Mais sobre integridade dos dados e conformidade

Esta leitura ilustra a importância da integridade dos dados usando um exemplo de dados de uma empresa global. As definições dos termos relevantes para a integridade dos dados serão fornecidas no final.

CENÁRIO: DATAS DO CALENDÁRIO PARA UMA EMPRESA GLOBAL

As datas do calendário são representadas em muitas formas curtas diferentes. Dependendo de onde você mora, um formato diferente pode ser usado.

- Em alguns países, **12/10/20** (DD/MM/AA) significa 12 de outubro de 2020.
- Em outros países, o padrão nacional é YYYY-MM-DD, então 12 de outubro de 2020 se torna **2020-10-12**.
- Nos Estados Unidos, (MM/DD/AA) é o formato aceito, então 12 de outubro de 2020 será **12/10/20**.

Agora, pense no que aconteceria se você estivesse trabalhando como analista de dados para uma empresa global e não verificasse os formatos de data. Bem, sua integridade de dados provavelmente seria questionável. Qualquer análise dos dados seria imprecisa. Imagine encomendar estoque extra para dezembro, quando na verdade era necessário em outubro!

Uma boa análise depende da integridade dos dados, e a integridade dos dados geralmente depende do uso de um formato comum. Portanto, é importante verificar novamente como as datas são formatadas para garantir que o que você acha que é 10 de dezembro de 2020 não seja realmente 12 de outubro de 2020 e vice-versa.

Aqui estão algumas outras coisas a serem observadas:

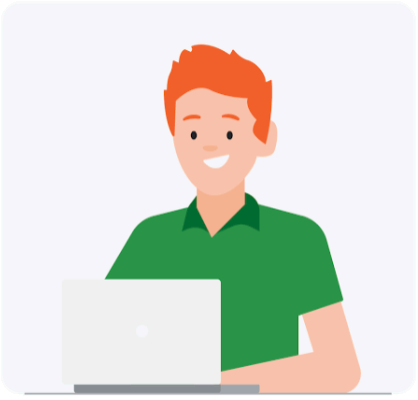
- **Replicação de dados comprometendo a integridade dos dados:** Continuar com o exemplo, imagine que você peça aos seus colegas internacionais para verificar as datas e manter um formato. Um analista copia um grande conjunto de dados para verificar as datas. Mas por causa de problemas de memória, apenas parte do conjunto de dados é realmente copiado. O analista estaria verificando e padronizando dados incompletos. Esse conjunto de dados parcial seria certificado como compatível, mas o conjunto de dados completo ainda conteria datas que não foram verificadas. Duas versões de um conjunto de dados podem apresentar resultados inconsistentes. Uma auditoria final dos resultados seria essencial para revelar o que aconteceu e corrigir todas as datas.

- **Transferência de dados comprometendo a integridade dos dados:** Outro analista verifica as datas em uma planilha e opta por importar os dados validados e padronizados de volta para o banco de dados. Mas suponha que o campo de data da planilha foi classificado incorretamente como um campo de texto durante o processo de importação (transferência) de dados. Agora, algumas das datas no banco de dados são armazenadas como sequências de texto. Neste ponto, os dados precisam ser limpos para restaurar sua integridade.

- **Manipulação de dados comprometendo a integridade dos dados:** Ao verificar as datas, outro analista percebe o que parece ser um registro duplicado no banco de dados e o remove. Mas acontece que o analista removeu um registro exclusivo da subsidiária de uma empresa e não um registro duplicado da empresa. Seu conjunto de dados agora está com dados ausentes e os dados devem ser restaurados para que estejam completos.

Conclusão

Felizmente, com um formato de data padrão e conformidade de todas as pessoas e sistemas que trabalham com os dados, a integridade dos dados pode ser mantida. Mas não importa de onde vêm seus dados, certifique-se sempre de verificar se eles são válidos, completos e limpos antes de iniciar qualquer análise.



REFERÊNCIA: RESTRIÇÕES DE DADOS E EXEMPLOS

Conforme você progride em sua jornada de dados, você encontrará muitos tipos de restrições de dados (ou critérios que determinam a validade). A tabela abaixo oferece definições e exemplos de termos de restrição de dados que você pode encontrar.

Restrição de dados	Definição	Exemplos
Tipo de dado	Os valores devem ser de um determinado tipo: data, número, porcentagem, booleano, etc.	Se o tipo de dados for uma data, um único número como 30 falharia na restrição e seria inválido.
Intervalo de dados	Os valores devem estar entre os valores máximo e mínimo predefinidos	Se o intervalo de dados for de 10 a 20, um valor de 30 falharia na restrição e seria inválido.
Obrigatório	Os valores não podem ser deixados em branco ou vazios	Se a idade for obrigatória, esse valor deve ser preenchido.
Único	Os valores não podem ter um duplicado	Duas pessoas não podem ter o mesmo número de celular na mesma área de serviço.
Padrões de expressão regular (regex)	Os valores devem corresponder a um padrão prescrito	Um número de telefone deve corresponder a ###-###-#### (nenhum outro caractere é permitido).
Validação de campo cruzado	Certas condições para vários campos devem ser atendidas	Os valores são porcentagens e os valores de vários campos devem somar 100%.
Chave primária	(Somente bancos de dados) o valor deve ser exclusivo por coluna	Uma tabela de banco de dados não pode ter duas linhas com o mesmo valor de chave primária. Uma chave primária é um identificador em um banco de dados que faz referência a uma coluna na qual cada valor é exclusivo. Mais informações sobre chaves primárias e estrangeiras são fornecidas posteriormente no programa.

Conjunto de membros	(Somente bancos de dados) os valores de uma coluna devem vir de um conjunto de valores discretos	O valor de uma coluna deve ser definido como Sim, Não ou Não aplicável.
Chave estrangeira	(Somente bancos de dados) os valores de uma coluna devem ser valores exclusivos provenientes de uma coluna em outra tabela	Em um banco de dados de contribuintes dos EUA, a coluna Estado deve ser um estado ou território válido com o conjunto de valores aceitáveis definidos em uma tabela de estados separada.
Precisão	O grau em que os dados estão de acordo com a entidade real que está sendo medida ou descrita	Se os valores dos códigos postais forem validados pela localização da rua, a precisão dos dados aumenta.
Integridade	O grau em que os dados contêm todos os componentes ou medidas desejados	Se os dados de perfis pessoais exigirem a cor do cabelo e dos olhos, e ambos forem coletados, os dados estarão completos.
Consistência	O grau em que os dados são repetíveis de diferentes pontos de entrada ou coleta	Se um cliente tiver o mesmo endereço nos bancos de dados de vendas e reparos, os dados serão consistentes.