Universidad técnica particular de Loja.

Nombre: Roberto Bustamante

Fecha: 23/04/2025

1. **¿Por qué es importante la calidad de los datos?**

La salud de sus datos afecta directamente la efectividad de varios marcos cruciales que fortalecen a su organización. Garantizar la exactitud de sus datos le permite fortalecer activamente las mismas herramientas que utiliza para gestionarlos y analizarlos. Es probable que su marco de gobernanza de datos no logre hacer cumplir los controles de acceso de manera adecuada o garantizar el cumplimiento total si sus datos están plagados de errores e inconsistencias. Lo mismo se aplica a la seguridad de los datos. Los datos sucios, con errores e información faltante, dificultan que sus equipos de datos identifiquen actividades sospechosas o aíslen amenazas.

La calidad de los datos también afecta la confiabilidad y usabilidad de su catálogo de datos.

Los datos de alta calidad conducen a un catálogo útil, y un catálogo de datos bien mantenido facilita prácticas efectivas de gestión de la calidad de los datos.

2. **¿Cuáles son los problemas de calidad de datos más comunes?**

Algunos de los problemas más comunes de calidad de datos incluyen datos incompletos o faltantes, formatos inconsistentes, datos inexactos, datos duplicados y datos obsoletos. Estos problemas pueden generar informes inexactos, una toma de decisiones ineficaz y mayores costos para las empresas.

- **Datos incompletos o faltantes**

Los datos incompletos o faltantes se refieren a situaciones en las que los campos obligatorios se dejan en blanco o no se proporcionan. Esto puede generar análisis e informes inexactos y dar lugar a decisiones comerciales incorrectas.

- **Formatos de datos inconsistentes**

Los formatos de datos inconsistentes se refieren a situaciones en las que los mismos datos se representan de diferentes maneras en múltiples sistemas o fuentes. Esto puede dificultar la integración de datos de diferentes fuentes y generar errores en el análisis y la generación de informes.

- **Datos inexactos**

Los datos inexactos son aquellos que son incorrectos, ya sea por errores de introducción o por información desactualizada. Esto puede generar informes

incorrectos, una toma de decisiones ineficaz y un aumento de los costos para las empresas.

- **Datos duplicados**

Los datos duplicados se refieren a múltiples instancias de los mismos datos existentes en diferentes sistemas o fuentes. Esto puede generar inconsistencias en los datos y errores en el análisis y la elaboración de informes.

- **Datos obsoletos**

Los datos obsoletos son aquellos que ya no son relevantes ni actuales. Esto puede generar análisis e informes incorrectos y, por lo tanto, tomar decisiones comerciales incorrectas.

3. **¿Cuáles son los principales atributos (métricas) de los datos de calidad?**

Estas métricas evalúan los datos en cuatro dimensiones clave:

Intrínseco: Se centra en la credibilidad, objetividad y reputación de los datos.

Contextual: Enfatiza la relevancia, puntualidad e integridad de los datos.

Figurativo: Se centra en el formato y presentación de los datos.

Accesibilidad: Se ocupa de la facilidad de acceso a los datos.

Estas dimensiones de calidad de los datos son esenciales para marco de calidad de datos y ayudar a garantizar que los datos sean completos y confiables. Al utilizar métricas de calidad de datos, puede establecer objetivos específicos para guiar a sus equipos a abordar los problemas de calidad de datos que ocurren comúnmente.

4. **¿Cuáles son los métodos que podemos usar para mejorar/corregir la calidad de los datos?**

1. **Gobernanza de datos**: Definir roles y responsabilidades claras (quién ingresa, aprueba y supervisa datos) y establecer estándares para su manejo.

2. **Validación en la entrada**: Implementar controles (campos obligatorios, formatos, listas de referencia) para evitar errores desde el origen.

3. **Limpieza y de duplicación**: Programar procesos periódicos (por ejemplo, trimestrales) que detecten y eliminen registros redundantes o inconsistentes.

4. **Perfilado y monitoreo**: Emplear herramientas que identifiquen patrones, valores atípicos y anomalías, alertando automáticamente sobre variaciones de calidad.

5. **Estandarización de formatos**: Crear un diccionario de datos con definiciones y formatos uniformes (por ejemplo, MM/DD/AAAA para fechas, capitalización adecuada).

6. **Gestión de datos maestros (MDM)**: Mantener una única fuente de verdad para datos críticos (clientes, productos, proveedores) con procesos de actualización regulares.

7. **Auditorías de calidad**: Realizar revisiones automatizadas y manuales periódicas, documentar hallazgos y corregir errores o mejorar procedimientos.

8. **Capacitación continua**: Formar al personal en políticas, impacto de la calidad de datos y mejores prácticas para mantener estándares altos.

9. **IA y aprendizaje automático**: Utilizar algoritmos para analizar grandes volúmenes de datos, detectar patrones y anomalías, y proponer alertas que faciliten la corrección rápida.

5. ¿Qué es Data Wrangling?

Data wrangling también se conoce como organización de datos.

Es un término general que describe varios procesos, todos diseñados para tomar datos sin procesar y transformarlos de conjuntos de datos complejos y desordenados en formatos más fáciles de usar. Cuando participas en la disputa de datos, encuentras y transformas datos para que puedas usarlos para responder una pregunta o producir una información valiosa necesaria para tomar decisiones.

Los profesionales llevan a cabo el data wrangling de una de dos maneras: manual o automatizada. En las empresas con un equipo de datos, los científicos de datos y otros miembros del equipo suelen encabezar el proceso de disputa de datos. En organizaciones más pequeñas, puede ser responsabilidad de los profesionales que no son de datos limpiar los datos antes de usarlos.

