente. Eliminar datos duplicados puede ayudar a evitar el acceso no autorizado a datos críticos para la empresa o información de identificación personal (PII).
- Las **copias de seguridad de los datos** son fundamentales para la seguridad y la integridad de los datos. Realizar copias de seguridad de los datos puede evitar que se pierdan de forma permanente y debe realizarse con la mayor frecuencia posible. Las copias de seguridad de los datos son especialmente importantes para las organizaciones que sufren ataques de ransomware, ya que les permiten restaurar versiones recientes de sus bases de datos y documentos.
- **Controles de acceso** La aplicación de controles de acceso adecuados también es importante para mantener la integridad de los datos. Esto depende de la implementación de un enfoque de acceso a los datos con los privilegios mínimos, que garantice que los usuarios solo puedan acceder a los datos, documentos, carpetas y servidores que necesitan para hacer su trabajo correctamente. Esto limita las posibilidades de que los piratas informáticos puedan hacerse pasar por usuarios y evita el acceso no autorizado a los datos.
- **Mantenga siempre un registro de auditoría** En caso de que se produzca una infracción, es fundamental que las organizaciones puedan descubrir rápidamente el origen del evento. Un registro de auditoría permite a las empresas realizar un seguimiento de lo que sucedió y cómo se produjo una infracción, y luego encontrar el origen del ataque.

Aplique reglas de validación claras que garanticen la exactitud y la coherencia de los datos.

La validación de datos es un componente fundamental de la gestión de datos, que garantiza que la información dentro de un sistema siga siendo precisa, confiable y coherente.

Este proceso de cuatro pasos es clave para lograr estos objetivos. A continuación, se detalla el proceso de cuatro pasos:

- Entrada de datos

- Definición de reglas de validación
- Proceso de validación
- Manejo de errores

Veámoslos en detalle.

1. **Entrada de datos:** el punto de partida de la calidad de los datos La validación de datos comienza en la etapa de entrada de datos, donde se recopila información sin procesar y se ingresa en un sistema. Este proceso implica:

- Recopilación de datos: la información se recopila de varias fuentes, incluidos formularios físicos, encuestas en línea, lecturas de sensores y más.
- Entrada manual frente a automática: los datos pueden ingresarse manualmente por personas o automáticamente a través de herramientas de software, según el volumen y la complejidad de los datos.
- Limpieza de datos: antes de ingresar, los datos pueden someterse a una limpieza para eliminar duplicados, corregir errores y estandarizar formatos, lo que garantiza un punto de partida limpio y confiable.
- Validación de datos en la entrada: algunos sistemas incorporan comprobaciones de validación básicas, como garantizar que se completen los campos obligatorios, que los tipos de datos sean correctos y que se respete el formato básico.

2. **Definición de reglas de validación:** Establecer estándares de datos El segundo paso en la validación de datos es definir reglas de validación, que establecen los criterios para determinar qué constituye datos válidos. Estas reglas pueden ser muy variadas e incluyen:

- Comprobaciones de tipo de datos: garantizar que los datos sean del tipo esperado (por ejemplo, texto, números, fechas).
- Comprobaciones de rango: verificar que los datos numéricos se encuentren dentro de rangos aceptables.
- Comprobaciones de formato: garantizar que los datos se ajusten a formatos específicos (por ejemplo, direcciones de correo electrónico o números de teléfono).
- Comprobaciones de integridad referencial: garantizar que se mantengan las relaciones y referencias de los datos.

3. **Proceso de validación:** Evaluar los datos en relación con las reglas Una vez que se establecen las reglas de validación, los datos se someten a una evaluación rigurosa en relación con estos criterios:

- Comparación de datos: los datos se comparan con las reglas definidas para determinar si son válidos. Los datos que cumplen con los criterios se consideran válidos, mientras que los que no los cumplen se marcan como no válidos.

4. **Manejo de errores:** Cómo manejar datos no válidos El paso final de la validación de datos implica manejar los datos que no cumplen con los criterios de validación:

- Solicitud de corrección: se puede solicitar a los usuarios o al personal de ingreso de datos que corrijan los datos no válidos.

- Corrección automática: en algunos casos, el sistema puede aplicar correcciones automáticas si los problemas son sencillos.
- Rechazo o notificación: se pueden rechazar los datos no válidos o se pueden enviar notificaciones al personal correspondiente para su revisión y corrección manual.

Al seguir estos cuatro pasos, la validación de datos garantiza que los datos sigan siendo precisos y confiables, lo que evita errores, inconsistencias e imprecisiones que podrían afectar la calidad de los datos y la toma de decisiones. Es una práctica fundamental para mantener la integridad y la confiabilidad de los datos.

Creado por:

Isabel Maniega