6. ¿Qué herramientas/funciones/paquetes R podemos usar para gestionar la calidad de los datos?

1. dplyr: Dominando la Manipulación de Datos

Desde filtrar filas de datos hasta modificar variables, dplyr es tu conjunto de herramientas. Familiarízate con funciones como filter(), select() y mutate().

2. ggplot2: Eleva tu Juego de Visualización de Datos

Con ggplot2, tus datos no solo hablan, ¡suenan! Entiende la función aes() y aprende a superponer tus gráficos para obtener visualizaciones más ricas.

3. GWalkR: Convierte tus Datos en una Aplicación de Visualización Interactiva

GWalkR es una herramienta interactiva de Análisis Exploratorio de Datos (EDA) en R construida por Kanaries. Integra los htmlwidgets con Graphic Walker. Puede simplificar tu flujo de trabajo de análisis y visualización de datos en R, convirtiendo tu marco de datos en una interfaz de usuario similar a Tableau para exploración visual.

4. tidyr: El Arte de Organizar Datos

tidyr asegura que tus datos sean ordenados y accesibles. Explora spread(), gather() y más para estructurar tus datos de manera perfecta.

5. readr: Simplifica la Entrada y Salida de Datos

Ya sea que estés leyendo un archivo CSV o exportando datos, readr lo simplifica todo. Sumérgete en funciones como read_csv() y write_csv().

6. caret: Aprendizaje Automático Simplificado

Para el entrenamiento de modelos y el aprendizaje automático en R, caret es indispensable. Entrena modelos, realiza predicciones y evalúa el rendimiento de forma fluida.

7. ¿Cómo gestionar la calidad de datos en la organización?

1- Establecer criterios de validación de datos:

La validación de datos es un paso crucial para garantizar su calidad. Define criterios y reglas claras para validar los datos en función de su integridad, consistencia y relevancia. Esto implica verificar la exactitud de los campos, la coherencia entre diferentes conjuntos de datos y la adecuación de los datos según los estándares establecidos. Al implementar criterios de validación sólidos, puedes prevenir la entrada de datos incorrectos o incompletos desde el principio.

2- Realizar una limpieza de datos periódica:

Los datos no siempre son perfectos y pueden contener errores, duplicados o información obsoleta. Llevar a cabo una limpieza periódica de los datos es esencial para mantener su calidad. Utiliza herramientas y técnicas de limpieza de datos para eliminar registros duplicados, corregir errores tipográficos, normalizar formatos y eliminar datos incompletos o irrelevantes. Esto mejorará la precisión y confiabilidad de tus análisis y reducirá la posibilidad de tomar decisiones basadas en información incorrecta.

3- Establecer mecanismos de control de calidad de datos:

Implementa un proceso de control de calidad de datos que garantice la precisión y la coherencia en todo momento. Esto puede incluir la revisión manual de registros críticos, la realización de auditorías periódicas de datos y la implementación de pruebas automatizadas para detectar posibles

problemas. Al tener mecanismos de control de calidad en su lugar, podrás identificar y corregir rápidamente cualquier desviación en la calidad de los datos.

4- Actualizar constantemente los datos:

La información empresarial está en constante cambio, por lo que es crucial mantener tus datos actualizados. Establece procesos para verificar y actualizar regularmente los datos relevantes. Esto puede implicar la actualización de información de contacto de clientes, la revisión de datos demográficos o la incorporación de nuevas fuentes de datos pertinentes. Mantener los datos actualizados garantiza que estés tomando decisiones basadas en información precisa y te ayuda a mantener una ventaja competitiva.

5- Capacitar al personal en la gestión de datos:

La calidad de los datos no es solo responsabilidad de un equipo o departamento; es una tarea que involucra a toda la organización. Proporciona capacitación y concienciación sobre la importancia de la calidad de los datos a todos los empleados. Enséñales cómo recopilar, ingresar y gestionar datos correctamente, y fomenta una cultura en la que se valora la integridad y la precisión de los datos. Al capacitar a tu personal en la gestión de datos, estarás fortaleciendo la calidad de los mismos en todas las áreas de tu empresa.

Bibliografía.

6 Grandes Paquetes de R que Todo Principiante Debe Conocer – Kanaries. (n.d.). https://docs.kanaries.net/es/topics/R/6-r-lib-for-beginners

Boomi. (2025, January 9). *9 Ways to Ensure Data Quality*. https://boomi.com/blog/9-ways-to-ensure-data-quality/

Cawsey, M. (n.d.). *Top 5 Most Common Data Quality Issues ➤*. Stibo Systems. https://www.stibosystems.com/blog/data-quality-issues

Haider, K. (2024, April 24). What Is Data Quality and Why Is It Important? | Astera. *Astera*. https://www.astera.com/es/type/blog/data-quality/

Shahid, A. (2025, March 7). 7 Data Quality Metrics to Assess Your Data Health | Astera. *Astera*. https://www.astera.com/es/type/blog/data-quality-metrics/

Sivori, A. (2023, July 21). *5 buenas prácticas para mejorar la calidad de los datos en una empresa*. WaveBI - Data Analytics. https://www.wavebi.com.ar/noticias/5-buenas-practicas-para-mejorar-la-calidad-de-los-datos-en-una-empresa/

Staff, C. (2023, November 29). *¿Qué es data wrangling y por qué es importante?* Coursera. https://www.coursera.org/mx/articles/data-wrangling